Longitudinal Surveys: from Design to Analysis

Proceedings

Statistics Canada's XXV International Symposium on Methodological Issues

October 27-30, 2009

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre	
Toll-free telephone (Canada and the United States):	
Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369
Local or international calls:	
Inquiries line	1-613-951-8116
Fax line	1-613-951-0581
Depository Services Program	
Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

Mail:
Statistics Canada
100, Tunney's Pasture Driveway
Ottawa, Ontario
K1A 0T6

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed *standards of service* that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" >"The agency" > "Providing services to Canadians."

Symposium 2009 — Catalogue no. 11-522-XCB

End-use licence agreement

Published by authority of the Minister responsible for Statistics Canada

For further information please contact:
Licensing Services
Client Services Division, Statistics Canada
R.H. Coats Building, 9th floor, section A
Ottawa, Ontario K1A 0T6, Canada
E-mail: licensing@statcan.gc.ca
Telephone: (613) 951-1122
Fax: (613) 951-1134

LONGITUDINAL SURVEYS: FROM DESIGN TO ANALYSIS

TABLE OF CONTENTS

Preface

Symposium 2009 was the twenty-fifth in Statistics Canada's series of international symposia on methodological issues. Each year the symposium focuses on a particular theme. In 2009, the theme was: "**Longitudinal Surveys: From Design to Analysis**".

Symposium 2009 was held from October 27-30, 2009, at the Palais des congrès in Gatineau, Québec, and it attracted more than 500 people from several countries. Two workshops and 69 papers were presented. Aside from translation and formatting, the papers, as submitted by the authors, have been reproduced in these proceedings.

The organizers of Symposium 2009 would like to acknowledge the contribution of the many people, too numerous to mention individually, who helped make it a success. The organizers would also like to thank the presenters and authors for their presentations and for putting them in written form. Finally, the organizers thank all the participants that went to the different presentations.

The Symposium 2009 Organizing Committee

Logistics and operations
Julie Trépanier, Chair
Asma Alavi
Jean-Luc Bernier
Sylvie Gauthier
Julie Girard
Denis Lemire
Caroline Rondeau

Program
Christian Nadeau, Chair
Abdelnasser Saïdi
Claude Turmelle

OPENING REMARKS

Opening Remarks

François Maranda[1]

Good morning,

On behalf of Statistics Canada, I would like to extend a warm welcome to you all, friends and colleagues, to Symposium 2009.

This year, our symposium, entitled "Longitudinal Surveys: From Design to Analysis," marks a special occasion. This is the 25th International Symposium organized by Statistics Canada on survey methodology. The first was held in 1984 and dealt with "Survey Data Analysis."

Whether you are familiar with this symposium or are joining us for the first time, we are pleased to have you with us to celebrate this 25th anniversary.

Between the first Symposium and this one, several topics have been discussed over the years, each as relevant as the next. I invite you to discover or rediscover them by visiting the exhibition in the Hall throughout the week.

Before addressing the topic for this year, let me offer an explanation for the longevity of this symposium. It was established to contribute to the advancement of survey methodology as a discipline and promote the resolution of issues shared by various statistical agencies while bringing together experts from various backgrounds and creating a collaborative environment. The longevity of the Symposium and the ever-increasing interest it draws lead me to believe that this objective has been fully achieved.

From the point of view of Statistics Canada, the Symposium is still today an invaluable source of information as part of efforts to address current and emerging challenges facing us in terms of survey methodology. The relevance and value of exchanges that occur at the Symposium are closely connected to your presence, thus enabling us to share views, experiences and diverse expertise for the benefit of all participants year after year.

By giving us the opportunity to learn how other organizations tackle challenges similar to ours, the Symposium helps to resolve our issues, critically examine our work and develop our staff.

Judging by the number of participants and how far they have travelled to be here, it seems that the benefits of our symposia are also recognized internationally. In addition to the 300 participants from Statistics Canada, there are about 175 people from 16 countries all over the world. Together, you represent more than 90 organizations from the academic community as well as the public and private sectors. It is truly impressive. Thank you for coming.

To get back to this year's topic, "Longitudinal Surveys: From Design to Analysis," this is not the first time that we have addressed longitudinal surveys at the Symposium. In 1989, a few presentations focused on the analysis of longitudinal data at a symposium with the topic "Analysis of Data Over Time." Then, in 1992, the Symposium "Design and Analysis of Longitudinal Surveys" was part of Statistics Canada's entry into the era of longitudinal surveys. The Survey of the Labour and Income Dynamics was to be launched a few months later, while the National Population Health Survey and the National Longitudinal Survey of Children and Youth were already planned for the following year. Together, these surveys provided a sound basis for addressing the dynamics of poverty, work, family, health and child development.

[1] François Maranda, Statistics Canada, 26-J RH Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6

Other initiatives subsequently emerged, still on the social statistics end, with the Youth in Transition Survey and the Longitudinal Survey of Immigrants to Canada, implemented in 2000 and 2001 respectively. These surveys made it possible to focus on the study of sub-populations in or on the verge of a major transition in their life, respectively the transition between education and the labor market and the integration into Canadian society following recent immigration. During the same period, a longitudinal survey initiative from an economic standpoint was born with the Workplace and Employee Survey, for which data collection began in 1999.

Longitudinal data sets thus became available in the late 1990s. Although the analysis was discussed at previous symposia, the Symposium held in 1998 on "Longitudinal Analysis for Complex Surveys" made it possible to review the wealth and analytical potential of the state of knowledge in analyzing such data. For Statistics Canada, the Symposium was also part of initiatives aimed at furthering longitudinal data analysis research.

So why address the topic of longitudinal surveys again in 2009? Given the diversity of experiences acquired over time, it seems that we are now at the assessment stage, with the goal of better meeting data needs. Time is the key element here. Unlike purely cross-sectional surveys, the relevance of the data gathered should increase with time. On the other hand, the phenomenon of attrition, which is closely related to time, reduces the accuracy of longitudinal data sets. Moreover, only time can provide assessment results that make it possible to determine the relevance of initial and subsequent choices, including those related to content as well as to methodology. It is now time to evaluate these experiences and determine future directions for such surveys. The challenges we face include those related to the length of longitudinal surveys, the representativeness of samples and the associated attrition, the integration of content and respondent burden, international comparability, data accessibility and privacy.

Over the next three days, there will be presentations on several of these aspects, plus some others on data collection, measurement errors and non-response, weighting and estimation, the use of administrative data, modeling and data analysis. I would like to thank all the presenters for sharing the fruits of their labor.

Yesterday, 140 of you attended one of the two full-day workshops. The first was on lessons and innovations from Statistics Canada's 15 years of experience with the Longitudinal Survey and the second was on the multilevel modeling of longitudinal data. I would particularly like to thank the leaders of yesterday's workshops for their significant contributions: Sophia Rabe-Hesketh, University of California and Anders Skrondal, Norwegian Institute of Public Health, as well as Michelle Simard and François Brisebois, Statistics Canada.

Finally, before I declare this 25th Symposium officially open, I would like to thank the organizing committee, session organizers, chairpersons and all the volunteers who have made this event possible.

My sincere thanks to each of you.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

A Methodological Research Agenda for Longitudinal Surveys

Peter Lynn[1]

Abstract

This presentation will propose an agenda for future methodological research on issues pertinent to longitudinal surveys. The agenda will be informed by a consideration of the methodological challenges that are unique to longitudinal surveys and a review of research designed to address those challenges. There will be a particular focus on recent and current research and discussion of the implications of recent research findings, of technological changes and of other innovations. The objectives are both to stimulate methodological research and to raise awareness of the limitations of current methodological knowledge.

Topics discussed will include sample design, between-wave intervals, keeping track of sample members and maintaining co-operation, adjustment for non-response and panel attrition, panel conditioning, and instrument design to minimise measurement error in measures of change.

[1] Peter Lynn, University of Essex, U.K. (plynn@essex.ac.uk)

EVENT HISTORY DATA COLLECTION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

A Triangulated Approach to Evaluating ELSA's Event History Calendar

Alice McGee, Hayley Cripps and Joanne Pascale[1]

Abstract

While the Event History Calendar (EHC) technique has been used in surveys for several decades, its popularity has begun to increase in part owing to recent research demonstrating that it can produce higher data quality than conventional questionnaires for some topic areas. However, there has been a call for systematic methodological research in this area in addressing two main areas: how the EHC works in practice, particularly with regard to the interviewer-respondent interaction, the use of, and receptivity to, the technique. The second area relates to landmarks, both internal and external. Little is known about the 'mechanics' of these landmarks, when they are introduced as memory aids, who introduces them, and how successful they are in helping respondents recall dates accurately.

This paper details a study that sought to address these unknowns through the evaluation of an EHC used as part of the English Longitudinal Study of Ageing (ELSA). A sample of 124 interviews was audio recorded and analysed using a range of quantitative and qualitative evaluation methods: an interviewer questionnaire; a respondent debriefing questionnaire; an interviewer debriefing; and behaviour coding.

This paper draws together and triangulates findings from these four evidence sources, addressing 6 specific research questions. It delves into the precise nature of the interaction between interviewer and respondent, examining receptivity to the method and how particular features played out in the field. It will also make reference to how landmarks were used, whether they assisted respondent recall and the nature of their use.

[1] Alice McGee and Hayley Cripps, National Centre for Social Research, UK; Joanne Pascale, U.S. Census Bureau, U.S.A. (joanne.pascale@census.gov)

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Tracing Life Courses with Prospective Panel Surveys – Lessons from the German Family Panel Study

Josef Brüderl, Laura Castiglioni, Ulrich Krieger, Volker Ludwig and Klaus Pforr [1]

Abstract

Collecting valid information on the occurrence, timing and spacing of events during the life course is a crucial issue for life course researchers. A prospective panel design can help to improve data quality, because only short periods of time between panel waves have to be recalled. At the "seam" of the two panel waves, however, misstated episodes are a common problem (seam effect).

In this presentation we argue that the seam effect can be reduced considerably by combining life history calendar and dependent interviewing techniques. Drawing on pretest data from the recently started German Family Panel Study, we provide evidence on the power of this approach in reducing the seam effect.

In wave two of our pretest study we administered a paper-and-pencil retrospective life history calendar from the age of 14 up until the date of the interview, covering relationship biography, fertility history, and residential mobility. In wave three, we asked for the events that occurred in between interviews. Using a split-ballot design we either presented a blank calendar or we presented a calendar, where the status at wave two was filled in (dependent interviewing). Differences between the two groups reveal how dependent interviewing helps to reduce the seam effect and thus can improve data quality.

[1] Josef Brüderl (jbruederl@sowi.uni-mannheim.de), Laura Castiglioni, Ulrich Krieger, Volker Ludwig and Klaus Pforr, University of Mannheim, Germany

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

A Comparison of Survey Reports Obtained via Standard Questionnaire and Event History Calendar

Jeffrey Moore, Jason Fields, Joanne Pascale, Gary Benedetto, Martha Stinson and Anna Chan [1]

Abstract

The US Census Bureau's Survey of Income and Program Participation (SIPP) provides monthly information about the nation's income, wealth, and program usage. Currently, SIPP administers an interview three times a year to each sample member; each interview's reference period covers the preceding four calendar months. In 2006 the Census Bureau initiated a SIPP re-engineering effort, a key component of which is a shift to a single annual interview covering the preceding calendar year. To accomplish this shift, the Census Bureau proposes to employ event history calendar (EHC) methods. Prior research, however, has raised some questions about EHC data quality for topics of key importance to SIPP, such as need-based program participation. In addition, the research base does not address the main SIPP design issue – the proposed shift from a four-month to a twelve-month reference period. To examine the implications of the switch to EHC methods and the expansion of the survey's reference period, the Census Bureau implemented the SIPP EHC Field Test in the spring of 2008. The essential feature of the test was an EHC reinterview of expired SIPP 2004 panel households. The reference period for the reinterview was calendar year 2007; the primary sample component consisted of cases which had already provided information about calendar year 2007 in the normal course of their final three SIPP interviews. The field test thus permits a direct comparison of standard questionnaire and EHC reports by the same people, about the same characteristics, and for the same period. The subject of this paper is the main component of the evaluation of the field test results: an examination of the correspondence of the two reports – one obtained from a standard questionnaire, the other from an EHC instrument – for several of the key characteristics of interest to SIPP, and for each month of 2007.

[1] Jeffrey Moore, Jason Fields (jason.m.fields@census.gov), Joanne Pascale (joanne.pascale@census.gov), Gary Benedetto (gary.linus.benedetto@census.gov), Martha Stinson (martha.stinson@census.gov) and Anna Chan, U.S. Census Bureau, U.S.A.

ATTRITION BIAS AND NONRESPONSE WEIGHTING

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Evaluation and Selection of Models for Attrition Nonresponse Adjustment

Eric V. Slud and Leroy Bailey[1]

Abstract

This paper studies weighting adjustment of nonresponse attrition in a longitudinal survey like the U.S. Survey of Income and Program Participation (SIPP). Adjustment factors are model-based, either constant over cells or derived from logistic regressions in terms of auxiliary covariates. Biases in estimated initialwave ('Wave 1') attribute totals are assessed between the survey-weighted estimator in the first wave and the weight-adjusted estimator for the same Wave-1 item totals based on later-wave respondents. New metrics of quality are defined for models used to adjust the longitudinal survey for attrition. The metrics combine cumulative maxima of estimated between-wave adjustment biases in survey attributes, based on randomly re-ordered subsets of the sample, relative to the estimated totals. Confidence bands for the metrics are estimated, and the metrics are applied to rank the quality of and to select among a collection of logistic-regression models for attrition nonresponse adjustment in SIPP 1996.

Key Words: Adjustment Cell, Logistic Regression, Weighting, Random Ordering, Raking, Subdomain.

1. Introduction

To measure the quality of adjustment for attrition in a longitudinal survey, one would certainly try to evaluate the biases of adjustments using external data on the sample frame and the survey variables whenever such external data are available. But external data for evaluation are seldom available, so survey investigators must usually evaluate and choose among competing adjustment methods based on comparisons internal to the survey. Yet there is little published methodological work on how to measure the biases of adjustment, from the internal evidence of a longitudinal survey.

There has been previous theoretical work on the large-sample behavior of model-based nonresponse weight adjustment methods. For example, Kim and Kim (2007) used weighted likelihood methods to choose between alternative parametric models for survey nonresponse. One of the rare papers explicitly assessing adjustment effectiveness using only evidence within the adjusted survey is Eltinge and Yansaneh (1997), which discussed several diagnostics and sensitivity checks for the definition of weighting adjustment cells in a survey. Dufour et al. (2001) studied internal longitudinal evaluation of nonresponse adjustment methods from a calibration perspective, assessing the magnitudes of calibration-based adjustments through a metric the authors defined in terms of weight changes through several stages of a weight-adjusted longitudinal survey.

Slud and Bailey (2006) studied the estimates and standard errors of differences, in the U.S. Survey of Income and Program Participation (SIPP), between Wave 1 totals of various 1996 cross-sectional survey items and the nonresponse-adjusted totals of the same Wave 1 items using response data from a later Wave (4 or 12) of the same survey. They studied nonresponse adjusted either via adjustment cells or by logistic regression models for the later-wave response indicators, finding that relative and standardized estimated biases varied considerably from one model to another. As in Dufour et al. (2001), many competing models could be defined, with different attribute variables used in constructing adjustment cells or as logistic regression predictors. Slud and Bailey (2006) noted that including `Poverty` as a predictor did have the effect, akin to raking, of making the sample-wide estimated Wave 1 adjustment bias of `Poverty` particularly small. They conjectured that this artificial effect would be removed by considering

[1] Eric V. Slud, US Census Bureau, SRD, 4600 Silver Hill Rd., Rm 5K004, Washington DC 20233-9100, and Math. Dept., Univ. of Maryland, College Park, Eric.V.Slud@census.gov; and Leroy Bailey, formerly of US Census Bureau, SRD.
The views expressed in this paper on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

estimated Wave 1 bias within a number of different subdomains. This suggests creating a composite metric by combining the estimated between-wave adjustment biases for various survey items across various subdomains.

Our goal in this research has been to devise metrics to aid in the comparison of different model-based methods of adjustment for nonresponse due to attrition, and thus to provide a basis for choosing among adjustment methods. Several earlier comparative investigations related to adjustment methods have been conducted — in particular Rizzo et al. 1994 who compared logistic-regression adjustment models in SIPP using many of the same variables we consider below — but none seem to have found a clear advantage for one adjustment method over others. Our approach is to adjust weights using models for later-stage response, of adjustment-cell or logistic regression type, to calculate the later-stage survey estimates of first-stage totals of population and other survey variables. We compare the relative biases of estimated late- versus early-stage weighted subtotals for various population subdomains, combining these estimates into two related metrics for adjustment effectiveness, and then use the metrics to choose among adjustment models within SIPP 1996.

The paper is organized as follows. Section 2 defines the metrics and presents bounds for them, theoretically valid when the adjustment model is correct, which can be applied in practice to flag inadequate weight adjustments. Section 3 applies the metrics to the comparison of a series of adjustment-cell or logistic regression models which might have been used in attrition adjustment of the SIPP 1996 data, expanding on those studied in Slud and Bailey (2006). The adjusted SIPP 96 totals and metrics in Section 3 are based upon first-stage weights (incorporating nonresponse adjustments up to Wave 1) and model-based later-wave adjustments, but no population controls. Then Section 4 provides the corresponding comparisons among Wave 1 totals based on Wave 1 and later-wave adjusted weights which are raked to Current Population Survey totals as was actually done in SIPP. Section 5 draws overall conclusions. The material of this paper is presented in greater detail in Slud and Bailey (2009).

2. Adjustment-Quality Metrics and Bounds

We formulate the survey design and nonresponse as a so-called quasi-randomization model (Oh and Scheuren 1983). Let S denote the sample of $n = |S|$ persons drawn from sampling frame U, with known or effectively adjusted single inclusion probabilities $\{\pi_i : i \in U\}$, and responding in Wave 1. For a set of cross-sectional survey measurements indexed by $k = 1,..., K$, such as the $K = 11$ items studied by Slud and Bailey (2006), denote by $y_i^{(k)}$ the Wave 1 item values and $\mathbf{x}_i$ a vector of auxiliary variable values for all $i \in U$. Let ρ_i denote individual response indicators in a specified later wave of the same survey, assumed independent and observed for all $i \in S$, and let $p_i = \mathrm{P}(\rho_i = 1|S)$ denote the unknown probabilities of individual response in that later wave. Let denote estimators based on known g of these unknown probabilities derived from a parametric $\hat{p}_i = g(\mathbf{x}_i, \hat{\vartheta})$ model using auxiliary data $\mathbf{x}_i$, with parameter estimators $\hat{\vartheta}$ obtained via estimating equations (Kim and Kim, 2007). For any population attribute $z_i, i \in U$, the frame-population total is $t_z = \sum_{i \in U} z_i$, and the corresponding Horvitz-Thompson estimator is $\hat{t}_z = \sum_{i \in S} z_i / \pi_i$.

For each survey item $y_i^{(k)}, i \in U$, and the adjustment strategy embodied in the estimated response probabilities $\hat{p}_i$, define the estimated nonresponse bias for each population subdomain $D \subset U$ as

$$\hat{B}_k(D) = \sum_{i \in D \cap S} \left((\rho_i / \hat{p}_i) - 1 \right) y_i^{(k)} / \pi_i \qquad (1)$$

In Slud and Bailey (2006) and earlier work of Bailey, the domain D was all of U, and $\hat{B}_k(U)$ was viewed as the difference between an adjusted estimator of $t_{y^{(k)}}$ based on data $(\rho_i, \rho_i y_i^{(k)}, \mathbf{x}_i : i \in S)$ and the ordinary Horvitz-Thompson estimator $\hat{t}_{y^{(k)}}$, and was regarded as an estimator of attrition nonresponse bias due to the method of adjustment.

When model-based adjustments are incorporated into the survey weighting for later waves, especially when model-terms related to specific survey items are introduced, our experience (Slud and Bailey 2006) suggests that the whole-population bias terms $\hat{B}_k(U)$ may be reduced much more than biases on smaller domains D. Our aim is to devise useful measures of maximum bias over many subdomains, favoring model-based adjustments which reduce biases over the interesting domains and discouraging models which correct for the biases of individual survey-item totals only over the whole population.

2.1 Relative Subdomain Bias

We now propose a measure of the typical relative bias in estimating item totals over subdomains. The idea is to consider the largest value of absolute relative bias $\hat{B}_k(D)/\hat{t}_{y^{(k)}}$ over a collection of subsets $D \subset U$. After re-ordering the elements of S by the random permutation $\tau = (\tau(1), \tau(2), ..., \tau(n))$, the largest absolute bias in survey variable *k* over consecutively τ-indexed subdomains of S is

$$\max\left\{\left|\hat{B}_k\left(\{\tau(1),...,\tau(\alpha)\}\right)\right| : 1 \le \alpha \le n\right\}.$$

Therefore, to measure the overall relative bias in estimating item *k* totals over subdomains, define

$$m_k = E_\tau\left(\max_{1\le\alpha\le n}\left|\hat{B}_k\left(\{\tau(1),...,\tau(\alpha)\}\right)\right|\right)/\hat{t}_{y^{(k)}} \qquad (2)$$

where the expectation is taken, for a fixed sample, over random permutations τ chosen equiprobably from the *n!* permutations of the elements of S.

In settings where the relative estimated bias $\delta^{(k)} \equiv \hat{B}_k(U)/\hat{t}_{y^{(k)}}$ over the whole sample is large, m_k and its estimator $\hat{m}_k$ turn out not to be much different from $\left|\delta^{(k)}\right|$. However, if $\delta^{(k)}$ is small, then $\hat{m}_k$ may be much larger, since the model-fitting may not simultaneously adjust for weighted $y_i^{(k)}$ totals over typical τ-ordered subsets of the sample.

The metric m_k depends only on the sample data. Although too complicated to evaluate exactly, it is estimated well by averaging in place of expectation over random permutations τ using the following strategy justified in Slud and Bailey (2009). With *R* denoting a moderately large number of Monte Carlo iterations, define an array $(V_{ri}, 1 \le r \le R, i \in S)$ of independent Uniform(0, 1) variates. Then a consistent Monte-Carlo estimator of (2) is given by

$$\hat{m}_k = \left(R\hat{t}_{y^{(k)}}\right)^{-1}\sum_{1\le r\le R}\max_{0\le x\le 1}\left|\sum_{i\in S: V_{ri}\le x}\left((\rho_i/\hat{p}_i)-1\right)y_i^{(k)}/\pi_i\right|. \qquad (3)$$

The quality of estimation of m_k in terms of $\hat{m}_k$, and relationships between these and $\left|\delta^{(k)}\right|$, are addressed in Section 2.3 below. We first modify (2) and (3) to allow only those random re-orderings of sample indices which preserve distinguished cells of the population, such as the cells to which population totals would be raked, or the population subdomains of particular interest to data users.

2.2 Metrics for Subdomain Bias over Distinguished Cells

Most random permutations of the sample completely shatter any meaningful sample subdomains. Yet the idea behind raking or calibration is that certain estimated subdomain totals, e.g., those over the cells of a specified geographic-demographic partition $\{A_c\}_{c=1}^{C}$ of the frame population, should be constrained equal to the cell totals (the controls) of a current (updated) census. For that reason, it makes sense to measure bias estimates $\hat{B}_k(A_c)$ over these cells, and to

assume from now on that a partition A of U into cells $A_c, c = 1,..., C$, has been fixed. Accordingly, we modify (2) so that the allowed permutations maintain the membership of elements in each cell A_c.

Now restrict the permutations τ in the expectation (2), denoting them by σ to reflect the new constraint, so that $\{\sigma(i): i \in A_c \cap S\} = A_c \cap S$, i.e., the units are re-ordered within each sample block $A_c \cap S$ but not mingled across blocks. The sample S can be indexed so that the $n_c \equiv |A_c \cap S|$ sampled elements in the c'th cell A_c are consecutively numbered from $L_{c-1}+1$ to L_c in the enumerated sample S, where $L_b = \sum_{j=1}^{b} n_j$ for each $b = 1,..., C$. The preservation of cells under σ then means that

$$\text{for all } 1 \le c \le C \text{ and } i \in A_c \cap S,\ L_{c-1} < \sigma(i) \le L_c. \tag{4}$$

The allowed random permutations σ of sample elements are chosen equiprobably from the $\Pi_{1\le c\le C} n_c!$ permutations σ of $\{1,2,...,n\}$ which satisfy (4). The modified metric is defined as the expectation over σ of the maximum absolute cumulative weighted sum of cellwise biases relative to $\hat{t}_{y^{(k)}}$:

$$m_k^* \equiv E_\sigma \left(\max_{1\le q\le n} \left| \hat{B}_k\left(\{\sigma(1),...,\sigma(q)\}\right) \right| \right) \Big/ \hat{t}_{y^{(k)}}. \tag{5}$$

A Monte Carlo estimator for the modified quantity (5) can again be implemented in terms of an array V_{ri} of independent Uniform(0, 1) random variates, according to the formula

$$\hat{m}_k^* \equiv \left(R\hat{t}_{y^{(k)}}\right)^{-1} \sum_{1\le r\le R} \max_{1\le c\le C, q\in A_c} \left| \sum_{\{i\in A_c\cap S: i\le L_{c-1} \text{ or } V_{ri}\le V_{rq}\}} \left(\frac{\rho_i}{\hat{p}_i} - 1\right)\left(y_i^{(k)}/\pi_i\right) \right|. \tag{6}$$

2.3. Confidence Intervals and Bounds for m_k and m_k^*

All of the quantities m_k, m_k^* are functions of the sampled survey data, and the probability statements made at this stage concern only the chance element introduced by the random variates V_{ri} used in defining (3) and (6), *conditionally given the sample*. In this Section, we provide a useful series of confidence bounds on the differences between the quantities m_k, m_k^* and their estimates $\hat{m}_k$, $\hat{m}_k^*$ and on the differences between these quantities and $|\delta^{(k)}|$.

First, the sampling standard error of $\hat{m}_k$ due to the random variates V_{ri} in (3) is

$$se(\hat{m}_k) = \left(\hat{t}_{y^{(k)}}\right)^{-1} \left[\frac{1}{R(R-1)} \sum_{1\le r\le R} \left(\max_{0\le x\le 1} \left| \sum_{i\in S} I_{[V_{ri}\le x]} \left(\frac{\rho_i}{\hat{p}_i} - 1\right) y_i^{(k)}/\pi_i \right| - \hat{t}_{y^{(k)}} \hat{m}_k \right)^2 \right]^{1/2}$$

and a central limit theorem applies for large *R*, so that with approximate 99% confidence

$$|m_k - \hat{m}_k| \le 2.576 \cdot se(\hat{m}_k). \tag{7}$$

Similar empirical confidence statements can be given bounding $m_k^* - \hat{m}_k^*$.

Next, it is shown in Slud and Bailey (2009) that with $\delta^{(k)} = \hat{B}_k(U)/\hat{t}_{y^{(k)}}$ and

$$\gamma^{(k)} = n^{-1}\sum_{i\in S}\left((\rho_i/\hat{p}_i)-1\right)^2\left(y_i^{(k)}/\pi_i\right),\ b_k = 1.2286\left(n\gamma^{(k)}\right)^{1/2}\Big/ t_{y^{(k)}} \tag{8}$$

defined from the sample, for large *n*

$$0 \le m_k - \left|\delta^{(k)}\right| \le \left(\hat{t}_{y^{(k)}}\right)^{-1} E_V\left(\max_{0\le x\le 1}\left|\sum_{i\in S}\left(I_{[V_{ri}\le x]} - x\right)\left((\rho_i/\hat{p}_i)-1\right)\left(y_i^{(k)}/\pi_i\right)\right|\right) \approx b_k . \tag{9}$$

The assertion (9) means that for any $\varepsilon > 0$, $m_k - \left|\delta^{(k)}\right|$ is bounded above with high probability for large *n* by $(1+\varepsilon)\cdot b_k$. A similar argument proves that $m_k^* - \left|\delta^{(k)}\right|$ is also positive and bounded above by $(1+\varepsilon)\cdot b_k$ for the same quantity b_k defined in (8).

For specific items *k*, we generally find when *n* is large that for moderate numbers *R* (say $R \ge 100$), the difference $\hat{m}_k - m_k$ (or $\hat{m}_k^* - m_k^*$) is very small compared to m_k (respectively m_k^*). Then $\hat{m}_k$ must be roughly b_k or smaller for the sample data to be compatible with a zero relative bias $\delta^{(k)}$. The objective of this analysis is to flag as 'inadequately adjusted' those items for which model-based attrition nonresponse adjustment yields estimated metric values $\hat{m}_k$ greater than $2.576\cdot se(\hat{m}_k) + b_k$. The main practical consequence is that whenever one of the three quantities m_k, m_k^*, $\left|\delta^{(k)}\right|$ is much larger than the bound in (9), then all three will be, indicating that this item *k* has been badly adjusted (for the Wave and model under consideration). When the quantities m_k, m_k^*, $\left|\delta^{(k)}\right|$ are of the same or smaller size compared with the bound b_k, the quantity m_k^* measures the quality of adjustment taking account of various population subdomains including all of the distinguished cells A_c.

3. Adjustment Metric Values in SIPP 96

In the specific example of the U.S. SIPP 1996 panel survey, we compare the estimated metric values $\hat{m}_k$, $\hat{m}_k^*$, individually or averaged over 11 cross-sectional survey items, across different adjustment models in order to choose a 'best' model. Response probabilities $\hat{p}_i$ are estimated by adjustment-cell and logistic-regression models described in detail by Slud and Bailey (2006, 2009). The survey items studied are: indicators that the individual's Household receives Food Stamps (`Foodst`), or Aid to Families with Dependent Children (`AFDC`); or indicators that the individual receives Medicaid (`Mdcd`), or Social Security (`SocSec`); and indicators that the individual has health insurance (`Heins`), is in poverty(`Pov`), is employed (`Emp`), unemployed (`UnEmp`), not in the labor force (`NILF`), married (`MAR`), or divorced (`DIV`).

In this data example, nonresponse is adjusted in one of two ways: either using a SIPP ratio-adjustment based on 149 standard cells defined in terms of variables including age, sex, race, Hispanic origin, educational level, and labor force status; or using one of a series of logistic regression models B–III summarized in Table 3-1. The variables used in these regression models are race, hispanic origin, Renter versus Owner of housing unit, indicator that individual is the Household Reference Person, indicator of college education, a 4-category variable of family type used to define raking cells for use in SIPP, U.S. Census region, an indicator of ownership of Assets, plus some or all of the 11 SIPP survey items listed above. See Slud and Bailey (2009) for additional details of model fitting. For comparison, the SIPP ratio adjustment 'model' with 149 cells yields deviance 76117 for 149 degrees of freedom.

We now assess models by flagging items with values $\hat{m}_k$, $\hat{m}_k^*$ large compared to b_k plus the right-hand side of (7). Metrics have been calculated for *R* = 100 Monte Carlo replications (ample because *n*=94444 is so large), with the results for Models B, D, and F presented in Table 3-2. The final columns of Table 3-2 display for Model B the bounds b_k (which turn out to be virtually identical within wave for the different adjustment methods) respectively

for Wave 4 and 12 nonresponse. For all items, adjustment methods, and waves, the right-hand bounds of (7) are at most 5% of the corresponding bounds b_k .

In Wave 12 in Table 3-2, always $\hat{m}_k > b_k$ under the adjustment-cell model and all of the logistic regression models, except for only a few *k*'s. One notable exception is Pov, where as in Slud and Bailey (2006), model B which includes `Pov` as a predictor does adjust effectively both in Waves 4 and 12. Similarly, Model D which includes `Pov`, `Mdcd`, `Heins`, and `UnEmp`, does well at adjusting the totals of these variables as measured by $\hat{m}_k$. Indeed, the most striking finding in Tables 3-2 and 3-4 is that including a variable as a predictor in these logistic regression models generally yields very good adjustment as measured either by metric $\hat{m}_k$ or $\hat{m}_k^*$. This is true even for Model F versus D, although Table 3-1 shows that model F actually has larger deviance than model D.

Table 3-1
Logistic regression models for Wave 4 or 12 response in SIPP 96. `Df` = number of coefficients in each model, and `Dev` the deviance for the 94444-record sample data in Wave 4. `AIC= Dev +2*Df`.

Model	Df	Variables	Dev	AIC
B	9	Wnotsp Renter College Pov Black RefPer Renter*College Black*College	76545	76563
C	14	same as B, plus Foodst Mdcd Heins UnEmp Div	76299	76327
D	14	same as B, moins Black*College plus Mdcd Heins plus UnEmp Pov*Heins Mdcd*Heins Heins*College	76242	76270
F	18	same as C, plus AFDC SocSec Emp Mar	76280	76316
I	20	College Assets hisp Region Famtyp Renter Refper Pov Mdcd Heins UnEmp Foodst SocSec Mar Renter*College	65678	65718
III	31	same as I, plus AFDC NILF Div Wnotsp Black plus 6 pairwise interactions	65638	65700

Table 3-2
Quantities $\hat{m}_k$ in (3) estimated from SIPP96 data, for wave 4 or 12 nonresponse adjustment, by either the Adjustment-Cell (A) or Logistic-Regression method (Models B, D, F) based on R = 100 replications. The last two columns are the bounds in (9), for model B.

Item	$\hat{m}^{4,A}$	$\hat{m}^{4,B}$	$\hat{m}^{4,D}$	$\hat{m}^{4,F}$	$\hat{m}^{12,A}$	$\hat{m}^{12,B}$	$\hat{m}^{12,D}$	$\hat{m}^{12,F}$	$b_{4,k}$	$b_{12,k}$
Foodst	.0052	.0186	.0076	.0039	.0442	.0130	.0110	.0093	.0056	.0123
AFDC	.0067	.0248	.0067	.0053	.1040	.0350	.0624	.0134	.0078	.0173
Mdcd	.0066	.0279	.0035	.0037	.0163	.0426	.0078	.0084	.0053	.0119
SocSec	.0191	.0116	.0125	.0027	.1118	.1068	.1066	.0073	.0041	.0086
Heins	.0085	.0065	.0013	.0012	.0197	.0133	.0027	.0028	.0019	.0040
Pov	.0187	.0033	.0032	.0032	.0372	.0091	.0074	.0085	.0047	.0097
Emp	.0016	.0017	.0015	.0014	.0082	.0122	.0161	.0034	.0020	.0041
UnEmp	.0534	.0594	.0095	.0098	.1176	.1280	.0207	.0184	.0131	.0250
NILF	.0032	.0034	.0029	.0023	.0333	.0462	.0456	.0063	.0033	.0069
MAR	.0111	.0018	.0018	.0017	.0508	.0226	.0213	.0037	.0025	.0051
DIV	.0124	.0201	.0168	.0048	.0235	.0390	.0334	.00098	.0067	.0133

Table 3-3 charts the progression of averaged $\hat{m}_k$ metrics (over $k = 1,...,11$ and Population Count) as the adjustment model varies over the adjustment-cell model and the logistic models in Table 3-1. The distinguished cells A_c used in our computations were a system of 101 cells defined by Sex, Age-intervals, and Race, on which SIPP rakes weights to population totals. The logistic regression models shown in Table 3-3 are all better than the cell-based model in adjusting at both waves 4 and 12. Since the models (except for F) are listed in order of decreasing AIC, ordering of models by AIC clearly differs from that by averaged $\hat{m}_k$. Model F seems the best adjustment model at both waves in Table 3-3, although Models I and III are strong competitors and would be preferred from examination of deviances.

The metric $\hat{m}$ rewards model F for including many of the SIPP items as predictors. Overall, neither metric m nor m^* rewards the much lower-deviance models I and III for their predictive accuracy.

Table 3-3
Metric values $\hat{m}_k$ calculated on SIPP 96 data for adjustment-cell and logistic regression models and averaged over $k = 1,...,12$, where item 12 is Population Count.

Model	AdjCell	B	C	D	F	I	III
$\hat{m}^4$	.0123	.0150	.0043	.0057	.0034	.0039	.0039
$\hat{m}^{12}$	.0474	.0389	.0248	.0281	.0078	.0137	.0083

Table 3-4
Estimated metric values $\hat{m}_k^*$ defined in (6) for Wave 4 adjustment (Wave 12 only in last row) with R = 100, based on adjustment cell and logistic regression models, using SIPP 96 data with partition A_c into 101 demographic cells. Metric $\hat{m}_k^*$ values in last 2 rows are averaged over items k.

Item	ModC	ModD	ModF	ModI	ModIII	AdjCell
AFDC	.0065	.0075	.0064	.0063	.0059	.0072
SocSec	.0115	.0127	.0038	.0048	.0045	.0192
Heins	.0021	.0022	.0020	.0027	.0026	.0086
Pov	.0050	.0052	.0051	.0049	.0049	.0191
UnEmp	.0097	.0092	.0098	.0103	.0102	.0535
MAR	.0025	.0025	.0025	.0038	.0035	.0112
DIV	.0057	.0170	.0056	.0055	.0054	.0129
Wav4.Avg	.0051	.0065	.0044	.0046	.0045	.0127
Wav12.Avg	.0273	.0304	.0110	.0163	.0113	.0492

Scrutiny of Table 3-4 for individual survey items, several of which are not shown, shows that the parametric adjustment models F, I, and III accomplish something that the adjustment cell model cannot: they generate response probabilities with good behavior over raking cells considered as subdomains. The logistic regression models, especially F and III, do better, item by item, with far fewer parameters than the 149 adjustment-cell response fractions by which the adjustment cell model divides the survey weights in the respective adjustment cells. Model F seems the overall best choice, with III a close second.

4. Models and Metrics using raked SIPP Weights

In practice, within government surveys such as SIPP, weights are adjusted and then raked so that population totals over certain demographic cells match the totals found through other, more accurate, censuses and surveys. For this reason, we recalculated the metrics for the cell-based adjustment model and the models displayed in Table 3-1 based on adjusted weights which were put through a final stage of raking, as discussed in Slud and Bailey (2009), based on cells defined in terms of sex, age, race, family structure and Hispanic origin. For the two metrics, averaged over survey items k (11 items plus population count), we provide the comparison between model-based adjustments with and without raking in Table 4-1. According to the metrics m_k and m_k^*, with the best models (F, I, III) it hardly matters whether raking is done or not, but there is no real benefit and may even be some loss in adjustment accuracy, especially for Models F and III.

Table 4-1

Estimated metric scores $\hat{m}_k$ and $\hat{m}_k^*$ from SIPP 96 data, averaged over survey items k, for Unraked and Raked adjusted weights from Logistic-regression (D,F,I,III) and adjustment-cell models.

Metric	Raked	Wave	ModD	ModF	ModI	ModIII	AdjCell
m_k	Yes	4	.0057	.0034	.0039	.0039	.0123
		12	.0281	.0078	.0137	.0088	.0474
	No	4	.0047	.0057	.0046	.0045	.0083
		12	.0163	.0120	.0166	.0113	.0181
m_k^*	Yes	4	.0065	.0044	.0046	.0045	.0127
		12	.0304	.0110	.0163	.0113	.0493
	No	4	.0046	.0056	.0046	.0045	.0082
		12	.0164	.0121	.0165	.0113	.0182

5. Conclusions

This paper has developed metrics for nonresponse-adjustment effectiveness, calculated by randomly reindexing the survey sample and calculating maximum discrepancies over consecutively indexed subdomains. One objective was to discount any advantage which an adjustment model might achieve toward eliminating whole-sample nonresponse biases by including survey attributes as predictors. However, when applied to SIPP 96 data, the metrics developed tended to favor models which overfit the data, and those which includedseveral of the survey items as predictors. While such an adjustment strategy could not be tried if the selected set of interesting survey attributes were too large, incorporating the survey attributes as nonresponse predictors seems to be a good idea in the SIPP setting with 11 attributes.

The best nonresponse models in the SIPP 1996 panel were found to perform almost equally well whether their adjusted weights are raked or not, and thus — from the vantage point of our metrics — there may not be much value in extensive raking when a highly effective nonresponse adjustment model is used.

References

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. and Särndal, C.-E. (2001). A better understanding of weight transformation through a measure of change, *Survey Methodology*, 27, pp. 97-108.

Eltinge, J. and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells, with an application to income nonresponse in the US Consumer Expenditure Survey, *Survey Methodology*, 23, pp. 33-40.

Kim, J.-K. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability, *Canadian Jour. of Statist.*, 35, pp. 501-514.

Oh, H. and Scheuren, F. (1983). Weighting adjustment for unit nonresponse, pp. 143-184 in *Incomplete Data in Sample Surveys*, vol. 2, eds. W. Madow, I. Olkin and D. Rubin. New York: Acad. Press.

Rizzo, L., Kalton, G., Brick, M. and Petroni, R. (1994). *Adjusting for panel nonresponse in the Survey of Income and Program Participation, ASA Surv. Res. Meth. Proc. paper, JSM 1994*.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer: New York.

Slud, E. and Bailey, L. (2006). Estimation of attrition biases in SIPP, *ASA Surv. Res. Meth. Proceedings paper, JSM 2006*, Seattle, WA.

Slud, E. and Bailey, L. (2009). Evaluation and selection of models for attrition nonresponse adjustment, *Survey Methodology*, to appear.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Nonresponse Weight Adjustments Using Multiple Imputation for the UK Millennium Cohort Study

John W. McDonald and Sosthenes C. Ketende[1]

Abstract

This paper discusses nonresponse weight adjustments for sweep 3 of the UK Millennium Cohort Study (MCS). Weight adjustments are available for monotone patterns of nonresponse, where the nonresponse weight is the inverse of the estimated probability of response based on a logistic regression model, which uses data from previous sweeps to predict response at the current sweep. For non-monotone patterns, some cases have missing data for previous sweeps and this approach cannot be easily applied. For MCS, 7.5% of the families took part in sweeps 1 and 3, but not sweep 2, i.e., a non-monotonic pattern of nonresponse for 1,444 families.

Our approach to estimate a nonresponse weight for MCS sweep 3 was to use multiple imputation to impute the required missing values at sweep 2 for these 1,444 families for the logistic model for response at sweep 3. This imputation used information from sweeps 1 and 3 and only involved imputing the missing values for time-varying variables shown to be predictive of nonresponse in MCS. This resulted in the multiple imputation of nonresponse weights at sweep 3, which can be averaged to produce a single nonresponse weight or the 10 imputed nonresponse weights can used for separate analyses and the results combined using Rubin's rules. We discuss the advantages and disadvantages of both approaches.

Key Words: Imputation, Longitudinal Survey, Wave Nonresponse, Weighting Adjustment.

1. Introduction

1.1 Wave nonresponse in longitudinal surveys

A common problem in longitudinal studies such as panel and birth cohort studies is wave nonresponse, which occurs when responses are obtained for some, but not all, waves of the study. Little and David (1983) discuss three types of wave nonresponse: attrition, reentry and late entry. Attrition occurs when a unit drops out of the study at one wave and remains out thereafter; reentry occurs when a unit drops out for one or more waves, but reenters the study at a later point; and late entry occurs when a unit does not participate in the first wave, but enters the study later. At wave 3 of the UK Millennium Cohort Study (MCS), the longitudinal pattern of response is complex, with attrition, reentry and late entry patterns of response. At wave 3, 2,210 families attrited at wave 1 and 1,664 attrited at wave 2; 692 families entered the MCS late at wave 2 and of these "new families" 124 attrited at wave 2; and 1,444 families reentered the MCS by participating at wave 1, not at wave 2, but reentered the MCS at wave 3.

The pattern of longitudinal participation may be monotone, where no participant returns to the study after missing a wave, or non-monotone, where some subjects return to the study after missing a wave. This paper discusses unit nonresponse weight adjustments for wave (sweep) 3 of the MCS. Weight adjustments are available for monotone patterns of wave nonresponse, where the nonresponse weight is the inverse of the estimated probability of response at the current wave based on a logistic regression model, which uses data from previous waves to predict response at the current wave. For non-monotone patterns, some cases have missing data for previous waves and this approach cannot be easily applied. For the MCS, the longitudinal pattern of response is complex and non-monotonic with families that enter late at wave 2 and reenter at wave 3 after not participating at wave 2. Our approach to estimate nonresponse weights for MCS wave 3 was to use multiple imputation to impute the required missing values at wave 2 for the logistic model for response at wave 3. This imputation used information from waves 1 and 3 and only

[1]John W. McDonald, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (John.McDonald@ioe.ac.uk); Sosthenes C. Ketende, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (S.Ketende@ioe.ac.uk)

involved imputing the missing values for time-varying variables shown to be predictive of nonresponse in MCS. This resulted in the multiple imputation of nonresponse weights, which can be averaged to produce a single nonresponse weight or the 10 imputed nonresponse weights can used for separate analyses and the results combined using Rubin's rules. In this paper, we discuss the advantages and disadvantages of both approaches.

1.2 Compensating for wave nonresponse in panel surveys

Kalton (1986) and Lepkowski (1989) review methods for compensating for wave nonresponse in panel surveys. Both discuss using weighting or imputation and the reasons why as a rule unit nonresponse is treated by weighting and item nonresponse is treated by imputation. Which strategy is used often depends upon the auxiliary information available for use in making the nonresponse adjustments. Weighting is often favoured when the auxiliary variables available are only weakly related to the variables with missing values, e.g. when only survey design information is available. Imputation is often favoured when the auxiliary variables available are strongly related to the variables with missing values. In the case of longitudinal surveys, two general imputation approaches are possible: cross-sectional imputations and cross-wave imputations. Cross-sectional imputations are imputations based on data from a single wave. Cross-wave imputations are imputations based on data from multiple waves. For example, for a two-wave panel with repeated variables, one would use the value of a variable on one wave to impute the missing value of the same variable on the other wave; if the missing value was for the second wave, this would be forecasting and if the missing value was for the first wave, this would be backcasting. Imputing for wave nonresponse for variables which are repeated at each wave may be the best solution if the responses for the repeated variable are highly correlated over time. In general, missing values may be forecast using values of the variable from previous waves or backcast using values of the variable from subsequent waves or `interpolated' if the missing wave is surrounded by adjacent non-missing waves. Of course, cross-wave imputation schemes would also typically use other variables as well as the cross-wave repeated variables in the imputation model.

Since both weighting and imputation each have their strengths (advantages) and weaknesses (disadvantages), both Kalton (1986) and Lepkowski (1989) consider that a combination of weighting and imputation methods may be the best solution. This combination strategy was adopted in this paper to estimate wave nonresponse weight adjustments. As our desired logistic regression model for response at wave 3 included predictor variables with missing data, we used cross-wave multiple imputation to deal with the item missing data on some of the predictor variables. This resulted in the multiple imputation of nonresponse weights at wave 3.

1.3 Nonresponse weight adjustments for the UK Millennium Cohort Study

The UK Millennium Cohort Study population is children born between 1 September 2000 and 31 August 2001 (for England and Wales), and between 24 November 2000 and 11 January 2002 (for Scotland and Northern Ireland), alive and living in the UK at age 9 months, and eligible to receive Child Benefit at that age; and then, after 9 months, for as long as they remain in the UK at the time of sampling. Previous Great Britain/United Kingdom birth cohort studies in 1946, 1958 and 1970 sampled all births in one week. In contrast, the MCS sampled births spread across a calendar year and used a complex sample design. The MCS was a stratified one-stage cluster sample of electoral wards. The UK population was stratified by country - England, Wales, Scotland and Northern Ireland. Using a ward-level measure of child poverty and ethnicity, wards in England were stratified into three strata labelled advantaged, disadvantaged and ethnic and wards in Wales, Scotland and Northern Ireland were stratified into two strata labelled advantaged and disadvantaged. A different proportion was sampled from each of the 9 strata, so that the smaller UK countries are over-represented relative to England, and families living in more disadvantaged areas are over-represented and in England, families living in areas with high proportions of ethnic minority families are also over-represented. Once the 398 wards were selected, a list of all 9 month old children living in these wards was required. The lists were generated by using government Child Benefit records. Child Benefit is a universal provision, payable from the child's date of birth, and Child Benefit claims cover virtually all of the UK population, except those who do not have a `right to reside' in the UK. All families, who were eligible to receive child benefit and whose child was living in the sampled ward at age 9 months, were invited to participate in the MCS via a letter from the government department which administered the Child Benefit records. The response rate for MCS wave 1 was 72%. For further details on the MCS sample design and implementation, see Plewis (2007a).

Families moved in and out of the selected sample wards. Families can continue to be paid Child Benefit without notifying the government of a change of address, especially if the benefit was paid directly into a bank account. As a

result, the list of children residing in the selected wards was out of date and some eligible children were missed as their existence in the selected ward was not known during the wave 1 fieldwork. The wave 2 sample was supplemented by 692 "new families" eligible at wave 1, but missed because their addresses were not up to date. The cohort child in new families entered the MCS late at age 3.

The productive sample at wave 1, i.e., with interview data from at least the main respondent or partner, formed the issued sample at wave 2, except for cohort members known to have died or emigrated. At wave 2, wave 1 refusals were not issued to the fieldwork agency. The response rate for wave 2 of the MCS was 79%. The correlates of unit nonresponse in the first two waves of MCS are described in Plewis (2007b) and Plewis et al. (2008). Plewis (2007b) found for MCS waves 1 and 2 that differences in response probabilities were small compared to the unequal selection probabilities in the sample design.

Plewis (2007b) describes the methodology used to generate the MCS wave 2 nonresponse weights. A set of potential predictors of nonresponse at wave 2 was studied using a logistic regression model. They included a range of socio-demographic and socio-economic explanatory variables along with a measure of residential mobility derived from the address database, which is part of the administrative side of the MCS survey operation (see Plewis et al. (2008) for more details of this measure). The new families who entered MCS late were excluded from this modelling and their nonresponse weights at wave 2 were defined to be one. All the explanatory variables (other than residential mobility) were measured at wave 1. Because of the issuing strategy, i.e. refusers were not issued, there was no missing data in the wave 1 variables used in the modelling. The predicted probabilities of responding based on this logistic regression model were inverted to generate the nonresponse weights at wave 2. The overall weights at wave 2 are the product of the wave 1 overall weights and the wave 2 nonresponse weights.

For wave 3, all families who were not interviewed at wave 2 were re-issued to the fieldwork agency, except for permanent refusals and families where the cohort member was known to have died or emigrated. Many families not interviewed at wave 2 were productive at wave 3. This yielded a non-monotonic pattern of longitudinal response with 1,444 families that were productive in waves 1 and 3, but not wave 2. As these 1,444 cases have missing data for wave 2 and the estimation of nonresponse weights using the logistic regression approach cannot be easily applied unless all the explanatory variables in the model were measured at wave 1. While some of the explanatory variables Plewis (2007b) found to be predictive of response from wave 1 to wave 2 were time constant such as the ethnic group of the cohort member, others were time varying such as family income, housing tenure (own, rent, other) and type of accommodation (house/bungalow, other). We would expect that the more recent values at wave 2 would be more predictive of response at wave 2, than the values at wave 1, and the values of these time varying explanatory variables are missing at wave 2 for the 1,444 families that were productive in waves 1 and 3, but not wave 2. However, one might think that this is the ideal situation for using imputation for the wave 2 missing values as we have the values of these variables at adjacent waves.

2. Imputation of missing wave 2 data

Multiple imputation was used to replace the missing wave 2 data with "plausible" values. Note that we are not imputing the entire wave, but only the missing values for time-varying variables shown to be predictive of nonresponse in MCS by Plewis (2007b). Multiple imputation is used to take account of the uncertainty in the imputation process. We impute multiple times, in our case 10 times, to create 10 datasets, where the missing values have been filled in. We use each of the 10 datasets to fit our logistic model for response at wave 3 using the same set of explanatory variables. Each dataset yields, typically different, estimates of the nonresponse weights for wave 3 of the MCS. So we end up with 10 estimates of the nonresponse weights. Various issues arise: 1) how should the imputations be carried out, 2) what are the assumptions of the imputation procedure and whether these assumptions justified and 3) how does the analyst deal with multiple estimates of the nonresponse weights.

2.1 Multivariate imputation by chained equations

Multivariate imputation by chained equations (MICE) is the name of the procedure for imputing missing multivariate data by fully conditional specification (van Buuren, 2007). The basic idea is to impute the missing values of variables on a variable-by-variable basis by specifying one imputation model per variable. For each missing variable,

a univariate conditional distribution for the missing variable, given other variables, can be specified, i.e., the imputation model is specified separately for each variable, involving the other variables as predictor variables. The method consists of iterating over these conditional distributions by means of Gibbs sampling. At each stage of the algorithm, an imputation is generated for the missing variable and this imputed value is used in the imputation of the next variable. Each iteration cycles through all the variables with missing data. This process is repeated until the process reaches convergence. For details, see van Buuren (2007). MICE has been implemented in a number of software packages including Stata, R and SPSS (Horton and Kleinman 2007, van Buuren and Groothuis-Oudshoorn forthcoming). We used Patrick Royston's implementation of MICE using his Stata ado files (Royston, 2004; Royston, 2005a; Royston, 2005b; Royston, 2007).

In theory, multiple imputation works, but in practice using MICE with Stata 10 was problematic as most of our variables were categorical variables and we had a large dataset. Having interactions in an imputation model meant we had to explicitly create a dummy variable for each category of categorical variables with more than two categories. The dummy variables were imputed when missing and then the original categorical variable was passively imputed from its imputed dummy variables. Issues we had to deal with in practice included predictors, which perfectly predicted binary variables so that we had to fix errors interatively by selecting a reference category other than the first or collapsing categories. Dealing with interactions in imputation models takes care (von Hippel 2009), in our case mainly because of sparse tables. For reasons of space we will not further discuss the details of the imputation models used, other than to say that we used cross-wave imputations and tried to include the sampling design in our imputation models by including the stratum variable as one of our predictors (Reiter et al., 2006). We were not able to account for clustering in our imputation models, but our logistic model for response included the stratum variables as predictors and took account of the clustered nature of the MCS sample. Note that Stata 11 has a new mi (multiple imputation) command, which resolves many of the above issues we had using Stata 10.

Our imputation model makes the assumption that the unobserved values are MAR (missing at random), i.e., given the observed data, the missingness mechanism does not depend on the unobserved data. In the context of a longitudinal study with a monotone pattern of nonresponse, the assumption that the missing longitudinal survey data are MAR is usually invoked when estimating nonresponse weights, so that the probability of response at wave t is assumed to be a function of variables measured at previous waves, but not on unobserved variables measuring changes between wave t-1 and t. Plewis et al. (2008) shows that residential mobility, i.e., any change of postal address, after wave 1 was an important predictor of nonresponse at wave 2 in the MCS, which casts doubt on this common assumption. Our nonresponse model for MCS wave 3 did include residential mobility between wave 2 and wave 3 as a predictor variable.

3. Handling multiply imputed nonresponse weights

3.1 Pooling the results using Rubin's rules

Multiple imputation fills in each missing value with a set of M plausible values to generate M completed datasets, in our case M = 10 datasets. Each of these 10 datasets were used to estimate 10 nonresponse weights for MCS wave 3, which were used in the standard way to produce cross-sectional or longitudinal weights. These weights can used to produce a weighted estimate of some quantity of interest, say a proportion, a mean or the regression coefficient in a logistic regression model. The results may be combined or pooled, using what have been termed Rubin's rules (Rubin, 1987), to give estimates and standard errors that take into account the uncertainty due to the imputed missing data values in the nonresponse model used to estimate the nonresponse weights.

3.2 Alternatives

One alternative is to use the mean of the M estimated nonresponse weights for analysis. Use of the mean weight has the following advantages: 1) it is simpler to use the mean weight with a single dataset rather than having use M datasets and then combine the results using Rubin's rules and 2) it is simpler to deposit one dataset in a data archive than M datasets. Use of the mean weight has the following disadvantages: 1) it does not take into account all the uncertainty in the results due to the incomplete data, 2) it is more complicated to deal with M datasets, both for the user, data provider and data archivist, and 3) users need to understand multiple imputation and be trained to use relevant software which deals with M datasets and pools the results.

Rather than invert the M estimated response probabilities from the logistic model for response at wave 3 and average, an alternative is to average the M estimated response probabilities and then invert. Note that the harmonic mean is always less than or equal to the arithmetic mean, with equality only when all numbers are equal. This implies that for each case, the mean weight is always greater than or equal to the inverse of the mean probabilities. As these two quantities are equal only when all M estimated response probabilities are equal. These two quantities will vary with the variability of the M estimated response probabilities. It is not clear which measure is to be preferred on statistical grounds.

3.3 Is it necessary to use multiply imputed nonresponse weights?

Is it necessary to use multiply imputed nonresponse weights with M datasets and pool the results? What are the differences if we used a single dataset and the mean of the weights or the inverse of the mean of the estimated response probabilities to weight each observation? We investigated these issues by estimating a mean, a proportion and a logistic regression coefficient using all three alternatives. At MCS wave 3, we estimated the mean weight for boys in kilograms, the proportion of boys overweight or obese and the regression coefficient for residential mobility in a logistic model for response at wave 3. Figures 3.3-1 to 3.3-3 present from left to right, a pooled estimate using the 10 datasets and Rubin's rule for pooling the results, estimates based on each of the 10 datasets, denoted IM1, IM2, …, IM10, an estimate using the mean weight and an estimate using the inverse of the mean of the estimated response probabilities. For the limited cases studied, the estimate using Rubin's rules and the mean weight gave very similar estimates, with the lowest estimate in all cases being for the inverse of the mean of the predicted response probabilities.

Figure 3.3-1
Estimated mean weight for boys in kilograms at Millennium Cohort Study wave 3

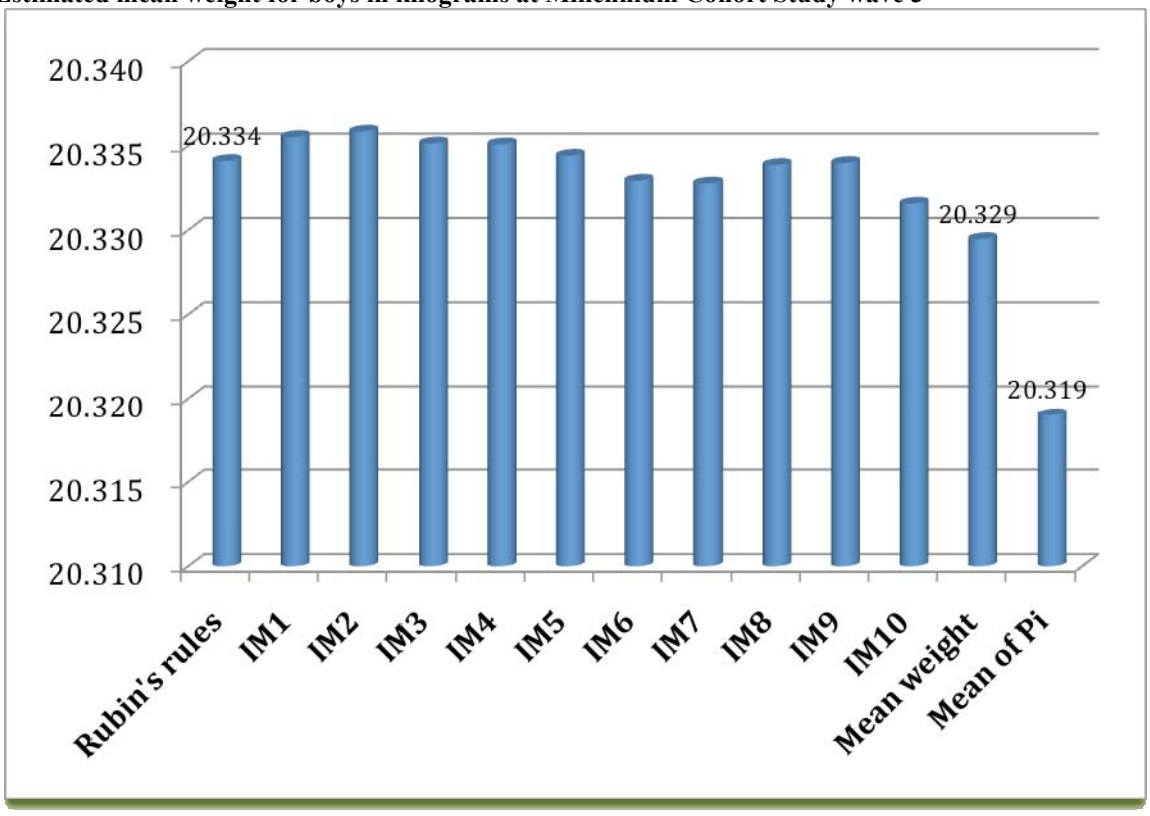


Figure 3.3-2
Estimated proportion overweight or obese at Millennium Cohort Study wave 3

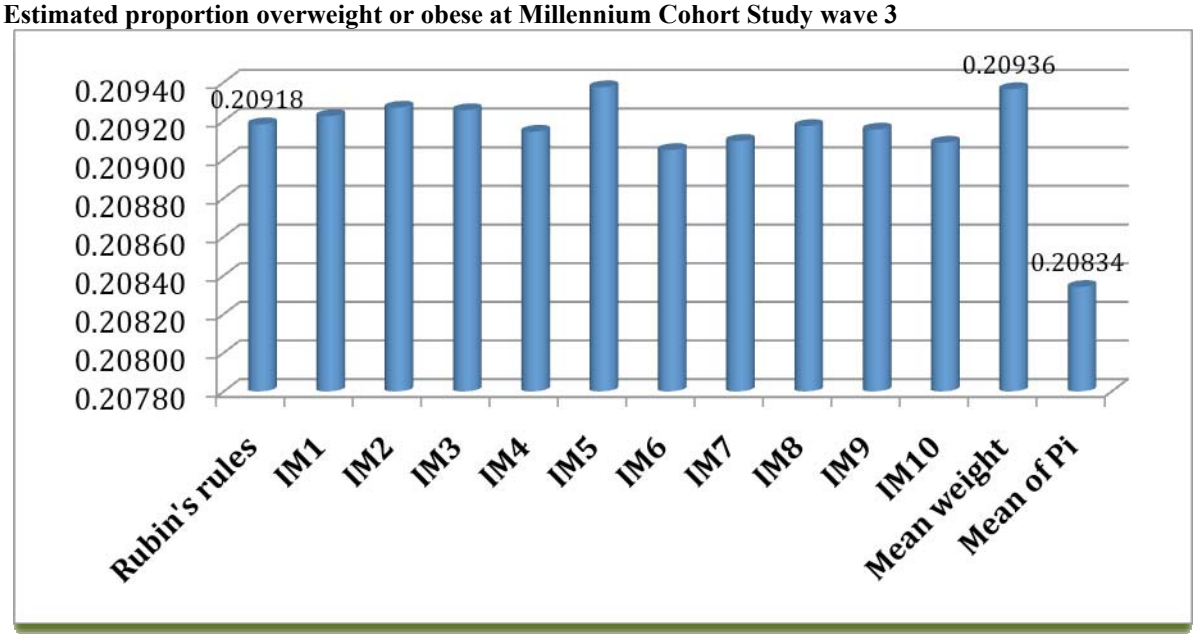
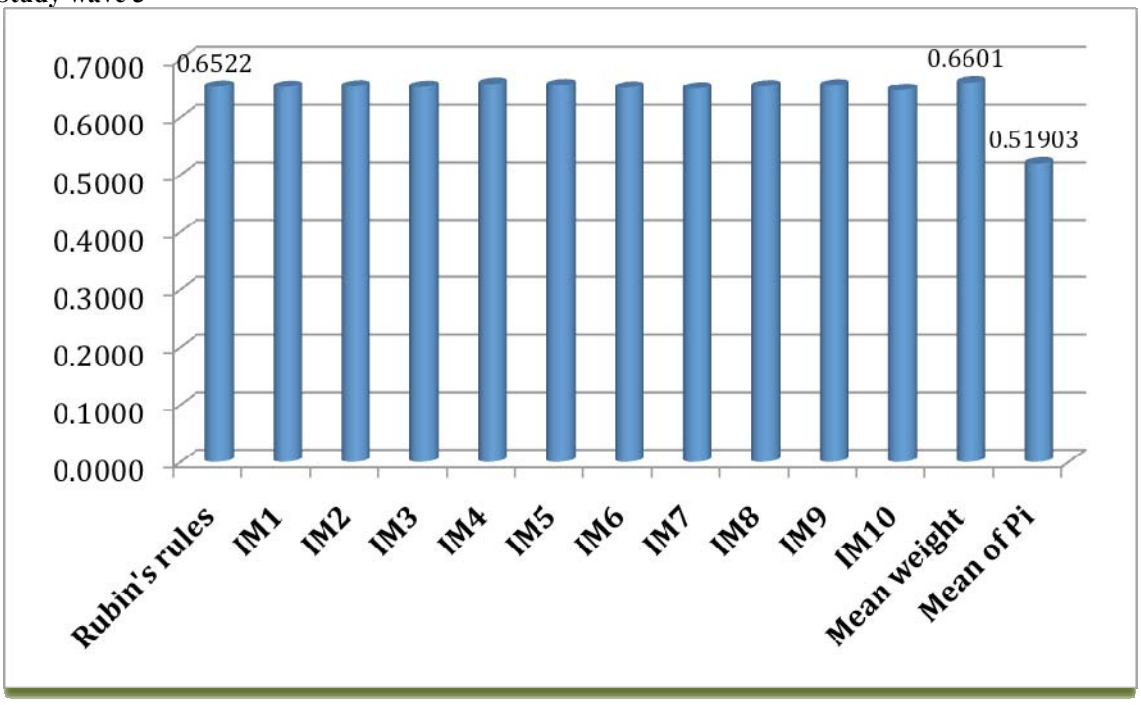


Figure 3.3-3
Estimated regression coefficient for residential mobility in logistic model for response at Millennium Cohort Study wave 3



Acknowledgements

The authors would like thank G. Kalton for helpful references and J. N. K. Rao for comments.

References

Horton, N. J. and Kleinman K.P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Regression Models, *The American Statistician,* 61, pp. 79-90.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys, *Journal of Official Statistics*, 2, pp. 303-314.

Lepkowski, J. M. (1989). Treatment of Wave Nonresponse in Panel Surveys, in D. Kasprzyk et al. (eds*.) Panel Surveys*, New York: Wiley, pp. 348-374.

Little, R. J. A. and David M. H. (1983). Weighting Adjustments for Non-response in Panel Surveys, Working paper, U. S. Bureau of the Census, Washington, D. C.

Plewis, I. (ed.) (2007a). Millennium Cohort Study First Survey: Technical Report on Sampling (4th ed.) , London: Centre for Longitudinal Studies.

Plewis, I. (2007b). Nonresponse in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology,* 10, pp. 325-334.

Plewis, I., Ketende, S.C., Joshi, H. and Hughes G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study", *Journal of Official Statistics* 24, pp. 365-385.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. (2006). The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data, *Survey Methodology*, 32, pp. 143-150.

Royston, P. (2004). Multiple Imputation of Missing Values, *Stata Journal*, 4, pp. 227-241.

Royston, P. (2005a) Multiple Imputation of Missing Values: update, *Stata Journal*, 5, pp. 188-201.

Royston, P. (2005b) Multiple Imputation of Missing Values: Update of ice, *Stata Journal*, 5, pp. 527-536.

Royston, P. (2007). Multiple Imputation of Missing Values: Further Update of ice, With an Emphasis on Interval Censoring, *Stata Journal*, 7, pp. 445-464.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

van Buuren S., Boshuizen, H. C. and Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis, *Statistics in Medicine*, 18, pp. 681-694.

van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification, *Statistical Methods in Medical Research*, 16, pp. 219–242.

van Buuren, S. and Groothuis-Oudshoorn K. (forthcoming). MICE: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*.

von Hippel, P. T. (2009). How to Impute Interactions, Squares, and Other Transformed Variables, *Sociological Methodology*, 39, pp. 265–291.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Non-response Bias in the National Longitudinal Survey of Children and Youth

Mike Tam and Agnes Waye[1]

Abstract

In this paper, we examine the non-response of the original cohort of the National Longitudinal Survey of Children and Youth (NLSCY). The NLSCY is a longitudinal survey that collects information on characteristics and factors that may affect the development and well-being of Canadian children and youth over time. An original cohort of children aged between 0 and 11 years old was sampled in 1994, with follow-ups occurring every two years. In view of the amount of non-response that has occurred over time, an extensive analysis was initiated to determine the extent of non-response bias that may be present and how such bias may be affecting analyses based on NLSCY data. To address these issues, we identify potential determinants of non-response.

Key Words: Longitudinal Survey, Non-Response Analysis, Logistic Regression.

1. Introduction

The purpose of this paper is to discuss some of the results of a recent evaluation of the original cohort of the National Longitudinal Survey of Children and Youth (NLSCY) with respect to non-response bias. As of 2007, original cohort members have been surveyed for seven cycles. During this time, response rates of those who responded at Cycle 1 have steadily declined, reaching 64.9% at Cycle 7 (see Table 3.1-1). A non-response bias assessment done as part of the ten-year methodological review of the survey at Cycle 6 indicated that there was evidence of bias in some estimates of the original cohort (Statistics Canada, 2007). The current paper aims to provide information on additional NLSCY variables that may be affected by the non-response that has occurred between Cycles 1 and 7. In the following section, an overview of the NLSCY is provided. The overview is followed by a discussion on the response rates and evolution of the sample of the original cohort over time. Then, some of the results of various assessments that we conducted will be presented. Finally, a brief conclusion encapsulating our current view of the non-response situation is given.

2. Overview of the NLSCY

The NLSCY is a longitudinal survey of children and youth that is sponsored by Human Resources and Skills Development Canada and conducted by Statistics Canada. At Statistics Canada, the survey is managed by Special Surveys Division with methodological services provided by Household Survey Methods Division. Some of the principal objectives of the survey include: determining the prevalence of risk factors, investigating how these factors affect development, and making the information available for policy-making. The NLSCY is also intended to be a general survey that can be used to support other types of analyses. Data for the NLSCY is made available at Statistics Canada's Research Data Centres located across the country. Under certain rules, university and other researchers may then access the data for their research.

The survey commenced in the fall of 1994 with a sample of 0-11 year olds living in one of the ten provinces in Canada. This sample is referred to, in the NLSCY, as the *original cohort*. Data for the survey was subsequently collected every other year, with each biennial collection period denoted as a *cycle*. Starting at Cycle 2, an early childhood development (ECD) cohort has been added to the NLSCY at each cycle. The ECD is a cohort of 0 and 1 year olds that is intended to be followed longitudinally for at least three additional cycles. The NLSCY is also used

[1]Mike Tam, Statistics Canada, R.H. Coats Building, 17-L, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (mike.tam@statcan.gc.ca); Agnes Waye, Statistics Canada, R.H. Coats Building, 17-I, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (agnes.waye@statcan.gc.ca)

for cross-sectional analysis. Thus, in addition to the longitudinal cohorts, a top-up sample of 2 to 5 year olds has also been drawn at each cycle to improve the representativeness of the cross-sectional target population. As of summer 2009, eight cycles of collection had been completed with the original 0 to 11 year olds reaching 14 to 25 years old. However, in this paper, we focus on the non-response of the original cohort at Cycle 7, which was completed in summer 2007.

In the survey, although the unit of study is principally the child (or children) in the selected household, the respondent is the adult who is the *person most knowledgeable* (PMK) about the target child when the child has not yet reached 18 years old. However, once the child has attained the age of 18, he or she becomes his or her *own respondent*.[2] The use of the term *respondent* within both contexts are used interchangeably in the remainder of the paper.

3. Analysis of non-response in the NLSCY

3.1 Response rates

In this section, we examine the evolution of the sample of original cohort members over seven cycles. Table 3.1-1 is comprised of the response rates of the original cohort members over Cycles 2 to 7 of the survey, relative to Cycle 1. At the first cycle of the NLSCY, there were actually 22,831 respondents. However, as a result of sample cuts due to budgetary reasons, eventually only 16,890 Cycle 1 respondents were followed longitudinally.

The rates have been divided into three age groups: 0 to 11, 0 to 5, and 6 to 11. The purpose of this delineation is to separately observe the response rates of the members of the original cohort where the interviewed member continues to be the parent and the members who are now their own respondents. Specifically, the 0 to 5 year olds who were 12 to 17 at Cycle 7 and so the interviewed member was the parent and the 6 to 11 year olds who were 18 to 23 at Cycle 7 and hence were their own respondents.

The rates are based on non-funnel respondents, that is, previous cycle non-respondents are included in the calculation if they subsequently respond. The cumulative response rates for funnel respondents are far lower. And though the response rates for both groups, 0-5 and 6-11, are low, the 0-5 group has suffered less non-response by Cycle 7. As shown in the table, the response rates have declined by substantial amounts.

Table 3.1-1
Unweighted Response rates of Cycles 2 to 7 relative to Cycle 1, by Cycle 1 age group

	0-11		0-5		6-11	
Cycle	Respondents	%	Respondents	%	Respondents	%
1	16,890	100.0	9,553	100.0	7,337	100.0
2	15,384	91.1	8,726	91.3	6,658	90.7
3	14,770	87.4	8,394	87.9	6,376	86.9
4	13,169	78.0	7,544	79.0	5,625	76.7
5	12,273	72.7	7,153	74.9	5,120	69.8
6	11,178	66.2	6,564	68.7	4,614	62.9
7	10,966	64.9	6,282	65.8	4,684	63.8

Since we are focusing on the non-response after Cycle 1, note that the decrease in response rates is also dependent on the number of individuals that we decide to return to the field. The number fielded at each cycle has decreased along with the number responding. This has been due to a combination of holding back members who we consider to be hard refusals or have had too many cycles of non-response. Of course, intensity of refusal or excessive consecutive cycles of non-response, upon which eligibility to be returned to the field is based, can themselves be related to individual or household characteristics. From the table, we also note that there was a small increase in the response rates for the 6-11 group between Cycles 6 and 7. This was a result of bringing back many members who had not been sent to the field for several cycles. Overall, 665 sample members who became their own respondents (i.e., 18 years

[2] In the event that a 16 or 17 year old youth is living independently, he or she would constitute the PMK.

old or more) were brought based on the Cycle 7 rule of not considering parental response history, and of these, 341 responded.

We also considered the response rates by various Cycle 1 variables to determine how the sample composition may be changing over time. The types of variables examined included demographic variables such as effective age of child[3], province of residence, and dwelling type; labour force variables such as socio-economic status, ratio of household income to the low income cut-offs (LICO), and current working status of the PMK and spouse; and social and health variables such as participation in an early childhood program and social support. Table 3.1-2 shows the response rates for three of the variables: effective age at Cycle 7 (ages 18-23 only), sex of the child, and the Cycle 1 social support score.

Table 3.1-2
Unweighted response rates by selected variables

Variable	Cycle					
	2	3	4	5	6	7
Effective Age at Cycle 7						
18	91.2	87.6	77.3	71.2	65.9	65.2
19	89.3	85.9	77.3	70.4	66.3	63.3
20	91.8	86.8	76.2	72.8	65.1	66.4
21	90.1	87.3	76.3	70.8	63.2	65.0
22	90.8	85.8	76.7	66.4	58.7	62.2
23	91.8	88.0	76.3	66.8	57.5	60.8
Sex of the child						
Male	91.0	87.0	77.5	71.9	64.5	63.2
Female	91.1	87.9	78.5	73.4	67.9	66.7
Social Support Score						
Low	84.7	86.4	71.1	64.3	56.2	54.5
Medium	89.9	85.5	75.0	69.7	62.6	61.3
High	92.7	89.3	80.7	75.4	69.5	68.2

The slight increase in response rates between Cycles 6 and 7 for the 20 to 23 year olds is due mainly to the bringing back of non-respondents who had, by Cycle 7, become their own respondents. The response rates between Cycles 6 and 7 decreased only slightly for 20-23 year olds at Cycle 7 who were also in the field at Cycle 6. By contrast, the response rates for 18-19 year olds at Cycle 7 who were also in the field at Cycle 6 decreased signficantly more, resulting in a slight decrease in response rates for *all* the 18-19 year olds at Cycle 7.

For sex of the child, there is a fairly large difference between males and females. Though not shown here, the same evaluation was performed, focusing on the part of the cohort who began the NLSCY at ages 0 to 5. In that age group, there is not much difference in the response rates by sex since there is a high degree of homogeneity of the sex of the parents who do the responding, namely that they are mostly female.[4] In the group who began the NLSCY at ages 6 to 11, the response rates are higher for females when the respondent is the youth himself or herself.

The third variable shown in Table 3.2-1 is the Cycle 1 social support score. The variable is based on a factor score of variables that measure perceptions of social support such as whether assistance and feelings of security are available from family and friends; whether someone is available to provide advice; and whether the member is part of a group who shares similar interests and concerns. Originally, the social support score variable is reported on a scale from 0 to 18, but here, it is separated into three categories: low, medium, and high.[5] The response rates for the low category has decreased at the most rapid rate, but there are relatively few respondents who actually fall into this category, partly accounting for the apparent anomalous and sudden increase at Cycle 3. However, between the categories

[3] In the NLSCY, the *effective age,* defined as the reference year minus the year of birth, is used in place of the actual age at the time of the interview. This is to ensure that members stay in the age groups to which they were assigned, regardless of whether collection takes place before or after their birthday.
[4] At Cycle 1, 93% of the responding PMKs were female.
[5] Low = 0 to 8, Medium = 9 to 13, and High = 14 to 18

medium and high, greater social support appears to portend greater response. This aspect is discussed a bit further later on in the paper.

Based on the response rates of some variables, there are indications of differential non-response with respect to some variables. Age and sex of the child are variables that would usually be taken into account at the non-response adjustment and post-stratification stages, so residual bias in those variables would be expected to be minimal. However, in the case of a variable like social support, it is not as clear. So, we continue with taking a closer look at different sets of Cycle 1 variables to acquire a sense of where additional potential bias might lie.

3.2 Comparison of respondents and non-respondents at Cycle 7

One method of evaluating potential non-response bias is by comparing the characteristics of respondents and non-respondents. A sample is divided into respondents and non-respondents and estimates of various parameters, such as averages for numerical variables or proportions for the levels of categorical variables, are computed separately for each group. See, for example, Fitzgerald et al. (1998). The estimates for the respondents and the corresponding estimates for the non-respondents are then evaluated for significant differences. We focus on the attrition since Cycle 1 of the NLSCY using the data from the Cycle 1 master file. By concentrating on attrition from Cycle 1, we have data available for both respondents and non-respondents at Cycle 7. For the purposes of this study, the non-respondents include refusals, non-contacts, and moves. They do not include deaths. To evaluate the statistical significance of the differences, we computed the standard errors using the Rao-Wu re-scaled bootstrap (Rao and Wu, 1988; Girard, 2007).

We first focus on basic socio-demographic variables such as: province, age of the parent, number of children in the household, parent's highest level of education, household income, tenure, and sex of the child. These are variables similar to those found in other non-response studies (Fitzgerald et al., 1998; Watson, 2003; Watson and Wooden, 2004). Because the group of 6-11 year olds have become their own respondents, we also considered separate analyses for the 0-11, 0-5, and 6-11 year olds. As can be seen in Table 3.2-1, there appear to be many significant differences between the respondents and non-respondents. Based on the PMK's age and level of education, the number of children in the household, housing tenure, and the sex of the child, the non-respondents are more likely to be young, have lower levels of education, have more children in the household, and rent their home. Because the table describes ages 0 to 11, the differences by sex of the child are statistically significant, as they also are for ages 6 to 11. By contrast, differences by sex of the child are not significant for ages 0 to 5 since it is the PMK's sex that is the more important factor.

Some other socio-demographic characteristics with significant differences were province of residence, where non-respondents were more likely to be from Ontario and less likely to be from Quebec; immigration status, where non-respondents were more likely to be immigrants; and parental status, where non-respondents were less likely to live with two parents. Some of the characteristics where there were no significant differences were household size and age of the child. Generally, the non-respondents appeared to have fewer members in the household, but none of the differences were significant. On the other hand, when restricting to the 0-5 group, there were significantly more respondents with household size equal to 3; but restricting to the 6-11 group, neither variable was significant. So, the difference in significance between number of children in the household and household size may be partly explained by the different impact of the variables on the two age groups. For the age of the child, since the specific survey questions asked are dependent on the age of the child, one conclusion that we could draw from the absence of differences is that different questions/components are not related to attrition.

When comparing respondents and non-respondents, but restricting the cohort to those who were 0 to 5 years old at Cycle 1, we obtain very similar results as for the entire cohort, with the exception of sex of the child, where the difference is no longer statistically significant. At this point, the PMK is still the individual who determines whether or not the interview will be granted. Finally, we consider comparisons amongst the 6-11 group only. Obviously, with these older children, there are less PMKs who are aged between 15 and 29. The age group 30-34 now has significantly more non-respondents, in contrast to the cohort of 0 to 5 year olds. In addition, the number of children in the household no longer has any effect, but the sex of child does, with respondents more likely to be females.

Table 3.2-1
Comparison of respondents and non-respondents (0-11 year olds)

Characteristic	Respondents	Non-respondents	Difference
Age of PMK			
15-24	4.4	6.4	***-2.1***
25-29	15.1	18.8	***-3.8***
30-34	31.9	32.4	-0.5
35-39	29.5	27.1	2.4
40+	19.1	15.1	***4.0***
Number of children in household			
1	21.1	16.8	***4.3***
2-4	76.1	79.4	***-3.3***
5+	2.8	3.7	-0.9
Highest level of education			
Less than secondary	14.0	21.2	***-7.2***
Secondary	16.8	21.4	***-4.6***
Beyond high school	29.3	26.5	***2.8***
College or university	39.6	30.4	***9.2***
Household income			
Less than 19,000	4.7	7.3	***-2.6***
20,000 - 39,999	20.3	22.2	-1.9
40,000 - 59,999	27.7	23.7	***4.0***
60,000 or more	34.5	26.0	***8.5***
Tenure			
Owned	75.3	63.5	***11.9***
Not owned	24.7	36.5	***-11.9***
Sex of child			
Male	50.3	53.0	***-2.7***
Female	49.7	47.0	***2.7***

Note: **Bold-faced** type indicates significance at 5%. Estimates were derived using the Cycle 1 survey weights.

3.3 Logistic analysis of non-response

3.3.1 Analysis using base variables

We also implemented a multivariate analysis of the same variables from Section 3.2 to determine the significance of specific characteristics when controlling for other variables. We constructed logistic regression models with the response status at Cycle 7 as the outcome variable and with the variables just discussed as the covariates. This was done for each age group: 0-11, 0-5, and 6-11. Table 3.3.1-1 provides a summary of the regression coefficient estimates for a subset of the variables. The first category for each variable is the reference category. For the 0-11 and 0-5 year olds, the PMKs who were older at Cycle 1 were more likely to respond relative to the younger PMKs, whereas for the 6-11 group, the age of PMK is no longer significantly related to response since the PMK no longer answers for the child. A larger household size as represented by the number of children in the household tend to decrease the probability to respond. However, there is no significant effect for five or more children in the household for the 6-11 group, possibly because the children are older. For education and tenure, we again see some confirmation of our previous results, where higher education and ownership of one's residence appear to be related to a greater likelihood to respond. In the case of the sex of the child, when he or she is the potential respondent is an important factor since we see that females tending to be more likely to respond relative to males.

Table 3.3.1-1
Parameter estimates of the base set of Cycle 1 variables

Characteristic	0-11	0-5	6-11
Age of PMK			
15-24	0.00	0.00	0.00
25-29	0.08	-0.03	-0.30
30-34	***0.26***	***0.31***	-0.35
35-39	***0.35***	***0.45***	-0.25
40+	***0.52***	***0.70***	-0.10
Number of children in household			
1	0.00	0.00	0.00
2-4	***-0.41***	***-0.49***	***-0.28***
5+	***-0.49***	***-0.63***	-0.34
Highest level of education			
Less than secondary	0.00	0.00	0.00
Secondary	0.03	0.19	-0.10
Beyond high school	***0.38***	***0.52***	0.25
College or university	***0.43***	***0.60***	***0.26***
Tenure			
Owned	0.00	0.00	0.00
Not owned	***-0.25***	***-0.22***	***-0.28***
Sex of child			
Male	0.00	0.00	0.00
Female	***0.12***	0.04	***0.19***

Note: **Bold-faced** type indicates significance at 5%. Estimates were derived using the Cycle 1 survey weights.

3.3.2 Analysis using variables from the Starting Out study

In the foregoing sections, basic demographic variables were assessed, the results of which were intended to provide a general idea of where potential non-response bias may lie. However, as diverse analyses are often done using NLSCY data, it is difficult to assert whether the variables in any particular analysis might have been affected by NLSCY non-response. It would therefore be useful to assess the variables in substantive analyses of the NLSCY to ascertain whether the current status of those variables currently have the potential for non-response bias. We considered several studies of the NLSCY such as McIntyre (1996) and Bohatyretz and Lipps (1999). In this paper, we discuss only results of the variables from McIntyre (1996).

The Starting Out study (McIntyre, 1996) was a study included in the *Growing Up in Canada* report. The investigation focused on the health of infants and toddlers in Canada by examining conditions that contribute to a healthy pregnancy and to the health of a child around the time of birth. For example, there are variables related to the mother's health during the pregnancy, the child's social behaviour, and the family environment. Each of the study's variables applies to a specific age range from 0 to 3 years old. In this section, we examine the variables that apply to each of the three age groups: 0 to 3, 0 to 1, and 2 to 3. We are interested in investigating the relationships between some of the variables from the study for members of the original cohort on their response status in cycle 7.

To begin, we considered the variables used to provide a demographic profile of the 0-3 year olds in the study. Some of the variables included were: sex of the child, parent status, household income, age of the mom at birth of the child. Since these variables were similar to our base set, we added three variables to look at further: age of the mom at the birth of the child, aboriginal status, and social support. We found that higher ages of the mom at the birth of the child were less likely to attrite. This is consistent with older PMKs, who were mostly moms, being less likely to attrite. PMKs with greater social support were also less likely to attrite and aboriginal status turned out not to be related to attrition. Not surprisingly, including the control variables, in particular age of the PMK, resulted in age of the mom to not be significant. However, social support remained significant.

The remainder of this section is a summary of some of the model results. A more detailed discussion can be found in Tam and Waye (2010). We first focused on the study of the health of the mother during pregnancy for 0-1 year olds. The main variables of interest were: smoking during pregnancy, breastfeeding, and post-partum depression. These were entered into a logistic model of response behaviour at Cycle 7. We found that mothers who smoked during pregnancy are more likely to attrite, mothers who breastfed their children were less likely to attrite, and that post-partum depression was not related to attrition. However, once we expanded the model to control for socio-demographic variables, the key variables were no longer significant. Other studies have shown that smoking during pregnancy was linked to lower levels of education (Millar and Hill, 2004) and that breastfeeding initiation was linked to both income and educational level (Heck et al., 2006). So, these are possible factors in causing the effects to fall to insignificant levels.

In another part of the study, social and emotional behaviours are considered. These variables included: hyperactivity, physical aggression, emotional disorder, positive interaction, and consistent parenting. Focusing only on these variables in a logistic analysis, we found that PMKs whose parenting style was positive interaction were less likely to attrite later on and those whose style was punitive were more likely to attrite. However, again, once the specification was expanded, the variables were no longer significant.

3.3.3 Analysis using of the oldest age group (22-23 year olds)

In this section, a logistic regression model was built to predict the probability of non-response at cycle 7 using survey data from cycle 5 for members of the original cohort who were 18 to 19 at cycle 5 and had responded. The rationale behind choosing this subset of the population is that cycle 5 is the first cycle at which part of the original cohort reaches the age of 18. We hypothesized that response behaviour at cycle 7 would be dependent on the child's data and not the PMK's data when the child was 18 or older. Therefore, we generated a new non-response model using the data available at cycle 5 to see if the individual characteristics of the child would be useful in predicting subsequent non-response at cycle 7. In addition, this model allowed us to use more recent data for our assessment.

Variables that were related to the child and not the PMK were chosen for this model. Some of the variables included in the model related to the child's psychological well being, health, and social skills. Additional variables that were included in the model were social support, depression, self image, parental conflict, smoking, club involvement, and education. However, our investigations showed that, in a model that included the base control variables, only club activities turned out to be significant in the logistic model. A possible explanation for this result is that club involvement indicated a higher degree of social interaction, which generally is related to a greater propensity to respond.[6]

4. Conclusion

In this paper, we describe some aspects of a non-response study that was conducted. Univariate analysis of an extensive selection of variables indicates that differential non-response exists, although this can often be explained by the relationship between the study variable and the socio-demographic variables. Multivariate analysis of socio-demographic and subject matter characteristics indicate potential non-response bias at varying degrees. Parameter estimates that are significant in logistic models of response when entered alone are often no longer significant when controlling for socio-demographic variables. Thus, it is clear that bias is dependent on the study at hand. If inferences do not need to be made on the portion of the sample which appears to be biased due to attrition, or if appropriate control variables are included in the analysis, bias may not be an issue. In the non-response study upon which this paper is based, additional analyses such as an investigation on whether attrition plays a role in some transitions in the NLSCY and an examination of possible differences between initial non-respondents and subsequent non-respondents are conducted.

[6] The significance of the variable may also be the result of Type I error. The actual *p*-value for the significant category was 0.007.

Acknowledgements

We would like to express our sincere thanks to Sarah Franklin and Joanne Moloney for their support in initiating this project. We would also like to convey our deep appreciation to Dany Faucher for his immense help and support throughout. Finally, we are indebted to Yves Lafortune and Michel Ferland for their thoughtful and insightful comments, which have vastly improved the paper.

References

Bohatyretz, S. and Lipps, G. (1999). Diversity in the Classroom: Characteristics of Elementary Students Receiving Special Education, *Education Quarterly Review,* 6(2), 7-19.

Fizgerald, J., Gottschalk, P. and Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics, *Journal of Human Resources*, 33(2), 251-298.

Girard, C. (2007). How to Avoid Getting All Tied Up Bootstrapping a Survey: A Walk-Through Featuring the National Survey of Children and Youth, working paper, Ottawa: Statistics Canada.

Heck, K., Braveman, P., Cubbin, C., Chavez, G. and Kiely, J. (2006). Socioeconomic Status and Breastfeeding Initiation among California Mothers. *Association of Schools of Public Health*, 121 (1), 51-59.

McIntyre, L. (1996). Starting Out, In *Growing Up in Canada*, Ottawa: Statistics Canada, 47-56.

Millar, W. and Hill G. (2004). Pregnancy and Smoking. *Health Reports*, 15(4), 53-56.

Rao, J.N.K. and Wu, C.F.J. (1988). Re-sampling Inference with Complex Survey Data, *Journal of the American Statistical Association*, 83(401), 231-241

Statistics Canada. (2007). The National Longitudinal Survey of Children: 10-Year Review of the Methodology, internal report, Ottawa: Statistics Canada.

Tam, M. and Waye A. (2010). Non-Response Analysis of the National Longitudinal Survey of Children and Youth, internal report, Ottawa: Statistics Canada.

Watson, D. (2003). Sample Attrition Between Waves 1 and 5 in the European Community Household Panel, *European Sociological Review*, 19(4), 361-378.

Watson, N. and Wooden, M. (2004). Sample Attrition in the HILDA Survey, *Australian Journal of Labour Economics*, 7(2), 293-308.

DATA COLLECTION AND LINKAGE

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

An Experimental Test of a Strategy to Maintain Contact With Families between Waves of a Panel Study

Katherine McGonagle, Mick Couper, and Robert Schoeni[1]

Abstract

The Panel Study of Income Dynamics (PSID) is a nationally representative longitudinal survey of approximately 9,000 families and their descendants that has been ongoing since 1968. Families are routinely sent a mailing asking them to update or verify their contact information to keep track of their whereabouts between waves. Analysis shows that having updated information prior to data collection is cost effective, resulting in less tracking, special refusal conversion efforts, fewer contacts to be interviewed, and lower attrition. This paper reports findings from an experiment designed to maximize response to a request for updated contact information with study families.

1. Introduction

1.1 Overview

Keeping track of sample persons between waves of data collection helps minimize attrition in longitudinal studies. All things being equal, the longer the time between data collection waves, the greater the likelihood that sample persons have moved, and the greater the difficulty in locating movers (Couper & Ofstedal, 2007; Duncan & Kalton, 1987). In 1997, the Panel Study of Income Dynamics (PSID) changed from annual to biennial interviewing. To capture residential changes between waves, PSID began sending families a "contact information update" mailing in the year between data collection waves. Families who update or verify the address and telephone information and return it receive a $10 post-paid check. About half the families respond to this mailing in recent waves. During 2007, families providing this information needed far less tracking or refusal conversion efforts, and half as many contacts to be interviewed, underscoring the cost effectiveness of the mailing.

Given these advantages, a study was designed before 2009 production interviewing with the goal of improving the response rate of the contact update mailing. Families were randomly assigned to the following conditions: $10 as a pre- versus post-paid incentive, mailing design (traditional versus updated), being sent a study newsletter, and timing and frequency of the mailing (July versus October versus both times). This paper reports findings with regards to response rates to the mailing by these different conditions. Overall, there is no effect of incentive type implying that post-paid incentives are more cost effective. Traditional design performs better than updated design. Families receiving a second mailing have higher response rates than those receiving one mailing. There are some interaction effects by timing-of-mailing, with October-only families having higher response rates when also receiving a pre-paid incentive and newsletter. Hypotheses for these findings and next steps for analysis are described.

1.2 Rationale for experimental conditions

Various types of contact strategies to improve panel retention have been used by ongoing panel studies (see Couper & Ofstedal, 2007; Laurie et al., 1999). However, we know of no experimental tests of particular strategies. While the issue of contact strategy effectiveness is an understudied one, the survey methods literature can be drawn upon for suggestions of ways to improve response rates. First, under some conditions a pre-paid incentive increases cooperation by heightening the salience of the incentive, as well as the respondent's sense of reciprocity (Singer et al., 1999). Moreover, an updated design of respondent contact materials should increase the salience of the request which may enhance cooperation (Dillman, 2000).

[1] Katherine McGonagle, Mick Couper, and Robert Schoeni, Survey Research Center, Institute for Social Research, University of Michigan.

Further, the timing and frequency of the request may affect cooperation. Ideally, the request occurs close enough to the upcoming data collection that most residential changes are captured, but not so close that perceptions of respondent burden are increased, which could occur with too many contacts. PSID interviewing occurs in odd years between March and November. Thus, two times of year were chosen for the mailing: midway between the end of the prior wave and beginning of the next (i.e., July) and as far into the year as was feasible to update addresses before production began (i.e., October). A third condition, July with an October re-mail for non-responders, was chosen to examine the effect on response rates of two contacts versus one.

Finally, evidence on the ideal amount of respondent contact is scant. In the two-year timeline of a biennial survey, respondents participate in a lengthy interview, receive a study newsletter, are asked to update contact information, and receive a letter alerting them to the upcoming interview. At what point do these multiple contacts become burdensome, or do they in fact enhance perceptions of identification with the survey? It is likely that these perceptions vary by characteristics of sample members. Thus, in this study the manipulation of the mailing involved modifying aspects of the incentive, the design, the timing, and the amount of contact being made with guidance from evidence in the survey methods literature.

2. Methods

PSID families eligible for the 2009 interview (n=8,512) were randomly assigned to four conditions which defined a 2 ("newsletter") x 3 ("timing") x 2 ("design") x 2 ("incentive") experimental design (Table 2-1). To manipulate number of respondent contacts and burden, half the families were sent a study newsletter a year before interviewing began. The second condition was the timing of the mailings with one-third mailed in July, one-third in October, and one-third mailed initially in July with a follow-up mailing in October for nonresponders. Mail design was the third condition, with half the families receiving the traditional black and white design, and half an updated design (Appendix I). The final condition varied whether the $10 incentive was pre-paid or post-paid.

Table 2-1
Return rates by experimental condition

Newsletter	Timing	Design	Incentive	Sample size	Return Rate (%)
Yes	July	Traditional	Pre-paid	330	57.9
Yes	July	Traditional	Post-paid	371	58.8
Yes	July	Updated	Pre-paid	422	53.8
Yes	July	Updated	Post-paid	339	50.7
Yes	July with October follow-up	Traditional	Pre-paid	383	65.0
Yes	July with October follow-up	Traditional	Post-paid	329	64.1
Yes	July with October follow-up	Updated	Pre-paid	298	65.4
Yes	July with October follow-up	Updated	Post-paid	331	68.0
Yes	October	Traditional	Pre-paid	395	60.8
Yes	October	Traditional	Post-paid	334	58.4
Yes	October	Updated	Pre-paid	330	57.9
Yes	October	Updated	Post-paid	394	50.0
Total				**4256**	
No	July	Traditional	Pre-paid	366	61.5
No	July	Traditional	Post-paid	346	63.8
No	July	Updated	Pre-paid	351	55.6
No	July	Updated	Post-paid	324	54.0
No	July with October follow-up	Traditional	Pre-paid	329	65.1
No	July with October follow-up	Traditional	Post-paid	386	65.8
No	July with October follow-up	Updated	Pre-paid	313	67.1
No	July with October follow-up	Updated	Post-paid	376	67.8
No	October	Traditional	Pre-paid	325	59.4
No	October	Traditional	Post-paid	401	54.4
No	October	Updated	Pre-paid	369	52.6
No	October	Updated	Post-paid	370	47.0
Total				**4256**	

3. Results

3.1 Overall treatment effects

Examination of return rates during 2009 showed that 60 percent of families provided updated or verified contact information. This did not vary by whether the family was sent a newsletter or by incentive type. There was a significant effect of timing with a higher response rate for the July-October follow-up condition (66 percent) compared to the July-only (57 percent) and October-only (55 percent) conditions. Finally, families in the traditional design condition had a significantly higher response rate than those in the updated design (61 percent versus 57 percent).

Traditional design performed significantly better regardless of the newsletter or incentive type. There were no differential effects of incentive type by newsletter or design of the mailing.

Logistic regression models examined effects of newsletter, incentive, and design within each timing condition. Among July-only cases, traditional design performed significantly better than updated design and there was no effect of incentive type or newsletter. Among October-only cases, a significantly higher response rate was associated with traditional design, pre-paid incentive, and being mailed the newsletter. Finally, there were no significant effects of newsletter, incentive type, or timing among July-October follow-up cases.

3.2 Treatment effects by key demographic variables

A central goal of the current experiment is to design a respondent contact strategy that maximizes the odds that contact information for all study families will be up-to-date prior to production interviewing, with the ultimate goal of minimizing panel attrition. Initial analyses were conducted to examine whether the overall treatment effects had differential impacts across key demographic statuses known to be associated with panel attrition, including mover status and family income. Respondents in families that moved or had strong intentions to move had lower response rates to the mailing than non-movers (56 percent compared to 66 percent). As in the overall sample, both groups were more likely to respond to the traditional design compared to the updated design, and to the July-October mailing compared to either of the one-time mailings. Interestingly, non-movers who received pre-paid incentives were significantly more responsive to the mailing than those receiving post-paid incentives (68 percent compared to 65 percent, respectively), while incentive type did not affect the response rate of movers. Analyses examining the effects of the experiment by highest versus lowest income quartile confirmed the overall patterns reported in the total sample, with no effects of incentive type or receiving a newsletter for either group, and increased response rates in the traditional design and July-October mailing conditions for both groups.

4. Discussion

What lessons can be learned from this experiment? First, a follow-up mailing for nonresponders is an effective, low cost strategy that may ultimately reduce the need for expensive tracking during production. This condition yielded response rates between 9 and 11 percentage points higher than the one-time mailing conditions.

Second, and unexpectedly, traditional design performed better. In both designs, the last known contact information was preprinted on a card that respondents folded over, sealed with an attached sticker, and mailed. The updated design also included lengthy instructions describing the necessity of tearing away the card before folding it. Perhaps the additional instructions made returning the updated card seem complicated and discouraged compliance. We will test this in 2010 by removing the "tear step" and instructions and modifying only the color and design.

Third, the pre-paid incentive performed better than post-paid only in the October condition. Perhaps it was comparatively appealing by October as the U.S economy began to unwind. However, overall pre-paid incentives were not cost effective as they did not increase response rates compared to post-paid. The finding that families who were sent the newsletter had higher response rates in October than those who were not may reflect the importance of maintaining contact with study families. October-only families who were not sent the newsletter had no study contact for at least 10 months, with most having no contact for 16 months or longer. These families had a four percent lower response rate than families who received the newsletter. Taken together with the finding of highest response rates for the mailing condition with follow-up suggests that the most promising timing is an initial contact approximately eight months before production starts, with a follow-up three months later if needed.

The next stage of analysis will examine the impact of the various treatment conditions on field effort and production outcomes during 2009 PSID production interviewing, including tracking and refusal conversion rates, contact attempts needed to obtain a final result, and attrition. In order to understand whether certain subgroups are more sensitive to aspects of the treatment protocol than others, perhaps leading to a tailored approach toward developing contact strategies, multivariate analyses of treatment effects by various socio-demographic conditions will be conducted given the strong intercorrelation of characteristics such as income, mover status, and age. This information will help us design an effective strategy for keeping track of panel families.

References

Couper, M.P. and Ofstedal, M.B. (2009). Keeping in Contact with Mobile Sample Members, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, New York: Wiley, pp. 183-203.

Dillman, D.A. (1987). *Mail and Internet Survey: The Tailored Design Method*, New York: Wiley.

Duncan, G.J. and Kalton, G. (2000). Issues of Design and Analysis of Surveys across Time, *International Statistical Review*, 55, pp. 97-117.

Laurie, H., Smith, R. and Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey, *Journal of Official Statistics*, 2, pp. 269-282.

Singer, E.S., Gebler, N., Raghunathan, T., Van Hoewyk, J. and McGonagle, K. (1999). The Effect of Incentives on Response Rates in Face-to-Face, Telephone, and Mixed Mode Surveys: Results of a Meta-Analysis, *Journal of Official Statistics*, 15, pp. 217-230.

APPENDIX I: DESIGN CONDITIONS

Traditional Design

Please accept this $10.00 check as a token of our appreciation!

If your name, address and phone number on the label below are correct, please check the "No Changes" box.

OR

If your name, phone and/or address on the label have changed, please make the necessary changes on the form below.

Then fold the postcard in half and seal the open edge shut with the sticker on the right.

Thank you for your help!

1.1 No Changes.

1.2 ***(Please check this box if the information below is correct.)***

Affix label here

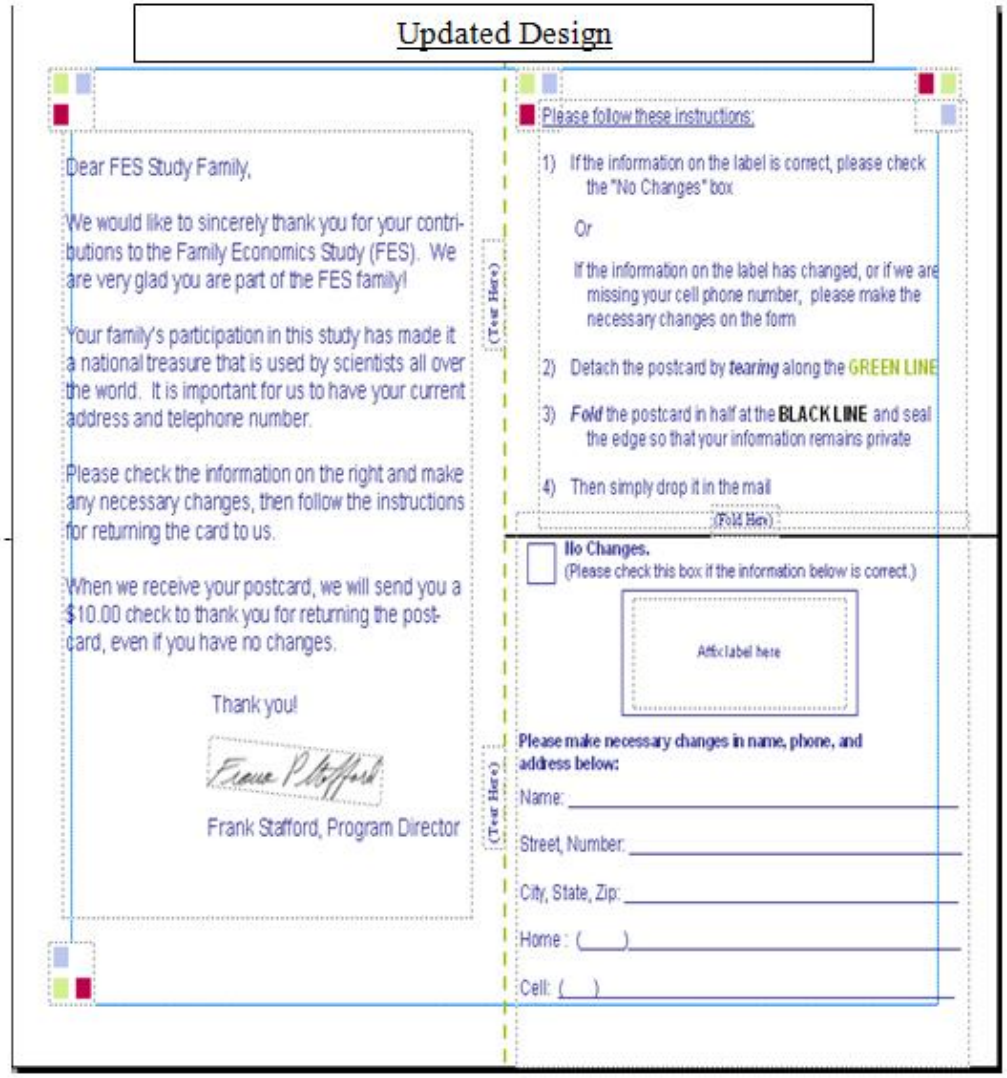
Please make necessary changes in name, phone, and address below:

1.2.1.1 Name: _____

1.2.1.2 Street, Number: _____

1.2.1.3 City, State, Zip: _____

Updated Design



Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Mixed and Multiple Collection Modes: The HILDA Survey Experience

Nicole Watson and Mark Wooden[1]

Abstract

This paper uses the experience of the Household, Income and Labour Dynamics in Australia (HILDA) Survey to examine whether the use of different data collection modes interact to produce unintended consequences. Specifically, the paper addresses four key questions. First, what respondent characteristics best predict choice of survey mode? Second, does choice of mode have any effects on response, especially in later waves? Third, does mixed and multiple modes affect item non-response? Finally, is there any evidence that the answers provided by respondents differ systematically with mode?

Key Words: Mixed Mode, Multiple Modes, Non-Response, Data Quality, Longitudinal Surveys.

1. Introduction

It is generally accepted that mixed mode designs in survey research can improve response rates (e.g., Day et al., 1995; Voogt & Saris, 2005). It has also been well established that choice of survey mode can influence respondent answers (e.g., De Leeuw, 1992). Far less, however, is known about the impact of mixing survey modes in longitudinal surveys. Dillman (2009) draws largely on the experience from cross-section surveys to speculate about the impact changing survey modes over time might have, but otherwise published research into this issue is rare.

This paper uses the experience of the Household, Income and Labour Dynamics in Australia (HILDA) Survey to examine the possibility that in mixed and multiple mode surveys the different modes might interact to produce unintended consequences. Specifically, the paper addresses four key questions. First, what respondent characteristics best predict choice of survey mode? Second, does choice of mode have any effects, either favourable or harmful, on response, especially in later waves? Third, does mixed and multiple modes affect the quantity of data collected? Finally, is there any evidence that the answers provided by respondents differ systematically with mode?

2. The HILDA Survey

The HILDA Survey is a household panel survey with a focus on employment, income and the family. Modelled on household panel surveys undertaken in other countries, and described in more detail in Wooden and Watson (2007), it began in 2001 with a large national probability sample of Australian households occupying private dwellings. All members of those responding households in wave 1 form the basis of the panel to be pursued in each subsequent wave (though interviews are only conducted with those household members aged 15 years or older), with each wave of interviewing being approximately one year apart. Like many other households panels, the sample is extended each year to include new household members resulting from changes in the composition of the original households. Like other household panel surveys, the HILDA Survey collects data via multiple modes. More specifically, the HILDA Survey employs both *multiple modes*, in that data are gathered from all respondents via a personal interview as well as a self-administered paper questionnaire, and *mixed modes*, in that respondents and interviewers have some flexibility to choose between face-to-face and telephone methods when delivering the interview component.

Interview lengths vary, both across individuals and across survey waves, but in general the aim is to ensure that the average time spent by interviewers in a two-adult household does not exceed 83 minutes. A self-completion questionnaire (SCQ) is provided to all persons who complete a personal interview, and is typically collected by the interviewer at a later date, or failing that, interview respondents are asked to return the completed form by mail. The SCQ consists mainly of questions that

[1] Nicole Watson, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (n.watson@unimelb.edu.au); Mark Wooden, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (m.wooden@unimelb.edu.au).

are difficult to administer in a time effective manner in a personal interview or that respondents may feel slightly uncomfortable answering in a face-to-face interview.

The design of the HILDA Survey also allowed for the possibility of interviews being conducted by telephone rather than in person. In wave 1 the telephone was mainly used as a method of last resort, and hence the incidence was low (just 0.5% of all interviews were conducted on the telephone). In subsequent waves, greater use of telephone interviewing was necessary given the cost of visiting sample members that move to locations beyond the 'usual' reach of our interviewer network. Furthermore, while not explicitly offered as an option, telephone interviews were also conducted with sample members indicating a strong preference for this mode of survey delivery. The incidence of telephone interviews has thus risen over the course of the panel, and by wave 8 stood at 10.1% (see Table 5-1).

3. Who is interviewed by telephone?

We first examine whether the types of respondents interviewed by telephone differ systematically from those interviewed face-to-face. Since our outcome is binary, we estimated probit models distinguishing between these two modes that used data pooled across waves 2 to 7 and allowed for correlation across observations from the same individual. (Data from wave 1 is excluded, both because the design of the study precluded many telephone interviews being conducted in wave 1, and because we include characteristics measured at the previous wave in our list of controls.) We estimated two models, one using the observations from previous wave respondents and the other all remaining observations (though for reasons of brevity only report and comment here on the former). Dynamics are allowed for through the inclusion of a variable identifying interview mode at the previous wave. A summary of results on coefficients of interest is presented in Table 3-1.

We hypothesised that interview by telephone will be more likely, relative to the face-to-face option, among: (i) respondents that changed address, especially if they moved into rural Australia; (ii) respondents whose working hours arrangements would make them harder to find at home; (iii) persons with disabilities and illness conditions; and (iv) persons who exhibited a greater reluctance to be located and respond. Strong evidence in support of all four hypotheses is found. All other things constant, the probability of a mover being interviewed by telephone next wave is more than twice that of a non-mover (8.9% vs 4.3%) greater than that of a non-mover. Further, independent of moving, the probability of a telephone interview increases the more distant the respondent is from a major city. Also as expected, we found that busy working people are significantly more likely to be interviewed by telephone, though what was unexpected was the relatively high probability of telephone interviews among the unemployed – their predicted probability of telephone interviews is about the same as an employee who works a 36 hour work week. Telephone interviewing, however, is not related to irregular work schedules. We also found that persons with long-term health conditions classified as 'serious' (that is, precluded the respondent from undertaking any work) are much more likely (about 1.4 times more likely) to be interviewed by telephone. With respect to hypothesis (iv), interviewer assessments of whether a respondent had been 'cooperative' or was suspicious about the aims of the survey were both found to be significant predictors of choice of mode at the next survey wave. We also found that non-return of the SCQ in one wave, which we argue is an indicator of the extent of respondent engagement with, and interest in, the survey, is a significant predictor of being interviewed by telephone the next.

Also of note, interviewer workloads were found to exhibit a u-shaped relationship with the probability of a telephone interview. Finally, it can be seen that after controlling for a wide range of factors, we still find clear evidence that the incidence of telephone interviewing has been rising over time, suggesting that there are other factors at work, such as changing preferences on the part of both respondents and interviewers.

Table 3-1
Probit model predicting telephone interview outcome among previous wave respondents, HILDA Survey waves 2 to 7 – Selected coefficients and summary statistics

Explanatory variable	Coeff.	Robust se	Explanatory variable	Coeff.	Robust se
Previous wave interview situation			Respondent characteristics		
Telephone interview	1.202**	.034	Moved	.461**	.022
Non-return of SCQ	.233**	.032	Age	-.052**	.009
Cooperative	-.134*	.068	Unemployed	.109*	.051
Suspicious	.196**	.057	Usual weekly work hrs (x 10^{-2})	.301**	.001
Understanding	.110*	.053	Irregular hours	-.012	.027
Interviewer workload (x 10^{-2})	-2.395**	.091	Severe LT health condition	.216**	.074
Ivwr workload squared (x 10^{-4})	.917**	.038	Location (base = Sydney)		
Wave dummies (base = wave 2)			Melbourne	-.273**	.038
Wave 3	.173**	.036	Brisbane	-.195**	.045
Wave 4	.240**	.034	Adelaide	-.264**	.055
Wave 5	.212**	.034	Perth	-.285**	.055
Wave 6	.258**	.034	Other major city	-.143*	.044
Wave 7	.436**	.033	Inner regional	.119**	.032
			Outer regional	.286**	.037
			Remote	.737**	.054
Pseudo R-squared	.229				
N	Observations = 69548; Individuals = 15520				

* p <.05; ** p < .01. Also included, but not reported, were controls for: gender, country of birth, Aboriginality, marital status, household size, education, equivalised household income, renter status, and neighbourhood socio-economic status.

4. Interview mode and panel attrition

As previously noted, it has been well established that providing sample members with alternative modes of survey delivery typically results in improved response rates. What is not clear is whether such findings extend to improved attrition rates within a longitudinal setting. As we have previously noted (Watson & Wooden, 2009, p. 164), the evidence from longitudinal surveys mostly suggests that telephone interviewing, when used in what are primarily face-to-face surveys, is predictive of higher (rather than lower) attrition at the next wave. Such findings suggest that use of alternative survey modes can only delay the decision to cease cooperating. However, it is also true that, in most of these surveys, telephone interviewing was being used primarily as a method of last resort. Responding by telephone was, therefore, itself an indicator of a lack of interest, or unwillingness to participate, in the survey.

The HILDA Survey, on the other hand, is different in that, with the exception of wave 1, telephone interviewing is mostly pursued for cost reasons. And it is perhaps this that best explains why we reached quite different conclusions from our analysis of HILDA Survey data from waves 2 to 4 (Watson & Wooden, 2009). We estimated independent probit models predicting survey response in wave *t*, conditional on contact being made, as a function of interview and respondent characteristics at time *t-1*. After inclusion of a large number of covariates, including a range of measures assumed to be correlated with respondent attitudes towards survey participation, a telephone interview outcome at time *t-1* was found to be associated with a higher probability of response at time *t*. This result, however, was only weakly significant, and re-estimation of these models using data up to wave 7 has seen the magnitude of the coefficient on this variable decline sharply and become statistically insignificant.

We also tested for the significance of interaction effects between survey mode and other covariates. For the most part, the interaction terms were insignificant, though there was one notable exception. Item non-response in one wave is generally found to be predictive of attrition at the next wave, but this was found to be especially so when the interview is undertaken by telephone.

Finally, we included a regressor identifying all interviews undertaken in the final phase of data collection. As expected, this regressor was negative and highly significant, indicating that the most difficult to contract respondents in one wave are far less likely to respond the next. Nevertheless, the inclusion of this variable had minimal impact on the estimated impact of survey mode.

On balance, there is little evidence from the HILDA Survey data experience to suggest that mode of interview is associated with attrition at later waves. The effect is positive but small and highly imprecise. That said, we admit that in the absence of an experiment we have not identified the appropriate counterfactual.

5. The effect of interview mode on item non-response

Even if survey mode has no consequences for unit response, the amount of survey data collected might still vary across survey modes because of differences in item non-response rates. This can occur in the HILDA Survey both as a result of differences in the rate with which SCQs are returned or because of differences in the incidence of refusals and don't know responses.

5.1 The impact of telephone interviewing on SCQ response

An obvious consequence of respondents electing for a telephone interview is a decline in the likelihood of an SCQ being returned. When interviews are conducted face-to-face the SCQs are handed directly to interviewees, with arrangements made for the interviewers to physically collect the completed questionnaire. Indeed, in many cases, interviewers are able to take the completed SCQs away with them on the same day the interview is conducted. With telephone interviews, however, the SCQ has to be mailed to interviewees with instructions to return the completed questionnaire in the post. Mail-out surveys, of course, are well known to be associated with much lower rates of response than surveys delivered using more personal forms of contact, and the HILDA Survey is no exception.

As shown in Table 5-1, SCQ response rates have been falling more or less continuously since the HILDA Survey commenced. Table 5-1 also suggests that much of this decline is explained by the increase over time in the incidence of telephone interviewing. The SCQ response rate among telephone interview respondents only averages 63%, whereas among face-to-face interview respondents it averages 93%, though the trend in the latter is still distinctly downwards. This is confirmed by results from the estimation of a pooled data probit model, which revealed, after controlling for interview wave as well as other interview and respondent characteristics, that telephone interview respondents are associated with an SCQ response rate that is 23 percentage points lower than otherwise comparable face-to-face interviewees. The results also confirm that independent of this mode effect, SCQ response has been declining over time. By wave 7 the SCQ response rate is estimated to have fallen by about 3.3 percentage points after controlling for observed differences in individuals.

Table 5-1
SCQ response rates (%) by wave, HILDA Survey waves 1 to 8

	W1	W2	W3	W4	W5	W6	W7	W8
Overall response rate	93.5	93.0	92.3	91.9	89.9	90.8	89.0	87.6
Face-to-face interview respondents	93.7	93.9	93.5	93.3	91.8	92.6	91.4	90.7
Telephone interview respondents	52.7	63.3	68.1	68.2	62.3	64.9	63.2	59.4
Telephone interviews as a % of all ivws	0.5	3.0	4.6	5.6	6.5	6.6	8.5	10.1

5.2 Interview mode and the incidence of refusals and don't know responses

In Table 5.2 we report the incidence of refusals and don't know responses (i.e., item non-response) for selected sub-sections of our interviews, as well as for the returned SCQs. For the five interview components the rates of item non-response are higher for respondents interviewed by telephone. The general level of item non-response, however, is very low, suggesting that this issue is trivial.

Table 5-2
Item-non response (%) by interview mode, HILDA Survey waves 2 to 7

Questionnaire component	Face-to-face	Telephone	Prob. diff. = 0
Child care / Housing (HQ)	0.97	1.84	.000
Employment / Labour force	0.07	0.11	.000
Income	0.55	0.87	.000
Family formation / Relationships	0.24	0.32	.002
'Special' interview modules	0.79	1.26	.000
SCQ (if returned)	1.83	0.94	.000

In contrast, rates of item non-response are, as we would expect, more substantial in the SCQ. And in this case the telephone interviewees actually perform better. We believe that this is simply a function of self-selection. That is, the respondents most likely to not answer questions will be the same persons who failed to return an SCQ, bearing in mind that the incidence of SCQ non-response is much higher among telephone interviewees.

6. Interview mode, data quality and respondent answers

The final questions we would like to examine are whether the quality of responses differs by, and whether answers to questions are sensitive to, delivery mode. We begin by first presenting, in Table 6-1, a short array of measures that might be indicative of data quality. As expected, telephone interviews are considerably shorter than face-to-face interviews which might be suggestive of less considered answers, but alternatively it might just reflect the lesser time spent on social interaction in telephone interviews. We can also see that the length of some answers (i.e., on occupation) are slightly shorter on the telephone. Interviewer assessments of respondents' understanding of survey question, however, suggest that it is the telephone respondents who are most knowledgeable. This, however, is largely a reflection of differences in education levels.

Table 6-1
Data quality indicators by interview mode, HILDA Survey waves 2 to 7

Indicator	Face-to-face	Telephone	Prob. diff. = 0
Mean PQ interview length (minutes)	32.1	28.4	.000
% with 'excellent' understanding of questions	69.2	77.2	.000
% with 'fair', 'poor' or 'very poor' understanding of questions	4.4	4.9	.132
Mean text length (characters) of occupation answers	77.4	72.1	.000
Mean text length (characters) of industry answers	22.7	21.7	.108

Second, for a small selection of outcome variables we examined whether there was any evidence that responses varied with interview mode. It might be expected, for example, that responses to life satisfaction questions would be subject to social desirability bias, with respondents more inclined to overstate satisfaction in the physical presence of a third party. As it turns out, however, and as shown in Table 6-2, if anything the reverse is true. Respondents in face-to-face interviews tend to provide lower scores. Part of the explanation for these results again lies in systematic differences in the two sub-populations. Once we control for individual heterogeneity (using a fixed effects framework) some of these differences decline in size, and become insignificant. This, however, is not true of satisfaction with financial situation, own health and the amount of free time. In each of these cases telephone interviewees are found to report significantly higher scores (between 0.17 and 0.19 of a point), after accounting for both a small number of observables (e.g., age, sex, education, household income, marital status, employment status, disability, and various interview characteristics) and unobserved heterogeneity (individual fixed effects). Why telephone interviewees should report more positive responses, and only on these three domains, is not obvious.

Table 6-2
Mean life satisfaction scores (on 0-10 scale) by interview mode, HILDA Survey waves 2 to 7

Satisfaction with:	Face-to-face	Telephone	Prob. diff. = 0	Coeff. (std. err.) on telephone in FE model
Home	7.83	7.82	.794	.016 (.029)
Employment opportunities	7.48	7.67	.000	.088 (.046)
Financial situation	6.50	6.66	.004	.191 (.031)
Personal safety	8.14	8.26	.001	-.047 (.025)
Feeling part of the local community	6.71	6.88	.002	-.025 (.032)
Health	7.60	7.77	.000	.171 (.025)
Neighbourhood	7.90	7.95	.174	.027 (.027)
Amount of free time	6.11	6.22	.062	.192 (.038)
Overall life	7.89	7.88	.837	.008 (.021)

Testing for mode effects with respect to more objective outcome variables in the HILDA Survey data set is more difficult, largely because many variables will be correlated with the choice of interview mode. It might be argued, for example, that social desirability biases could lead respondents (especially men) to exaggerate the hours worked each week. But as previously discussed, we expect and find that persons who work long hours are more likely to be interviewed by telephone. It is thus not possible to determine whether responses to questions about hours worked are subject to mode effects by simply comparing

group means. However, data on hours worked are also collected in the SCQ, and we can use the difference between the reports in the SCQ and in the personal interview to identify the possibility of response biases mode. The simple descriptive results are reported in Table 6-3, and reveal, counter to expectations, that it is telephone interview respondents that are most likely to exaggerate the hours worked each week. In hindsight, this might be explained as a function of panel conditioning. But again we emphasise that the size of the differences involved are quite small, and only significant among men.

Table 6-3
Usual weekly work hours and interview mode, HILDA Survey waves 2 to 7 (employed persons)

	Males			Females			Persons		
	F2F	Tel.	P	F2F	Tel.	P	F2F	Tel.	P
SCQ	41.5	44.3	.000	30.5	33.0	.000	36.2	39.1	.000
Personal interview	42.0	45.5	.000	30.9	33.6	.000	36.7	40.0	.000
Personal interview less SCQ	0.50	1.15	.033	0.47	0.57	.698	0.48	0.88	.051

7. Concluding remarks

The findings reported on here, while based on simple and arguably crude analyses, should be comforting to panel survey designers. While selection into different survey modes is clearly non-random, we nevertheless uncovered very little evidence that the use of multiple and mixed modes had any harmful effects. Panel attrition rates, if anything, are enhanced by the presence of two alternative modes of survey delivery, though the evidence presented here is extremely weak. We could also find little evidence that the way respondents answer different questions varies markedly with survey mode. While far from conclusive, our results suggest that changes in mode are unlikely to have marked effects on the longitudinal consistency of data. The only evidence of any obvious negative impact came from the interaction of mixed modes with multiple modes. Allowing respondents to respond by telephone had a clear damaging impact on the proportion of self-completed questionnaires that were returned.

References

Day, N.A., Dunt, D.R. and Day, S. (1995). Maximizing response to surveys in health-program evaluation at minimum-cost using multiple methods, *Evaluation Review*, 19, pp. 436-450.

De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone and Face-to-Face Surveys*, Amsterdam: Vrije University.

De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys, *Journal of Official Statistic*s, 21, pp. 233-255.

Dillman, D.A. (2009). Some Consequences of Survey Mode Changes in Longitudinal Surveys, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester (UK): Wiley, pp. 127-140.

Voogt, R.J.J. and Saris, W.E. (2005). Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects, *Journal of Official Statistic*s, 21, pp. 367-387.

Watson, N. and Wooden, M. (2009). Identifying Factors Affecting Longitudinal Survey Response, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester (UK): Wiley, pp. 157-182.

Wooden, M. and Watson, N. (2009). The HILDA Survey and its Contribution to Economic and Social Research (so far), *The Economic Record*, 83, pp. 208-231.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Managing Complex Inexact Matching in Coding and Linkage Applications

Michael J. Wenzowski [1]

Abstract

As the pressure increases to indirectly acquire data from administrative and other sources, the requirement to identify, and possibly even link these records expands dramatically. For example, data collected for other purposes may not be coded as required, and may also be incomplete without performing a record linkage. Both of these activities fall into the realm of inexact and probabilistic matching. As examples: the data present may not have been elicited in a manner which facilitates coding to the appropriate standard; similarly, we are rarely presented with a unique key with which to perform a deterministic linkage. These scenarios lead us to a requirement to perform inexact, or "fuzzy" matching, coupled with a probability-based method for identifying "correct" matches.

We present the results of a recent Statistics Canada initiative to re-engineer our generalized coding and record linkage systems in order to enhance their applicability across a wide range of processing problem and subject matter domains. They are typically adapted for a particular use by the methodology team responsible for creating the application, and are routinely run in production by survey operations staff. We will demonstrate how we have increased the general usability of these packages by offering more intuitive controls over managing the complexity of their internal processing, and have simplified their installation, set-up and processing models. The approach taken will be to highlight the "end-user" perspective on using this software – whether that user be IT, methodology or survey operations staff.

[1] Michael J. Wenzowski, Statistics Canada, Canada (michael.wensowski@statcan.gc.ca)

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Managing Respondent Relations on the National Population Health Survey

Andrew MacKenzie and Natasha Zaletel [1]

Abstract

The National Population Health Survey (NPHS) is a longitudinal survey that has been collecting information on the health of the Canadian population and related socio-demographic information since 1994. By the fall of 2009, the NPHS project will have collected eight cycles of data spanning 15 years and will be preparing to collect the ninth cycle in 2010. Response rates to the NPHS have remained high throughout the first 8 cycles of data collection but have trailed off similar to most other social surveys in recent years. Being a longitudinal survey, maintaining a sample of willing participants is critical, particularly as respondents are lost through mortality, relocation and other forms of non-response. NPHS has enjoyed very positive respondent relations and the purpose of this presentation is to share the methods that have been used to foster and develop these ongoing relations with respondents. NPHS has invested heavily in respondent relations over the years including focus group testing of introductory letters and brochures as well as gifts and thank you letters for survey participants. This presentation will also discuss the balance sought by NPHS in deciding when to remove a respondent from the collection sample when they are untraceable or have repeatedly refused to participate.

[1] Andrew MacKenzie (andrew.mackenzie@statcan.gc.ca) and Natasha Zaletel, Statistics Canada

ANALYSIS OF LONGITUDINAL SURVEY DATA

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

A Simulation Study of Calibration Methods for Estimation of Gross Flows

Marcel de Toledo Vieira and Gad Nathan [1]

Abstract

Methods traditionally used for the analysis of longitudinal data, such as those based on the application of generalized linear models or multilevel models to repeated measures and the use of generalized estimating equations, are primarily model-based. We consider the application of calibration methods for the estimation of gross flows from longitudinal data. Calibration can be carried out to known totals of cross-sectional variables or to longitudinal auxiliary variables and the choice of suitable distance functions allows a wide range of both design-based and model-based estimators, such as GREG estimators. The simulation study is based on data from the British Household Panel Survey and is planned to compare the efficiency of calibration to cross-sectional variables with calibration to longitudinal auxiliary variables as well as of traditional estimators of gross flows.

[1] Marcel de Toledo Vieira, Federal University of Juiz de Fora, Brazil (marcel.vieira@ufjf.edu.br); Gad Nathan, Hebrew University of Jerusalem, Israel

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Loss to Follow-up and Cox PH Modeling of Jobless Spells from SLID

Dagmar Mariaca Hajducek and Jerry Lawless [1]

Abstract

The Survey of Labour and Income Dynamics (SLID) provides information to help understand the dynamics in the Canadian population pertaining to the employment, income, health, and many other aspects of human life. As a multipurpose survey, which intends to provide information on a diverse spectrum of interests, SLID offers many statistical challenges and in particular, an opportunity for the development of the extension of duration analysis theory to longitudinal survey data. As an illustration, the study of the duration of jobless spells from SLID requires the extension of classical duration analysis techniques to survey data. One of the features that characterize these data pertains to loss to follow-up (LTF), which may be related to the durations under study. For example, the probability that an individual is lost to follow-up may be related to their unemployment experience. On the one hand, Cox PH modeling techniques deserve attention due to their appeal, both in the classical and survey settings. Conditioning on past event history may be used to deal with within-individual correlation in spell durations. On the other hand, weighting techniques such as inverse probability weighting (IPC) based on LTF modeling allow us to accommodate dependent LTF. This talk will discuss the use of LTF based IPC weights in Cox PH conditional modeling of jobless spells durations.

[1] Dagmar Mariaca Hajducek (cdmariac@uwaterloo.ca) and Jerry Lawless (jlawless@math.uwaterloo.ca), University of Waterloo, Canada

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Issues in Latent Growth Modeling with Longitudinal Public-release Data

Laura M. Stapleton[1]

Abstract

This manuscript outlines some of the issues faced by the applied researcher when analyzing longitudinal public-release data. Research questions that address the amount of growth over time, the shape of growth, and differences in growth across groups can be answered with such data; however the applied researcher must first determine how the specific data collection method used by others should be accommodated. This paper addresses the decision points for the researcher within the context of latent growth modeling.

Key Words: Analysis, Modeling, Measurement.

1. Introduction

This paper reviews the challenges that face the applied researcher who is examining longitudinal data from a probability sample with structural equation modeling (SEM.) Of specific interest will be modeling the same construct over time, to examine changes in that construct. The paper begins with a discussion of the research questions addressed by these models and the data that are required to undertake the analyses. Following the introduction of the general latent growth modeling (LGM) method, I identify the practical issues in using longitudinal data collected from probability sample designs. Some issues are of a theoretical nature while others are statistical and are relevant to both applied researchers and those who prepare public-release data sets.

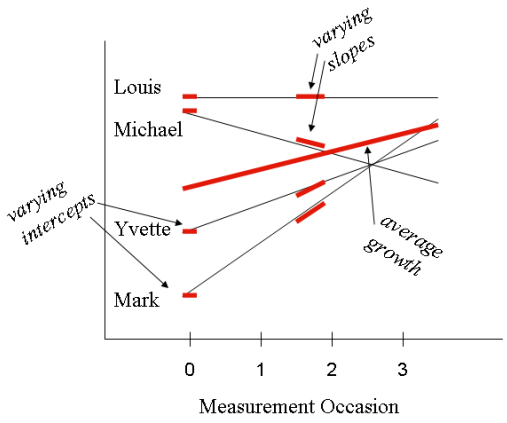
2. Research questions addressed by latent growth modeling

Longitudinal data offer the opportunity to ask research questions about change over time. Although repeated measures analysis exists in the traditional ANOVA framework, the statistical assumptions, such as sphericity and homogeneity of error variance across time, are often not met with empirical data. Additionally, such traditional methods focus on averages and do not examine individual differences in amount and trajectory of change and predictors and outcomes of those individual trajectories. Therefore, for researchers interested in examining change in a construct over time (the amount of change, the shape of the change, and moderation of change) the use of growth modeling provides an alternative. Two approaches to growth modeling include the use of a hierarchical linear modeling (HLM) framework or an LGM framework (Duncan, Duncan, Strycher, Fuzhong, & Alpert, 1999). The two approaches share many components. While HLM provides a flexible data structure for a small set of models, LGM accommodates a larger set of models but constrains the analyst to a small number of data structures (Raudenbush, 2001). Differences in the modeling approaches are beyond the scope of this paper and the remainder of the paper focuses on the LGM approach.

Example research questions addressed using an LGM analysis might include how much growth occurs on average, over time, in students' abilities (see bold line in Figure 2-1), whether there is variability in where students start on ability (see varying intercepts in Figure 2-1), and whether the growth rate differs across students (see varying slopes highlighted in Figure 2-1). Additionally, one might examine predictors and consequences of students' initial starting places and growth rates.

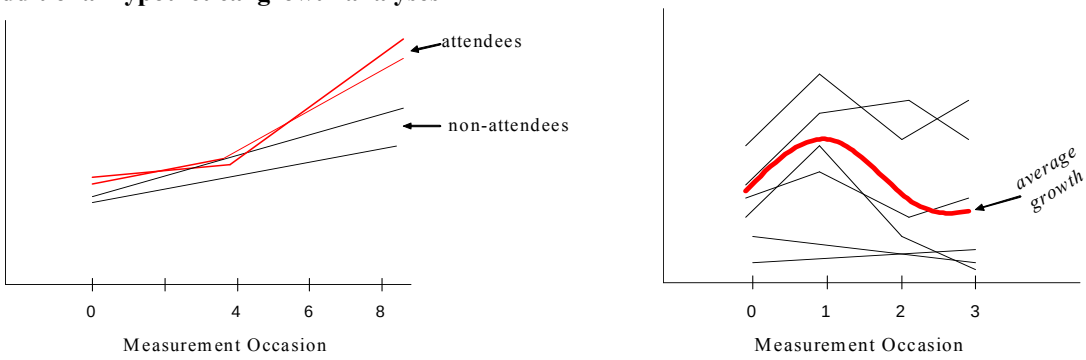
[1]Laura M. Stapleton, University of Maryland, Baltimore County, Psychology Department, 1000 Hilltop Circle, Baltimore, MD, 21230 USA

Figure 2-1
Basic hypothetical growth analysis



As another example, Panel A of Figure 2-2 shows income levels for individuals measured at 18 years old, then at 22, 24, and 26 years old. The question might be posed whether the shape of growth is different for individuals who attended college versus those who did not attend college. In this example, the non-attendees have a linear income trajectory while the college graduates have piecewise linear growth. Finally, as shown in Panel B of Figure 2-2, a non-linear growth function might be needed to model the change over time, in this case, the growth is cubic. Additionally, the shape of the growth appears to be different across individuals; there are some individuals with cubic growth and some who demonstrate basically flat lines.

Figure 2-2
Additional hypothetical growth analyses

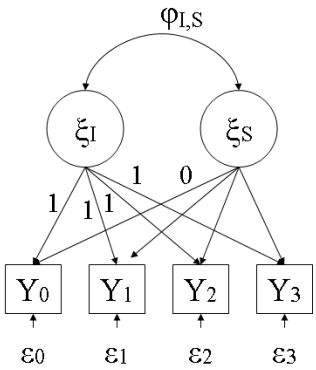


Panel A Panel B

3. Basic latent growth model identification

A latent growth model with four time points, reflecting the basic growth data shown in Figure 2-1, is laid out in Figure 3-1 (note that only the covariance structure, and not the mean structure, is shown). Typically, the data structure of an LGM for an individual is specified as $y = \lambda\xi + \varepsilon$, where **y** is a ($t$ x 1) vector of measures over t time points, λ is a (t x 2) matrix of loadings on intercept and slope factors, ξ is a (2 x 1) vector of latent intercept and slope values, and ε is a (t x 1) vector of occasion-specific residuals. This model reflects the hypothesis that each observation is a function of a latent intercept (level) and a latent slope (shape). It also may be hypothesized that the latent intercept and slope covary with each other (as shown by the double-headed arrow in Figure 3-1). The first column in the λ matrix is a vector of 1s to indicate that each measure is a function of one unit of intercept, and the values in the 2nd column of the λ matrix are defined as units of slope. If this first loading is set to 0 as shown in Figure 3-1, then the remaining loadings reflect the relative amount of growth added at each additional measurement occasion. Assuming that the data were captured at equal intervals (perhaps, each year), and if the remaining loadings in the 2nd column are set to be consecutive integers (1, 2, and 3 in the example), then the shape of growth is hypothesized to be linear; the construct increases one unit of growth at each wave. Because individuals will vary in their amount of the latent growth factor (ξ_2), each individual trajectory can take on a different slope (as shown in Figure 2-1). Definitions of the components of λ depend on the model hypothesized, for example, linear, non-linear, and with data collected at equally or non-equally spaced time points.

Figure 3-1
Example of latent growth model identification for an assumed four time point data structure



$$\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{I1} & \lambda_{S1} \\ \lambda_{I2} & \lambda_{S2} \\ \lambda_{I3} & \lambda_{S3} \\ \lambda_{I4} & \lambda_{S4} \end{bmatrix} \begin{bmatrix} \xi_I \\ \xi_S \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

Maximum likelihood estimation can be used to estimate the most likely parameters to result in the observed sample data via a fit function

$$F_{ML} = \ln|\hat{\boldsymbol{\Sigma}}| + tr(\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}) - \ln|\mathbf{S}| - p$$

where *S* represents the sample covariance matrix, *p* is the number of parameters to be estimated, and $\hat{\Sigma}$ is the model implied covariance matrix defined as

$$\hat{\Sigma} = \lambda\phi\lambda' + \theta$$

where ϕ is a (2 x 2) latent variable covariance matrix, and θ is the (*t* x *t*) covariance matrix of residuals. This fit function is covariance based, but can also include the mean vector. Also, full information maximum likelihood (FIML) finds the log-likelihood summed across individuals (and can be used to accommodate missing data, as discussed in a subsequent section).

4. Issues that arise when using public release data for latent growth modeling

In this next section, I provide a general overview of the challenges that face the applied researcher when using secondary datasets to examine latent growth. The challenges can be loosely grouped into two areas, one of measurement (or how and when the construct was measured) and one of sampling (or how the individuals who comprise the sample were identified.) The secondary researcher, by definition, has no control over how and when the data were collected and has no control regarding from whom the data were collected. Explicitly acknowledging this lack of control is essential. Deciding whether the measurement procedure results in adequate measurement of constructs of interest is the first issue to address; it is possible that the secondary researcher should not attempt the analysis given the measurement properties. Next, assuming adequate measurement, deciding how to generalize from the sample to the population of interest will require an understanding of the sampling design as the inferential statistics will depend on the sampling procedure used.

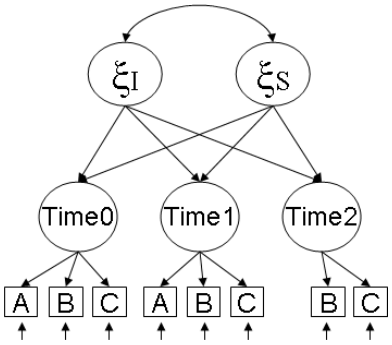
To give some context to the discussion of the issues that face the secondary researcher, I present here two datasets that have been used for latent growth modeling. The Early Childhood Longitudinal Study – Kindergarten (ECLS-K) is a survey program conducted by the National Center for Education Statistics in the U.S. This program started in Fall of 1998 with a probability sample of kindergarten students. The sampling design included a three-stage sample of geographic regions as primary sampling units (PSUs), then schools within those PSUs as secondary sampling units (SSUs), then the selection of students within those schools at different selection rates, depending on student characteristic. This survey program included seven waves of data collection between kindergarten and 8th grade. At each wave, children were assessed for their academic skills/abilities among other measurements. The National Longitudinal Study of Youth (NLSY) is conducted by the U.S. Bureau of Labor Statistics and two cohorts have been followed: 1979 and 1997. In this survey program, an area sample yielded households that were contacted to identify respondents between the ages of 12 and 17. Each year, these respondents are contacted again and currently with NLSY97, 11 years of data are available for secondary researchers. Questions were asked of these adolescents regarding their behaviors, schooling, college attendance, employment outcomes, and health. Datasets from these two survey programs have been used often as demonstrations of growth modeling using both SEM and HLM frameworks (e.g. Curran & Bollen, 2001; von Hippel, Powell, Downey & Rowland, 2007). Rarely have these demonstrations acknowledged the issues in sampling and few have explicitly addressed concerns regarding the measurement process as described below.

4.2 Construct measurement

Because longitudinal data are not necessarily collected for the purpose of LGM, several problems can arise in secondary data LGM analyses with regard to the measurement of the construct, including the consistency in how the measure was captured at each wave, whether ceiling or floor effects might exist in the measurement, and differences in the reliability of the measurement at each time point. Each of these issues is briefly discussed in this section.

One of the fundamental assumptions made in LGM is that at each measurement point, the same construct is being measured and, therefore, meaningful interpretation can be made about amount of change in the construct across time. There can be situations when questionnaires might be changed between time periods, possibly invalidating any comparison across time. The model identification addressed in section 3.1 assumed that one was modeling a single manifest variable across time. To address the issue of changing items across years, it is possible to model a latent construct over time, thus allowing for modeling with (slightly) different items at each time point. This modeling has been referred to as a curve of factors model (McArdle & Hamagami, 2001) or a second-order LGM (Hancock, Kuo & Lawrence, 2001). In this model, shown in Figure 4.2-1, the available items are used to identify the construct at each time period, where the values of the respective loadings are constrained to be the same across time period (e.g., $\lambda_{A0} = \lambda_{A1}$ and $\lambda_{B0} = \lambda_{B1} = \lambda_{B2}$). This model comes with a strong assumption of measurement invariance across time periods. For this model a researcher might also include covarying residuals of each of the respective manifest indicator variables to represent hypothesized method variance.

Figure 4.2-1
Curve-of-factors latent growth model



A second concern is the existence of ceiling or floor effects in assessments. When a construct has been measured across time, it is possible that the range of scores that was applicable at the start of the study is no longer relevant later in the study. For example, using data from ECLS-K, a research question might relate to growth in vocabulary. Theoretically, children have exponential growth in vocabulary. However, survey programs may use a vocabulary assessment on which children can top out. With such an assessment (displaying ceiling effects), growth will appear as non-linear with a negative quadratic term, thus suggesting a leveling off in vocabulary. The secondary analyst has no recourse but to recognize that the trajectory should not be interpreted as growth in vocabulary, but rather as growth in the score on the specific measure.

Finally, a measure may not have consistent reliability over time. It is possible that, at each successive wave of data collection, the measure becomes more (or less) reliable. For example, it has been found that self-report measures by children younger than 12 are less reliable than when those children are older (de Leeuw, 2005; Rebok et al., 2001). Thus, variability in measures may be greater at earlier waves of data collection for populations of children. In the LGM framework, this variability in reliability can be accommodated. As shown in Figure 3-1, the researcher specifically models error at each measurement occasion and can specify that the estimates of error variance at each time point are not constrained to be the same. The specification of varying levels of random error is one area where LGM in an SEM framework is more flexible over HLM, which typically assumes a constant residual variance across time points (Rovine & Molenaar, 2001).

4.3 Timing of data collection

A general set of concerns that researchers face when using LGM with secondary data regards the timing of the collection of the data and how to appropriately model the measurement occasions. Regardless of the decision of the type of growth to model (linear, piecewise, non-linear), the following four considerations need to be addressed: choice of loading values to reflect time intervals, consistency of time across individuals, whether the collection times are appropriate for hypothesized growth, and whether wave or another construct is the appropriate function of change. These four issues are addressed in this section.

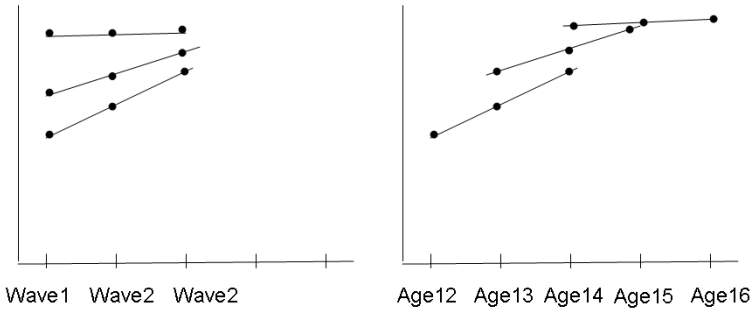
In the model identification process for latent growth models as shown in Figure 3-1, the researcher needs to explicitly model the time interval if a linear, piecewise linear, or specific non-linear model is selected (such as quadratic or cubic) by specifying the loading values. The loadings for the linear growth factor should represent the time period between measurement occasions and therefore it is crucial that the researcher understand these and appropriately model them. For example, with ECLS-K, the data were collected at the following times: Fall 1998, Spring 1999, Fall 1999, Spring 2000, Spring 2002, Spring 2004, and Spring 2007. A researcher hypothesizing linear growth should recognize that the first through fourth measurements occurred at approximately 6-month intervals. The 5th measurement occurred after a break of two years, the 6th after another 2 year break and the 7th after another 3 years and therefore the loadings might be: 0, .5, 1, 1.5, 3.5, 5.5, and 8.5 and estimates would be interpreted as yearly amount of growth.

Another concern regarding timing of data collection is consistency across individuals. When using LGM, the model is expected to be applicable for each individual and thus, each of the measurement occasions is assumed to be at the same time across individuals, referred to as balanced on time (Ware, 1985). In the ECLS-K data, however, the measurements are not taken at the same time point. In the Fall Kindergarten assessment, the assessments could have been made any time between September and December. Similarly, with the NLSY97 data, often researchers model the data given the age of the respondent (12, 13, 14, etc.) However, at the time of measurement, some children might have been 12 years and 4 months, others 12 years and 9 months. McArdle and Hamagami (2001) partially addressed the issue by using ½ year increments instead of yearly increments. Assuming data are balanced on time when they are not can jeopardize conclusions made about variability in trajectories of growth and therefore HLM is a reasonable alternative for growth modeling, as it does not have an assumption of balanced on time.

A focus of LGM is to evaluate hypotheses about the shape of the trajectory. The model estimates, however, are dependent on whether your data allow you to investigate the change at key points over time. For example, in examining BMI change over time in children using ECLS-K, von Hippel and colleagues (2007) examined whether growth rates were different during school and non-school periods (summer), however the ECLS-K data points were not collected at the beginning and end of the school year; on average, the data were collected in mid October and early May. The authors had to assume a linear trend across months within period. Because secondary researchers do not have control over timing of data collection, care must be taken to only evaluate hypotheses that can be appropriately tested given the data. Shapes can only be "seen" if a data collection occurred at a crucial turning point in the trajectory.

Finally, in addition to hypotheses about shapes of trajectories, with LGM, researchers can examine whether there is variability across individuals. Therefore, there should be some expectation that change could operate similarly across individuals at the modeled time points. Just because data are collected at the same wave from a set of people does not mean that all individuals should be expected to be similar at that wave. For example, with the NLSY97 accelerated longitudinal design (or cohort sequential design), a cohort of a range of ages is followed. One participant may be 14 years old at the first wave while another is 12 at the first wave and therefore 14 at wave three. There is no reason to expect that these individuals would be similar at wave one or would have similar trajectories. Data should be modeled based on the relevant function of change (such as age or years since beginning employment), not data collection time (McArdle & Hamagami, 2001). For example, in Figure 4.3-1, suppose that data were collected for three waves from children who were between ages 12 and 14 at the first wave. The panel on the left would likely be an inappropriate way to model the data, while the panel on the right may appropriately reflect a developmental construct. Similarly, for ECLS-K, to examine change in height over time, semester in school may not be relevant and age in months may be a better function of change to be modeled.

Figure 4.3-1
Choice of appropriate time referent for latent growth modeling



4.4 Missing data accommodation

Under longitudinal survey programs, participants may stop out for one wave or more and return eventually or may drop out from data collection entirely. Three options of working with missing data at waves are multiple imputation, FIML estimation, and use of non-response adjusted panel weights. In multiple imputation analysis, all data available for a participant (not just those in the model) can be used to estimate plausible values of the missing items (typically 5 or 10 are selected). These random draws are then used in multiple analyses and the results combined. This approach assumes that the data are missing at random or have "accessible" missingness (Graham, Taylor & Cumsille, 2001). In FIML estimation, available data are used to find the set of parameters to maximize the likelihood of the given data for an individual and only those parameters associated with the available data are estimated for that individual log likelihood. The individual likelihoods are then summed. This approach has been found to provide robust estimates when assumptions are met (Enders, 2001).

A final option to consider with missing data is to make use of non-response adjusted panel weights that are provided on data sets. These panel weights differ from the sampling design weight as they are adjusted to reflect participants who drop or stop out. These weights are carefully created to reflect the type of respondents who have dropped out, using techniques such as multiple imputation analysis and raking. If several panel weights are provided on the data file, the researcher needs to choose the one that best fits the research question. For example, on the ECLS-K dataset, there have been seven data collection times. Some researchers may want to model using only the kindergarten and first grade data. Others may desire to use the spring kindergarten, 1st, 3rd, 5th, and 8th grade data, while others may just want to model with kindergarten, 1st and 3rd grade data. For each of these analyses, a panel weight exists that is a value of zero if the individual did not participate in all relevant waves and is non-zero if he or she did. It is possible, however, that an analyst wishes to model with data from kindergarten, 1st and 3rd grade, but wishes to include in the analysis any participants who "stopped out" during 1st grade (but eventually returned to the study in 3rd grade). In this case, the researcher cannot use the panel weight and must address the missing data with either multiple imputation or FIML estimation.

4.5 Unequal selection probability

In an LGM context, ignoring differential rates of selection will likely lead to biased estimates of intercept and slope, as well as other model estimates, if related to the probability of selection. Of course, if weighted estimates differ from unweighted estimates, the researcher should consider examining moderation across explicit sampling strata. Recently, Asparouhov and Muthén (2005) and Rabe-Hesketh and Skrondal (2006) evaluated a pseudo maximum likelihood estimation with incorporation of sampling and non-response adjusted weights in SEM analyses and found the estimation to be robust.

4.6 Variance estimation

Because public-release data typically are not collected using simple random samples, special approaches to estimating standard errors (referred to as variance estimation) are needed. For SEM analyses, the dependency in observations not only affects standard error estimation, but also χ^2 tests of model fit (Muthén & Satorra, 1995). Researchers have two choices for appropriate variance estimation for latent growth models: linearization and replication methods. In the linearization option, current software packages use pseudo maximum likelihood estimation such that there is a sandwich estimator to "scale" the asymptotic covariance matrix of the estimates as well as the χ^2 model fit statistic (Asparouhov & Muthén, 2005; Rabe-Hesketh & Skrondal, 2006). Most SEM software can accommodate this estimation, assuming the sample PSU and stratum indicators are available on the dataset. Replication methods are not currently available in software packages for SEM. Jackknife, balanced repeated replicates, and bootstrapping can be accommodated and Stapleton (2008) provided SAS macro programming to do such analyses with the Mplus software and found these methods to be robust.

4.7 Clustered and nested observations

The prior discussion has assumed that the LGM analyses were undertaken at an individual level and interpretation was limited to the individual and, furthermore, was applicable to all individuals. A different type of analysis, a multilevel LGM, would address the clustered nature of the observations. Specific research questions would include, do student trajectories differ across schools? Are specific school characteristics related to the average growth rate of students in the school? A concern with this type of analysis using secondary data is the need to have sampling weights at each level of the analysis (Rabe-Hesketh & Skrondal, 2006). Currently, most public release data sets only provide individual level sampling weights, not the PSU weight or the conditional weights of the SSU and USU. Another concern is that although a multilevel analysis may accommodate one stage of the sampling, the full design has not been addressed. For example, with ECLS-K, if a multilevel LGM of student growth nested in schools is undertaken, then only part of the sampling design has been accommodated. Schools are SSUs, so

the dependency of schools within PSUs is still a concern, and therefore special variance estimation techniques would still be required (Rabe-Hesketh & Skrondal, 2006).

5. Summary

In summary, whenever undertaking latent growth modeling with secondary data, three general issues should be addressed. First, the research questions need to be framed in light of available data (specifically, constructs measured and the times available). Second, the model needs to be carefully built to accommodate the data characteristics (e.g., measurement occasions, availability of data at given time points) and if the model cannot address the research question, then it should be abandoned. Finally, parameter and variance estimation should address the sampling design as well as any non-response over time.

References

Asparouhov, T. and Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf, retrieved September 26, 2006.

Curran, P. J. and Bollen, K. A. (2001). The best of both worlds: combining autoregressive and latent curve models. In L. M. Collins, & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 107-135). Washington, DC: American Psychological Association.

de Leeuw, E. (2005). Surveying children. In Best, S. G. (Ed.) *Polling America: An Encyclopedia of Public Opinion.* Westport, CT: Greenwood Press.

Duncan, T. E., Duncan, S. C., Strycher, L. A., Fuzhong, L. and Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 313-342). Greenwich, CT: Information Age Publishing.

Graham, J. W., Taylor, B. J. and Cumsille, P. E. (2001). Planning missing-data designs in analysis of change. In L. M. Collins, & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 333-353). Washington, DC: American Psychological Association.

Hancock, G. R., Kuo, W.-L. and Lawrence, F. W. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8,* pp. 470-489.

McArdle, J. J. and Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 137-175). Washington, DC: American Psychological Association.

Muthén, B. O. and Satorra, A. (1995). Complex sample data in structural equation modeling. In Marsden, P. V. (Ed). *Sociological Methodology*, pp. 267-316.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series B,* 60, pp. 23-56.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 35-64). Washington, DC: American Psychological Association.

Rovine, M. J. and Molenaar, P. C. M. (2001). A structural equations modeling approach to the general linear mixed model. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 65-6496). Washington, DC: American Psychological Association.

Rebok, G., Riley, A., Forrest, C., Starfield, B., Green, B., Robertsoon, J. and Tambor, E. (2001). Elementary school-aged children's reports of their health: A cognitive interviewing study. *Quality of Life Research, 10,* pp. 59-70.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling, 15,* pp. 183-210.

von Hippel, P. T., Powell, B., Downey, D. B. and Rowland, N. J. (2007). The effect of school on overweight in childhood: Gain in body mass index during the school year and during summer vacation. *American Journal of Public Health, 97(4),* pp. 696-702.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician, 39,* pp. 95-101.

ISSUES IN ECONOMIC SURVEYS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

The Construction of a Prototype of the Italian LEED Based on Administrative Data: Main Methodological Aspects

Colace Andrea, Congia M.Carla and Rizzi Roberta[1]

Abstract

A first prototype of an official Italian longitudinal Linked Employer-Employee Data base (LEED) has been designed and realised by ISTAT mainly using administrative data, in particular the tax declarations transmitted by employers to the Tax Office. With the entire population of employers and employees coverage, the LEED results from a complex integration process with other administrative sources and the Italian Statistical Business Register. As the entities within the LEED are linked longitudinally, it is possible to track workers over time and link this to longitudinal firm dynamics allowing the analysis of important aspects concerning the labour market. In this paper the main methodological aspects for the construction of such a LEED are discussed.

Key Words: Linked Employer-Employee Data, Administrative Data, Longitudinal Linked Data.

1. Introduction

Since many years the Italian Institute of Statistics (ISTAT) has been exploiting administrative data for statistical purposes extensively. This work focuses on recent developments in the use of administrative data aimed at making progress in the production of official data sets suitable for analysis of labour market dynamics. Creating a longitudinal linked employer-employee data set (LEED) based mainly on administrative files represents a great challenge for the Italian NSI. After a brief focus on the relevance of a longitudinal LEED, the international and national context in which this ISTAT experience takes place is pointed out. A description of the administrative sources chosen for setting up an official Italian longitudinal LEED is given as well as the basic infrastructure of the first prototype and the main linkage issues are illustrated.

2. Linked Employer-Employee Data: relevance and previous experiences

Analyses of labour market dynamics which aim at capturing the complex interactions of employers and employees at micro level require the availability of adequate data. The creation of data bases that link firm and workers characteristics is not an easy task, specially when they are designed with a longitudinal approach. However, the relevance of linked employer-employee data for analysis of labour market policies is widely recognised (Abowd, 2008; Contini, 2005). Such a data set enables policy analysts and researchers to consider information relating to both the supply and the demand sides of the labour market simultaneously. When the entities within the LEED can be linked longitudinally, the informative value of these data sets increases enormously. Such a panel data set tracks the same employees and their workplaces over time, providing crucial information to investigate a wide variety of labour market outcomes like worker and job flows, employment tenure and multiple job holding, wage dynamics, firm demography and productivity, effects of persons and firm characteristics in the determination of compensation, individual mobility in relation to firm-specific employment adjustments, etc.. This kind of LEED may also allow to extend or improve the production of official statistics and to fill possible information gap in labour market dynamics studies, which still exist in Italy.

Many European and other advanced countries with a long tradition of micro-econometric analysis of labour market has developed suitable LEEDs for statistical purposes. Looking at the existing international experiences, many different types of matched employer-employee data sets have been created. They can be classified according to several dimensions. First, some are cross-sectional data sets while others are longitudinal. Cross-sectional LEED contain both employer and employees data but most of them relate to a single point in time. This is the reason why that approach does not allow to really analyse the labour market dynamics. On the other hand, LEED with longitudinal component track workers and their firms or workplaces over time and are suited to study dynamic interrelationships among firms and workers. The second dimension that can be used to classify the existing LEED is their representativeness. Some LEED are based on a sampling design focused only on

[1]Colace Andrea, ISTAT-Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (colace@istat.it); Congia M.Carla, ISTAT- Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (congia@istat.it); Rizzi Roberta, ISTAT- Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (rizzi@istat.it)

employees, while others are representative only of the population of firms. It may arise that a dataset matching employees with their employers is not designed to be representative cross-sections and panels of workers and firms. Their design features and their coverage depends on the purpose of the user (Abowd, 1999). In this work a special attention is paid to the representative matched worker-firm panels. They can be based on statistical surveys or on administrative data. Generally, survey-based data contain smaller numbers of observations than administratively-based LEED but are richer in covariates. The value of these survey-based LEED lies in the detailed information they contain. However, many longitudinal representative matched employer-employee datasets are based on administrative files. Among them, official linked datasets developed from administrative data are becoming a form of LEED more common as National Statistical Institutes are more sensitive to the need for resources of this type and decide to invest in their production (Desai, 2008). This data suffers from a number of weaknesses related to statistical exploitation of administrative data but potentially allow the construction of higher quality datasets. This kind of LEED has been implemented in many European countries like Denmark, Sweden and France. Some leading experiences outside the European Union are those of Statistics New Zealand (Padapopolous, 2004) and U.S. Census Bureau (Abowd, 2004).
In Italy, an official experience is represented by the European Structure of Earnings Survey performed four-yearly by ISTAT on the basis of a two stage sampling frame that can assure only a cross-sectional representativeness of both firms and workers. The other main Italian experience in constructing a linked employer-employee data set is the Workers Panel based on the administrative data of the National Social Security Institute (INPS). The raw data, extracted directly by INPS from its administrative database, are a sample of workers nationally representative longitudinal (from 1985 to 2004) so that suffer from some shortcomings on firms representativeness. Then data are supplied for research purposes to the University of Turin and the experts of the Research Centre Laboratorio Revelli that have created the Working History Italian Panel-WHIP, to the Ministry of Labour to populate the database CLAP (Campione Longitudinale degli Attivi e dei Pensionati), to the Italian Institute for the Development of Vocational Training for Workers (ISFOL) and some other Italian Universities (Bocconi, La Sapienza).

3. The prototype of the Italian longitudinal LEED

3.1 The sources used

Why an "official" and "administratively-based" LEED?
In the last years, the Italian National Institute of Statistics has been more sensitive to the need of researchers and policy analysts for reliable and accessible matched employer-employee data sets and has started to invest in the first activities for the design and the implementation of a longitudinal linked employer-employee data set mainly based on administrative data. The Italian NSI access to administrative micro data is assured by national laws. This implies the access to a wider range of data sources than a researcher would have, potentially allowing the construction of datasets of higher quality, characterised by an homogeneity of treatment and methods, with larger population coverage and fewer biases. The Italian NSI is also supposed to have more resources to develop datasets that will be made available to academics and researchers.

Following a strategy of extensive exploitation of administrative sources for statistical purposes since 1990s ISTAT has been using administrative data developing a great experience in treating and combining these sources. It is in this context that the activity of using the administrative sources for the construction of the first official longitudinal LEED prototype takes place. In particular, two experiences have been crucial in the following design of the LEED: the construction of the Business Register (ASIA) integrating mainly administrative sources and the feasibility studies for the statistical exploitation of the tax individual records (Tax form 770) and the social security individual records (EMens form).

The informative contents
The development of ASIA is the result of the integration of several administrative sources and constitutes the first example of a statistical product built almost exclusively on administrative data. By integrating them it was possible to validate information present in individual sources in order to correct errors, but above all to transform administrative information into statistical data, consistent with the concepts and specific definitions of the official statistical system. ASIA allows the annual availability of the universe of industrial and service businesses that are active in the country, and to know their location, number of employees, economic activity and turnover of business. The Business Register makes it possible to have a unique framework of reference for all official economic statistics. The other administrative sources can be linked to it by the tax code number of the enterprise, available both in the tax and in the social security sources hereafter illustrated, acquiring a preliminary statistical valence.

The administrative source of Tax form 770 used in the construction of the LEED's prototype is the compulsory annual declaration of taxes, social security contributions and work injuries insurance premium deducted by the withholding agents (enterprises, public and private institutions and self-employed). The coverage includes workers who, during the year prior to the tax year, worked at least one period, whether as employees or self-employed.

BUSINESS REGISTER (ASIA)	TAX FORM 770	ANNUALISED E-MENS
Identification and structural data business fiscal code **[employer primary ID]** legal form economic activity sector (ATECO 2002-ATECO 2007) turnover class **Employment data** Number of workers Number of employees Number of self-employed workers **Geographical data** municipality (description and ISTAT code) province (description and ISTAT code) region (description and ISTAT code) macro-area (description and ISTAT code)	**Identification data** fiscal code of withholding agent (business) **[employer primary ID]** worker's fiscal code **[worker ID]** worker's date of birth worker's country of origin **Fiscal data regarding income data** labour incomes subject to law deductions data to calculate tax due number of days for which deductions are due supplementary social insurance contributions and for welfare purposes redudancy pays **Social security data regarding employment and income data** category of beneficiary (employee, collaborator, pensioner) professional qualification (blue-collar workers, employees, etc…) (a) working time (full-time, part-time) (a) duration of contract (open-ended, fixed-term, seasonal) (a) gross wages paid days (a) paid weeks (a) paid months (detailing individual months) collective bargaining agreement (code) (a) collective bargaining agreement (type) (b) level of salary scheme (b) contributions paid by workers wages paid to collaborators contributions payable to collaborators hire and termination date of the collaboration activity (b) months paid to collaborators (detailing individual months) (c) **Insurance data regarding employment data** hire and termination date of the job **Geographical data** province where the work activity is carried out (b)	**Identification data** business INPS register number **[employer secondary ID]** business fiscal code **[employer primary ID]** worker's fiscal code **[worker ID]** **Data regarding employment data** professional qualification (blue-collar workers, employees, etc…) working time (full-time, part-time) duration of contract (open-ended, fixed-term, seasonal) contribution type (identifies workers whit specific contribution situations) worker type (identifies specific types of workers) collective bargaining agreement (code) gross wages paid days paid weeks paid months (detailing individual months) hire and termination date of the job **Geographical data** municipality where the work activity is principally carried out (ISTAT code) province where the work activity is principally carried out (ISTAT code)

The use for statistical purposes of the Tax form 770 started in ISTAT officially in 2006, following the creation of an inter-institutional working group consisting of some Institute's researchers with different expertises in the processing of administrative sources, in the use of fiscal data, in the utilisation of social insurance variables, etc., and also specialist staff of the Ministry of Economy and Finances and the National Social Security Institute (INPS). The activity carried out by the working group aimed also to integrate, improve and extend ISTAT official informative offer on important economic phenomena, in particular on net wages and tax/contribution wedge of workers (Calzaroni et al., 2008). Starting from the reference year 2005, the Tax form 770 changed as a result of modifications to the legislation; some information that it contains has thus been migrated into other administrative sources. In order to ensure that the Tax form 770's information remains available over time, it has therefore become necessary to acquire and to begin to study and to process a new source provided by the National Social Security Institute: the Annualised E-Mens. It is the result of a processing phase conducted by INPS in order to develop an archive containing aggregate annual data, drawn from the data of the monthly declaration of social security contributions (E-Mens form) paid by employers (enterprises, public and private institutions) to INPS.

Beginning from the original informative contents of each source used, to develop the prototype of LEED, the most relevant variables were selected and some other statistical variables have been retrieved from the administrative information. Given the large number of variables contained especially in the tax source, those of greatest relevance to the construction of the LEED and the development of labour market analysis are grouped and summarised above. Every source is characterized by variables that can be classified as identification data (where the ID variables are included), employment data and geographical data, while the Tax form 770 and the Annualised E-Mens are also characterised by fiscal or insurance data regarding income data. In the scheme, in the column concerning the Tax form 770, some relevant notes are also reported to indicate: the data not gathered since 2005 and migrated in the other source (a); the data not gathered since 2005 and not migrated in the other source (b); and the data included since 2005 (c).

3.2 The main methodological aspects

The treatment process of administrative data

A very complex treatment process has been needed to exploit the administrative data used. First, an examination of the source's informative contents and an analysis of the administrative rules through which it was possible to identify the general categories of subjects who must fill in the tax and the social security forms (reference universe), the relevant subjects involved (analysis units) and the income categories included in the form has been carried out. The comparison of the definitions and classifications, provided by the tax law in relation to those of the official statistics, preceded the phase of integration with the

Business Register to assess the coverage of the source's universes of reference in terms of withholding agents (businesses) and receivers (employees).

Then, the retrieving of the statistical variables from the administrative information has been necessary. Some of the statistical variables listed in the above summary are the result of extremely painstaking operations, as in the case of the wage variables obtained from the fiscal data declared with an high level of detail in the Tax form 770, taking into account the way in which in Italy net income tax is calculated in order to determine net as well as gross wages. On the other hand, some variables not included in the sources, but essential to integrate the input archives, has been retrieved. This is the case of the business fiscal code in the Annualised E-Mens, obtained through a linkage with the INPS business register number.

Official statistical classifications has been assigned through the linkage with the Business Register ASIA, also translating the administrative codes as in the case of geographical information.

A continuous monitoring of law and administrative rules which may change over time implying the changing of the informative contents of the administrative sources, as the migration of some information from the Tax form 770 to the Annualized E-Mens, has been carried out.

Finally, the multiple forms for the same job has to be treated and each administrative source has to be validated before their integration.

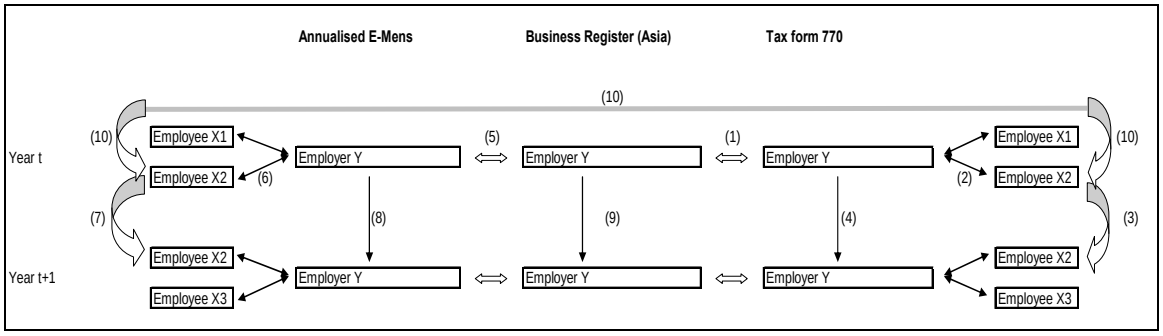
The integration of sources

In order to construct the longitudinal LEED it has been necessary to take several aspects of integration into account: between sources, between units of analysis and lastly integration in relation to time (reference years).

1) With reference to sources, integration was carried out among the sources described previously. In particular with regard to integration between the two sources which also contain data on workers (Tax form 770 and Annualised E-Mens), a comparison activity between the values taken on by the variables common to the two sources, as well as a verification of the magnitude's values of the migrated variables compared with the original ones, has been added to the linkage operation.

2) With reference to units of analysis, it is necessary to distinguish ASIA from the other two sources. In fact the Business Register's unit of analysis is the business, while in the other two archives the units of analysis are both the business and the worker. The presence of the business fiscal code in the two administrative sources enable them to be linked to ASIA, making it possible to have an official frame of the characteristics of the business for which workers provide their labour. At the same time, the availability of both tax codes makes it possible to integrate the two archives, assuring the completeness of the informative potential.

3) Lastly, with reference to the temporal aspect, it is pointed out that the official Italian LEED prototype is referred to the years 2005 and 2006. The decision to construct the LEED on the minimum duration to speak of a longitudinal database (two years), even thought the multi-year experience in the utilisation for statistical purposes of the administrative sources, in the management of the Business Register and in the availability of the data, was dictated by issues of data compatibility following modifications to legislation. The reconstruction of the time series of data may represent a significant undertaking with regard to data processing following the process of interpreting and reconciling the variables gathered from different administrative archives. To show the complexity of the set of links that it is necessary to manage among sources, units of analysis and reference years in order to construct the LEED prototype, an outline is provided in the following diagram.



In the Tax form770 archive the links to be managed are: the link between worker and business during a specific period of time (link 2); the link between the same workers in different years (link 3), in order to track the individual's working career; the link

between the same businesses in different years (link 4), in order to track the dynamic of businesses (and their consequent births/deaths or merger-acquisition).

Similarly the same links are found in the Annualised E-Mens archive and consist of: the link between worker and business during a specific period of time (link 6); the link between the same workers in different years (link 7); the link between the same businesses in different years (link 8).

The ASIA archive simply contains the link between the same businesses in different years (link 9), in order to track the dynamic of business.

Between the Tax form 770 archive and ASIA and, in a parallel way, between the Annualised E-Mens and ASIA there is a link between the same businesses in a specific period of time (link 1 and link 5). Combining links 1 and 5 it is then possible to link all three archives, highlighting the businesses common to all three sources as well as the businesses common to only two of the three sources (for aspects connected to issues of different scope or to problems of under-coverage of the source). Finally, between the Tax form 770 archive and the Annualised E-Mens archive there is also a link between the same workers in a specific period of time (link 10).

The longitudinal linkage issues

Managing the previous links is not the simplest of operations. Several difficulties may be encountered – for instance, in tracking businesses from one year to the next or from one source to another, in particular for businesses involved in merger-acquisition which may lead to distortions in the analysis of business demographics in terms of births and deaths of businesses. Failure to identify businesses undergoing such processes – which are, incidentally, also relatively frequent in Italy – would risk identifying as distinct some units which are actually characterised by statistical continuity (insofar as they are economically and structurally stable, although having different legal forms) and vice versa, with clearly negative effects on the determination of real employment flows. In particular, in the case of larger businesses, the consequences in terms of erroneous employment flows would be particularly significant. Of lesser but certainly not insignificant impact is the problem of correctly identifying worker's fiscal codes. A missing or incorrect fiscal code may in fact have significant consequences in the process of reconstructing an individual's working career.

4. Final remarks

This paper shows that in the recent years some advancements in the production of a longitudinal LEED suitable for analysis of labour market dynamics has been reached by the Italian Institute of Statistics. Using administrative data to design this official LEED has been a natural choice considered the experience of the Italian NSI in exploiting administrative files for statistical purposes. A lot of work has been carried out to overcome a number of weaknesses related to the administrative nature of the data. That has implied a very deep analysis of the information contents of the sources used, the study and monitoring of the changes of the administrative rules that have caused some discontinuities over time in information migrated to another administrative source. The retrieving of the statistics variables has been a hard and painstaking task, requiring an accurate comparison between the definitions and classifications provided by tax legislation with those provided by official statistics. The data integration has been carried out through unit record linkage between sources and over time. The number of different links (ten) gives an idea of the record linkage procedure complexity, also because of the supplementary work needed for harmonising different administrative businesses identifiers in the two administrative sources used. The longitudinal linking of employers and employee data has been tested and the first results are reported in another paper of the authors (Colace, Congia and Rizzi 2009). The very earliest analyses on employment tenure show the longitudinal potential of such a LEED and the results confirm a number of basic and well known evidences on the labour market dynamics. The longitudinal quality of the LEED has to be assessed more because some issues related to the tracking of workers and businesses from one year to the next have to be handled as merger or split processes or the correct identification of multiple job holders avoiding double-counting. Further efforts have to be done to make these official longitudinal LEED reliable and accessible to researchers ensuring the confidentiality of data. However, to achieve this ambitious task the Italian NSI is needed to invest more in both human and IT resources.

References

Abowd, J. M., Kramarz, F. and Woodcock, S. (2008). Econometric Analyses of Linked Employer–Employee Data, in Matyas, Sevestre (eds.) *The Econometrics of Panel Data*, Berlin Heidelberg: Springer-Verlag, pp. 727-760.

Abowd, J. M. and Kramarz, F. (1999). The Analysis of Labor Markets using Matched Employer-Employee Data, in O. Ashenfelter & D. Card (eds.) *Handbook of Labor Economics*, edition 1, Volume 3, Number 3, Elsevier.

Abowd, J. M., Haltiwanger, J. and Lane, J. (2004). Integrated Longitudinal Employer-Employee Data for the United States, *American Economic Review*, Volume 94, Issue 2, pp. 224-229.

Calzaroni, M., Congia, M.C., Montebugnoli, M.E., Rizzi, R. and Tronti, L. (2008). ‚Net wages and tax/contribution wedge of employees from administrative data, paper presented at the 29th Annual Conference of the International Working Party on Labour Market Segmentation, Porto, Portugal.

Colace, A., Congia, M.C. and Rizzi, R. (2009). A Longitudinal Analysis on Italian Employees using Administrative Sources, paper presented at the 30th Annual Conference of the International Working Party on Labour Market Segmentation, Tampere, Finland.

Contini, B., and Trivellato, U. (2005). *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*, Bologna: Il Mulino.

Desai, T. (2008). A Guide to Linked Employer-Employee Data Sources in the EU and Beyond: 1st Edition" in final report of the Project for the European Commission DG Employment "European Labour Market Analysis using Firm-level Panel Data and linked Employer-Employee Data".

Padapopolous, T. (2004). Linked- Employer-Employee Data (LEED), in *Labour Market Statistics 2003*, Statistics New Zealand.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Are Prices Surveys Sample Designs Robust to Aging Weights? A Simulation Study

Zdenek Patak, and Daniele Toninelli[1]

Abstract

The importance of the prices' study has increased quickly in the last years: several national statistical agencies started developing projects based on longitudinal studies to measure the movements of the prices of products and services. Many methodological issues are subject of intensive research to improve the quality of the whole indexes' production process. The target of this work is to contribute to improve the quality of the data collection process by studying the temporal evolution of the survey data. Starting from a simulated population, a number of commonly used sample selection methods are compared underlining their relative efficiency and evaluating how and how much the change in the size measure over time affects the estimates and the bias of the results.

Key Words: Prices Index, Sampling Weights, Simulation Study, Sampling Methods, Index Bias.

1. Introduction

With the release of the Boskin Report (Boskin *et al.*, 1996) on the state of U.S. price indexes, there has been an intense debate on ways to improve their quality. Major strides have been made in assessing the applicability of standard business survey methodologies in an area that has long been dominated by judgment and subject-matter expertise. In the U.S., in particular, the coverage, collection and sampling processes have experienced major innovations. With the growing importance of the service sector, Statistics Canada is developing a new set of Service Producer Price Indexes that may incorporate many of these innovations. Statistics Canada is also using this opportunity to investigate issues such as the impact of sampling design on the precision of an index, and the effect of prolonged use of aging weights on reliability. By means of a simulation study, this paper compares a number of probability sampling strategies, such as simple random sampling without replacement (SRSWOR), and several probability proportional to size (PPS) methods, commonly used in practice in the context of estimating a price index. The study was done to assess the robustness of various designs to aging samples and basket (a set of goods or services whose prices are being monitored over time) weights in terms of bias and mean squared error of the estimated index. The target population for the simulation study was generated based on parameters computed from the Canadian Wholesale Services Producer Price Index survey.

1.2 Price index computation

The first step in obtaining the final index is the computation of the *elemental index* at the establishment level (lowest level of aggregation). Let p_{hik}^{t} be the profit margin observed at time t for product k (k = 1, …, m_{hi}) of establishment i belonging to group h, where m_{hi} is the number of observed products for establishment i (m_{hi} = 1, 2, 3). The elemental index $\left(P_{hi}^{t/t-1}\right)$ between times t and t-1 is computed as the geometric mean of the profit margin ratios,

$$P_{hi}^{t/t-1} = \left[\prod_{k=1}^{m_{hi}} \left(\frac{p_{hik}^{t}}{p_{hik}^{t-1}} \right) \right]^{1/m_{hi}} . \qquad 1.2\text{-}1$$

This index is also called *Jevons index*[2]. The *aggregated index* at the stratum level of aggregation (stratum h) is a weighted average of elemental indexes. It is computed using the following Laspeyres formula[3],

[1]Zdenek Patak, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (Zdenek.Patak@statcan.gc.ca); Daniele Toninelli, University of Bergamo, via dei Caniana 2, Bergamo, Italy, 24127 (daniele.toninelli@unibg.it).

[2] See the *Producer Price Index Manual* (2004) and the *Consumer Price Index reference paper* (1995).

[3] For further information about the Laspeyres index see: the *Producer Price Index Manual* (2004), the *Consumer Price Index reference paper* (1995), and Laspeyres (1884). For more details about the use of the Laspeyres index in the SPPI project context, see also Patak and Lothian (2007).

$$\hat{I}_h^{t/t-1} = \frac{\sum_{i=1}^{n_h} \left[z_{hi}^{t-1} \cdot w_{hi}^{t-1} \cdot P_{hi}^{(t/t-1)} \right]}{\sum_{i=1}^{n_h} \left(z_{hi}^{t-1} \cdot w_{hi}^{t-1} \right)}, \qquad 1.2\text{-}2$$

where h is stratum, n_h is the number of units of stratum h, z_{hi}^{t-1} is *economic weight* of unit i (belonging to the stratum h) at time t-1, and w_{hi}^{t-1} is *sampling weight* of unit i at time t-1. More generally speaking, for a year t, the economic weight z_{hi}^{t} of an establishment i belonging to the group of units h is computed as the ratio of the revenue of the considered establishment $\left(R_{hi}^{t}\right)$ to the total revenue of all the establishments belonging to the same group $\left(R_h^t\right)$: $z_{hi}^{t} = R_{hi}^{t} / R_h^t$. The sampling weight w_{hi}^{t} of a unit i for the year t is computed as the inverse of the probability of selection, $w_{hi}^{t} = 1/\pi_{hi}^{t}$. The probability of selection $\left(\pi_{hi}^{t}\right)$ of unit i is the probability of being included in the selected sample; it varies according to the sample selection methodology. In the following paragraph (1.3), we will test if sampling and economic weights vary over time. We will also try to answer the question: how often should weights be updated to get a value of the estimated index as close as possible to the value of the actual index?

1.3 How weights change over time

Both sampling and economic weights appear in the aggregated index formula (1.2-2), so potentially they both have an influence on the estimates. In fact, if a price index has to be computed, the system of weights to use has to be decided and this can affect the quality of the final result. The best choice for computing a price index for a certain year would be to use that year's weights: this would provide what we call an *actual index*; however this is not always possible as the actual weights are usually available with a delay. This means that weights that are not current are often used for computing a price index. Furthermore, it is important to understand if and how the weights vary over time, and what is the impact of their change on the bias and on the variance of the index. This paragraph studies the first aspects, while the impact of changing weights on the index is discussed in par. 3.1.

An analysis of the sampling (w_i) and economic weights (z_i) for the period from 2004 to 2009 showed that in the first three years the sampling weights remained relatively unchanged; the percentage variations stayed within 10%. The year over year variation started changing appreciably after four years, +73.6% between 2007 and 2008 and +46.1% between 2008 and 2009. If the first variation could be explained by a change in how the BR was structured, the same reason does not explain the 2008-09 change. The average annual change in sampling weights was +15.2%. Furthermore, the changes in sampling weights in the short term did not seem to follow a predictable path, i.e., the trend was negative between 2004 and 2005 and between 2006 and 2007, positive in the other cases. If we consider cumulative variations, there was an increase of 145.9% in the five years beginning in 2005 in comparison with 2004; the average cumulative change over the same period was 29.2%. This broad change in the sampling weights can understandably have a strong effect on the quality of the estimates if we are using extremely outdated weights.
The economic weights showed a similar pattern over the same six years (2004-2009), but with smaller peaks of variation compared to the sampling weights; the maximum year-over-year change was 39.8% between 2007 and 2008 (the impact of switching to the new BR is underlined again). The average annual variation was bigger (20.1%) than the one observed for the sampling weights.

The data show that sampling weights vary more in the long run, while economic weights vary more on an annual basis. Having said that, even if the biggest change (2007-08) is in part explained by an external event (that does not explain completely the 2008-09 change), we can conclude that these variations can potentially have a considerable impact on the precision of the final estimates. They can also reinforce themselves, given that both sampling and economic weights appear in the index formula. The next paragraph will introduce the methodology we used to evaluate the impact of changing weights (that is of weights getting aged) on the estimates of the index in terms of both bias and variance.

2. Methodology

Our study is based on a simulated population of price relatives. They were generated using parameters computed using price relatives observed in the 2006 to 2008 waves of the Statistics Canada's wholesale survey. The simulated relatives were then associated to the real frame population of establishments used to select the wholesale survey samples. Structural and fiscal data for each unit were obtained from the BR. Once the simulated frame population was generated, we selected samples using the following sampling designs: *Simple Random Sampling* (SRS), *Probability Proportional to Size* sampling (PPS), and *Sequential Poisson Sampling* (SPS). The selection of samples was carried out with a bootstrap-like methodology, selecting a maximum of

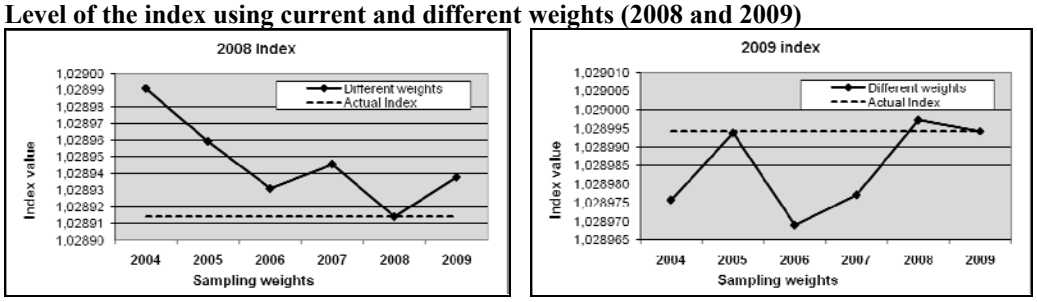
5,000 samples of 1,000 units each, in order to obtain convergence of the index' variance. For each of the selected samples we computed different versions of the index, varying the sampling weights (that is using sampling weights coming from all the BR' variables' considered years). The actual index was computed for each year of the simulated population (from 2004 to 2009) using the current period weights; this means that, for a year x (x = 2004, 2005, …, 2009), the system of weights of the same year (x) was used in computing the index. Furthermore, other *simulated indexes* were computed for each year, using weights coming from the other years: for example, to compute a simulated index of a year x, the weights of a year y (y = 2004, 2005, …, 2009, $y \neq x$) were used. For each version of the simulated index we also computed the variance and the bias obtained by comparing the simulated index itself with its actual value (that is the estimate obtained using current weights, i.e. weights belonging to the same year of the index). Thus, this bias stems from the use of not-updated weights. In this way, for each year's index we can evaluate both the bias of the estimates and the variance that we obtain using aging weights: this all gives us the evaluation of the estimates' precision. Furthermore, we can compare the results obtained for different sampling selection schemes, measuring their relative effect on the index bias as well as their robustness. The following part of the paper (par. 3.1) will focus, in particular, on the study of the effect of sampling weights (our experimental factor) on the estimates of the index. In the par. 3.2 we will discuss some results (in terms of both bias and variability) obtained with the implemented sampling selection methods. The comparative evaluation of such methods will follow (par. 3.3).

3. Some results

3.1 Level and variance of the index

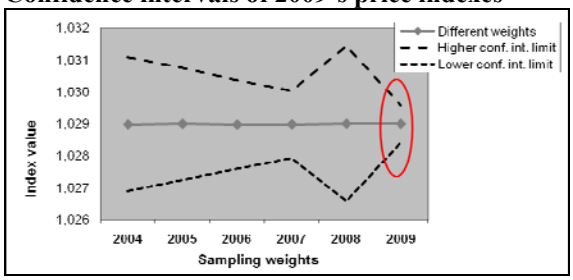
In this paragraph we evaluate the impact of sampling weights on the level of the price indexes. Using the whole frame simulated dataset, we computed, for each year (from 2004 to 2009), the actual index and the simulated indexes. In the graphs 3.1-1/2 the results about the 2008 index (3.1-1, on the left) and about the 2009 index (3.1-2, on the right) are shown. In both the graphs the broken line represents the actual index (that is the index computed using current sampling weights), while the black line shows the estimates of the simulated indexes (computed using not-current sampling weights).

Graph 3.1-1/2
Level of the index using current and different weights (2008 and 2009)



The most common situation (observed in 2005, 2006 and 2007) is similar to the one shown for the 2008 index (graph 3.1-1): the more we use weights that are far from the current year, the more the bias of the simulated index increases, in comparison with its actual value (2008 index computed with 2008's weights). However, sometimes (e.g. in 2004 and in 2009) we can also obtain a situation similar to the one underlined in the 2009's graph (3.1-2): even if we are using aged weights (i.e. weights of 2005) we can obtain estimates close to the value of the actual index. The graph 3.1-3 shows the confidence intervals built around the 2009 index values, obtained using different sampling weights. The depicted situation is common to the majority of the cases (2008 excluded): when we are using current weights, the computed index is more precise (that is, the confidence interval is thinner).

Graph 3.1-3
Confidence intervals of 2009's price indexes

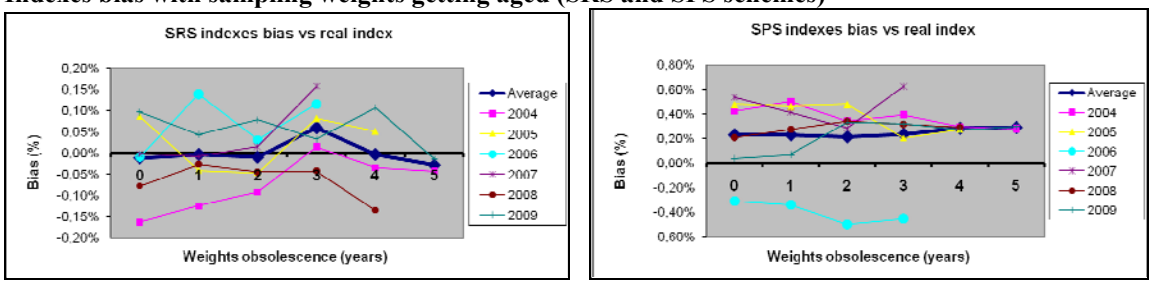


All this seen, it is possible to conclude that we do not always obtain extremely biased estimates when we do not use current weights, but usually these estimates are less precise than the ones obtained using updated weights. After analyzing what happens when the sample (and the sampling weights) become dated, in the next par. 3.2 we study the effect of the sampling methodology on the precision of the index.

3.2 Bias and standard deviation of the estimates

The results shown in this section are referred to a maximum number of 500 bootstrap iterations in selecting samples of 1,000 units from the simulated frame population. For the selection of the units the following methods were used: SRS, PPS and SPS. Seen that the results obtained with the PPS are similar to the one obtained with the SPS, only these last will be discussed in this paper. In the graphs 3.2-1/2 the bias obtained with the SRS (graph 3.2-1, on the left) and SPS (graph 3.2-2, on the right) is shown. In particular, we show the behaviour of the index while sampling weights are getting dated: on the *x* axis the distance (from 0 to 5 years) of the used weights from the year object of analysis is represented (so "0" means that we are using current weights).

Graph 3.2-1/2
Indexes bias with sampling weights getting aged (SRS and SPS schemes)



In the graph 3.2-1 the thick line represents the average path of all indexes while sampling weights are getting dated: the bias (and its variance) seems to increase more as we use weights that are far from the current year. However, we notice that the average situation is only the compensation of completely different paths shown from the thinner lines (referred to each single index). In fact, if we take a look at the details, we notice that for some indexes (e.g. 2007) if we use older weights, we obtain a bigger bias; but for others (e.g. 2004) the bias decreases more when we are using old weights. Observing the 2008 index, we notice that the worst performance is obtained using weights that are 4 years old, but we also notice that the bias of the index is reduced if we use 1, 2 or 3 years old weights, instead of using current weights. On the contrary, the most precise value for the 2009 index is obtained using 5 years old weights. The graph 3.2-2 shows the results obtained with the SPS selection scheme. At an average level (thick line) the bias is bigger than the one obtained with SRS (note the different scales of the two graphs). This is because the units are selected with a probability proportional to size and so the system of weights gives more importance to the bigger units that are usually characterized by bigger values of the index (Toninelli, 2008). What is in common within the two sampling methods is that we obtain a more biased average path when we use older weights. But, again, this is the result of compensations of the completely different paths of the various indexes. For some of them (e.g. 2006, 2008, and 2009) the bias is extremely reduced using current weights, while for some others (e.g. 2004 and 2005) the bias increases with older weights; other indexes (like the 2007's one) have more irregular paths. Seeing this, it is clear that, using different sampling schemes, we obtain very irregular paths of the indexes that often have not much in common with each other. For this reason, it is not possible to define a precise general consequence on the index's estimates, while weights are becoming dated. This is probably because each index considers a changing systems of weights that is referred to various economical sectors; each one of them can influence, more or less, the general structure and the distribution of weights and, then, the index estimates and their bias. In the next par. 3.3 the sampling selection schemes will be compared in term of bias and standard deviation.

3.3 Comparing sampling schemes

In the following analysis we summarize by year the effect of using current rather than dated weights. A bootstrap-like selection methodology was used, selecting 5,000 samples of 1,000 units each. The evaluation of the estimates, for each considered year, is made in terms of bias and standard deviation: they give us an idea of the precision of the estimates. The bias is obtained comparing the computed simulated estimates of the index (an average of the estimates given by the 5,000 selected samples) with the corresponding actual index. The standard deviation is computed for each index obtained through the 5,000 bootstrap-selected samples. The average of these sample values is then compared to the standard deviation of the actual index. Using these criteria we evaluated, by sampling selection scheme, the percentage reduction of the bias and of the standard deviation that we obtain using current weights (tables 3.3-1/2). In the tables the cells (year/sampling scheme) where the updated weights

giving a better performance are underlined in grey. The percentage reduction of the bias using updated weights is shown in table 3.3-1 (left). SRS performs better, using updated sampling weights, in two cases out of six (2005: -27.3%; 2008: -39.7%); in a third case (2007: -0,8%) using updated rather than not-updated weights we obtain round the same level of bias; in the other cases (2004, 2006 and 2009) we obtain a bigger bias (more than +47%) using updated weights; the same happens at an average level (average of the six years: +28.4%). PPS and SPS give better results using updated sampling weights in the same years (2006, 2008 and 2009), that is, in three cases out of six. Also, on average the bias is reduced by, respectively, 10.7% and 6.2%, in comparison to the bias obtained using not updated weights.

Table 3.3-1/2
Updated Vs not-updated weights estimates (bias and standard deviation percentage reduction)

	BIAS		
Index (year)	SRS	PPS	SPS
2004	56.4	5.8	17.3
2005	-39.7	17.7	27.8
2006	47.5	-43.1	-29.1
2007	-0.8	29.5	30.9
2008	-27.3	-32.6	-28.6
2009	76.3	-65.3	-83.2
Average	*28.4*	*-10.7*	*-6.2*

	ST. DEV.		
Index (year)	SRS	PPS	SPS
2004	0.2	-10.8	-15.7
2005	-3.0	-25.1	-25.9
2006	-0.5	-12.3	-17.9
2007	5.1	-14.8	-16.1
2008	-8.0	-29.3	-44.0
2009	1.5	43.8	49.1
Average	*0.1*	*-12.6*	*-14.2*

The percentage reduction in the standard deviation (in comparison with the actual index's one) using updated sampling weights rather than not updated weights is shown in table 3.3-2 (right). With SRS we usually don't get considerably better results when we use updated weights: on average, we obtain a variability that is almost the same. In fact, the difference between the standard deviation obtained with updated rather than not updated weights is reduced to 0.1% only. Anyway, for the six considered indexes (2004 to 2009) we obtain a slightly smaller standard deviation, using updated weights, in three cases on six (2005, 2006 and 2008) while for other years the standard deviation is smaller with not-updated weights. On the other hand, with the probability proportional to size selection schemes (PPS and SPS) we obtain better results in most of the cases (five out of six years) if we are using updated weights. Moreover, the standard deviation level is strongly reduced (from about -10% to -44%); over the six years PPS leads to an average reduction of -12.6%, and SPS performs even better, resulting in an average reduction of -14.2%. For 2009 only, the standard deviation increases considerably using updated rather than not updated weights (PPS: +43.8%; SPS: +49.1%).

3.3 Final remarks

Studying the variability of the weights over time, we found that sampling weights vary more in the long term, while economic weights vary more on an annual basis. Moreover, the over/underestimation of the actual index and its variance are highly different year by year: it is not possible, on average, to identify a clear trend in the evolution of the weights. Further research is needed as soon as new BR data become available. Moreover, a deeper study, taking into consideration the different economical sectors, could be useful to identify paths that may escape detection when studied globally. This will allow us not only to understand if we can define a more regular evolution of the weights of a certain economic area over time, but it will also help us to forecast their future change. To understand how often weights should be updated, we compared the values of the actual index (computed with current weights) and the values of the simulated indexes (computed with not-current weights). We also studied the path of indexes while weights are becoming dated. Considering the SRS results, we noticed that, even if at a single level the paths of the indexes are really different from one another, on average the bias of the index (and its variance) is usually reduced using weights that are no older than 2 years. For proportional to size sampling schemes, we usually obtain a bigger bias (seen the bigger importance given to the biggest units indexes) and it slightly increases too, while weights become dated. But, again, this is an average result of completely different paths. We then compared some sampling selection schemes in terms of the percentage reduction in the bias and standard deviation obtained using updated/current weights. If compared to SRS, the probability proportional to size sampling schemes give us better results in terms of bias (3 cases out of 6 and on average). Moreover, when we are using updated weights, the standard deviation of the index is reduced too (5 cases out of 6 and on average), and this means that the index is more precise. On the contrary, SRS method does not usually work better with the current weights: it gives better results more "randomly". So, we should always use updated weights. But the main issue in choosing a weights' system is that current weights are usually not available for a while, so how often should the weights be updated? As often as possible, of course. Ideally, they should be updated annually (or as soon as new data are available); in fact, sampling weights older than two years are not recommended (see graphs 3.2-1/2). If weights are updated less frequently, one could sometimes be lucky, using a system of weights that can be similar to the current one. This is because the weights do not follow a linear trend. But it is not possible to identify the system of weights more similar to a certain year (i.e. the current

year) if we do not know the corresponding weights. Furthermore, the variance of the index is higher if weights are not updated resulting in an index that is less precise. We focused the second part of our work on the change in the sampling weights, in particular, so further studies are needed to measure the impact of economic weights on index estimates and the combined impact of both sampling and economic weights, eventually taking into consideration the different economic areas.

References

Boskin, M.J., Dulberger, E., Gordon, R., Griliches, Z. and Jorgenson, D. (1996). Toward a More Accurate Measure of the Cost of Living. *Final Report to the Senate Finance Committee* from the Advisory Commission To Study The Consumer Price Index. December 4. http://www.ssa.gov/history/reports/boskinrpt.html.

Consumer Price Index reference paper (1995). Statistics Canada, Price Division. Ottawa, Canada: Minister of Industry.

Laspeyres, E. (1884). Hamburger Warenprise 1851-1860 un die californisch-australicher soldent dukung seit 1848, Ein beitrag zur dehre von des Geldentwertung, in *Jarbüucher fur Nationalöokonomie.*

Patak, Z. and Lothian, J. (2007). *Enhancing the Quality of Price Indexes - A Sampling Perspective*. Ottawa, Canada: Statistics Canada.

Producer Price Index Manual - Theory and Practice (2004). Washington, D.C.: International Monetary Fund (http://www.imf.org).

Toninelli, D. (2008). Survey Techniques: an application to prices data for the computation of price indexes, unpublished PhD thesis, Bergamo, Italy: University of Bergamo.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Adding a Longitudinal Component to Statistics Canada's Agriculture Tax Data Program

Terri Blanchard and Peter Xiao[1]

Abstract

The Agriculture Tax Data Program (TDP) at Statistics Canada is primarily designed to produce cross-sectional estimates on financial variables. The administrative data are obtained from tax forms sent to the Canadian Revenue Agency (CRA) by unincorporated, incorporated and communal farms in electronic or paper format. A longitudinal component was added to the TDP starting with tax year 2001. The goal of the new component is to follow changes to individual farms over time. The 2001 tax year panel has been followed on a yearly basis. Annual panels of cohorts have been created starting in tax year 2006 and have also been followed yearly.

This paper describes the longitudinal component of the TDP, the imputation strategy designed specifically for the longitudinal units and the weighting methodology.

Key Words: Longitudinal Data, Administrative Records, Tax Data.

1. Introduction

The Agriculture Tax Data Program (TDP) at Statistics Canada produces cross-sectional estimates of financial variables such as average operating revenues and expenses, net operating income and off-farm income for Canadian farm operators, farms and farm families. The TDP obtains data from the Canadian Revenue Agency (CRA) through three tax data sources: unincorporated farm operators (who file a T1 form), incorporated farms (T2 filers), and communal farms (T3 filers). Starting with the 2001 tax year, a longitudinal component was added to the TDP with the goal of creating a research database that would be used to follow changes within individual farms over time. For example, the database could be used to look at whether strong or weak financial performance is a transitory or permanent feature specific to some farm types. The 2001 cohort is followed on a yearly basis and a longitudinal database currently exists with 6 years of data for the 2001 cohort. Two additional cohorts starting with tax years 2006 and 2007 have been created.

This paper will focus on the longitudinal component of the TDP, in particular the creation of the longitudinal database for the 2001 cohort. Section 2 gives an overview of the sample design. Section 3 focuses on the imputation strategy developed to treat non-response within the longitudinal cohort and the calculation of the estimation weights is described in section 4.

2. Sample design

The Agriculture TDP has produced cross-sectional tax-based estimates annually and therefore the sample design and processing systems are well established. When the longitudinal component was introduced in 2001, it was decided to maximize the overlap between the longitudinal cohort and the cross-sectional sample in order to use the existing infrastructure for sampling and data processing as much as possible. The sample for the first year of the longitudinal cohort is actually the same as the cross-sectional sample for that year. Therefore, the sample design described in the following sections is the sample design for both the first year of the longitudinal sample and the yearly cross-sectional samples.

Even though separate samples are drawn and different data fields are collected for the T1, T2 and T3 filers, the data for the three types of filers are consolidated into a common set of variables for the production of estimates. The estimates produced by the TDP use the combined data from the T1, T2 and T3 filers.

The agriculture sector reports their tax data using variables that are different from the variables used by other business sectors to report their tax data. Within Statistics Canada, the Agriculture Division is fully responsible for the in-depth editing,

[1]Terri Blanchard, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, terri.blanchard@statcan.gc.ca,; Peter Xiao, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, peter.xiao@statcan.gc.ca

imputation and analysis for the production of estimates related to the agriculture sector. Other sectors using tax data from CRA have their data fully cleaned and processed through the Tax Data Division at Statistics Canada.

A census of the approximately 300 filers in the T3 universe is taken every year. The rest of this paper will focus on the T1 and T2 filers only.

2.1 Sample design for T1 portion

The target population for the T1 portion of the longitudinal sample is all individuals reporting farming income from self-employment (T1 filers) in the first year of the cohort. Data are obtained from CRA through tax forms remitted by the individuals. CRA receives data for the T1 filers reporting farming income in three formats: paper forms mailed to CRA, electronically submitted forms and AgriStability/AgriInvest[2] forms. Paper forms mailed to CRA include forms completed by hand and those completed using a tax software program which contain a barcode on the first page. Paper forms that have a barcode (barcode filers) are treated as electronic filers by the TDP. CRA sends Statistics Canada the data for all electronically submitted T1 forms, all barcode forms, all AgriStability/AgriInvest forms and a sample of the paper forms that are not barcode forms. The sample selection of all types of forms and the data capture of the paper forms is done by CRA according to specifications provided by the TDP.

A sample is selected from the electronic and AgriStability/AgriInvest filers even though the TDP receives data for the complete universe of these filer types. The sampling is done because it would be costly and time consuming to edit, correct and impute erroneous or incomplete data for all electronic and AgriStability/AgriInvest filers using the existing system. In order to maximize the use of the electronic and AgriStability/AgriInvest filer data, the data are put through a series of automated edits to assess the quality of the data. Units that pass all of the edits and therefore do not require manual intervention are considered clean. The clean units that are not randomly selected are added to the cross-sectional sample. These additional units are not part of the longitudinal cohort.

The T1 population is stratified by Canada's 10 provinces and three territories and then into a maximum of eight size strata based on Gross Farm Income (GFI) using the cumulative square root-F method. The sampling rates per stratum are determined by allocating a predetermined sample size to the provinces/territories using square-root proportional allocation and then to the size strata within each province/territory using the Neyman allocation method. Statistics Canada's generalized sampling system, GSAM, is used to perform the stratification and allocation. For more information on GSAM see Statistics Canada (2006). A census is taken in Newfoundland and Labrador and the three territories due to the small population sizes. The remaining nine provinces are sampled.

The Poisson sampling technique is used to select the sample. Each T1 tax return that is eligible for sample selection is assigned a pseudo-random number, based on Social Insurance Number, using a hash function that returns a value between 0 and 999. The sampling rate within each stratum is also converted to number between 0 and 999 called the hash upper limit number. A unit is selected for the sample if its pseudo-random number is less than or equal to the hash upper limit number for the stratum in which the unit belongs. This method is used so that CRA can capture selected paper forms as they are received and so that the TDP can process the data from the sampled units as they are received from CRA. Waiting until forms for the complete universe are received before sampling and processing the data would mean that the estimates would not be published in a timely manner.

2.2 Sample design for T2 portion

The target population for the T2 portion of the longitudinal sample is comprised of all incorporated farming operations (T2 filers). Data are obtained from CRA through tax forms remitted by the corporations to CRA. Statistics Canada receives data from CRA for the complete T2 universe. A sample is drawn for the T2 filers because the data received is not always complete and sometimes requires intensive manual data correction which is both time consuming and costly.

The T2 population is stratified into Canada's 10 provinces and three territories and then into 11 farm types. The 11 farm types are oilseed and grain, potato, other vegetable, fruit and tree nut, greenhouse and nursery, other crops, beef cattle, dairy cattle, hog and pig, poultry and egg production and other animal production. Each geography/farm type stratum is further stratified into a maximum of four size strata based on sales using the cumulative square root-F method. The sampling rates are determined by first allocating a predetermined sample size to the geography/farm type strata using square-root proportional allocation. Secondly, the sample within each of these geography/farm type strata is allocated to the size strata using the

[2] The AgriStability/AgriInvest program was formerly known as the Canadian Agriculture Income Stabilization (CAIS) program and the Net Income Stabilization Account (NISA) program.

Neyman allocation method. The stratification and allocation are done using GSAM, the same generalized system that was used for the T1 filers. A census is taken in the four Atlantic provinces and the three territories due to their small population sizes. The remaining six provinces are sampled. As with the T1 filers, the Poisson sampling method is used. The pseudo-random number generated for each unit using a hash function is based on the unit's business number.

2.3 Following the longitudinal cohort units

As mentioned in previous sections, the longitudinal sample for the first year of a cohort is all randomly selected units of the cross-sectional component of the same year. Once these units are identified, they are to be followed on a yearly basis for 10 years. In order to ensure that Statistics Canada receives data from CRA for the longitudinal units of a cohort in subsequent years, the longitudinal units are specified for inclusion in the cross-sectional sample files even if they are not randomly selected. The hash function used to calculate a unit's pseudo-random number for sampling purposes maximizes the overlap among the randomly selected units from one year to the next. This stability in the sample minimizes the number of longitudinal units that are added to the cross-sectional sample files.

The 2001 cohort contains 26,086 T1 filers, 11,674 T2 filers and 287 T3 filers. The 2006 cohort contains 20,345 T1 filers, 7,523 T2 filers and approximately 300 T3 filers. The 2007 cohort contains 19,856 T1 filers, 7,803 T2 filers and 300 T3 filers. The decrease in the sample size for T2 filers in tax years 2006 and beyond is due to a reduction in the cross-sectional sample sizes starting with tax year 2006.

3. Treatment of non-response

As is the case with surveys, the longitudinal cohort experiences both partial and total non-response. Partial non-response and total non-response are dealt with in different ways as described in the following sections.

3.1 Partial non-response

Partial non-response can be due to the fact that it is not mandatory for tax filers to complete all fields on the CRA forms. For example some filers may provide only generic information such as totals instead of providing detailed information on the components of these totals. Partial non-response was dealt with through deterministic, historical, donor and manual imputation. The donor imputation used the nearest-neighbour approach and was used mainly to allocate total values to a more detailed breakdown when details are not provided. The detailed breakdown of the donor record was applied to the recipient record. All partial imputation for the longitudinal cohort units was done as part of the processing system used to produce the cross-sectional estimates.

3.2 Total non-response

Total non-response can occur for a variety of reasons such as when a unit did not file a tax return for a particular year or when the unit filed a tax return but it was received after the data processing by the TDP was complete. Units that did not file a tax return for a particular year did not appear in the universe file and were coded as non-farm units. For the longitudinal database, data was imputed for units that appeared on the universe file but whose data was received after the TDP processing. For cross-sectional estimation, the late filer units are dealt with through re-weighting.

The majority of the missing units were imputed using the nearest neighbour donor imputation method. The Mass Imputation module of Banff, Statistics Canada's generalized edit and imputation system, was used for the donor imputation. For more information on Banff, see Statistics Canada (2007).

The sections below focus on the imputation strategy used for the 2001 cohort data for tax years 2001 to 2005 which were all imputed at the same time and for tax year 2006 which was imputed later.

3.2.1 Donor imputation strategy for total non-response for 2001-2005

The data for tax years 2001 to 2005 of the 2001 longitudinal cohort were imputed as a group because the imputation was not performed until after data for the 2005 tax year was received. The imputation strategy was developed to take advantage of the fact that five years of data was available. Having five years of data meant that the response pattern over the five years could be used to create imputation groups. In addition, there were a large number of matching variables that could be used to help find a good donor.

The imputation strategy used the most recent year in which the recipient reported data, prior to the missing year(s), to identify a donor whose data could be used to impute for the missing year(s). The same donor was used to impute all consecutive missing years. This provided consistency in the imputed data. The donor's data was adjusted based on a common variable before being copied to the recipient in order to bring the donor data in line with the existing recipient data.

The first step in the imputation process was to group both recipients and donors into groups based on their response patterns from 2001 to 2005. For example, recipients that had data in 2001, 2002 and 2003 but not in 2004 and 2005 were grouped with potential donors that had data in 2003, 2004 and 2005. Data from 2003 for both the recipient and donors were used as matching variables to identify a donor. There were 10 response pattern groups created based on the response patterns from 2001 to 2005. See Blanchard (2008) for more detailed information on the response pattern groups.

Within each response pattern group, imputation classes were created. The imputation classes are groups of homogeneous records within which donors will be found for recipients. The first run of donor imputation was done using province by farm type for the imputation classes. The same farm type groupings were used for all provinces. For the second run of donor imputation, the imputation classes were the farm type groupings only.

The matching variables used to identify a donor were total operating revenue, total operating expenses and farm type revenue from the most recent year the recipient reported data as well as GFI (T1 filers) or sales (T2 filers) for the missing year(s). For example, if a recipient reported data in 2003 but was missing data in 2004 and 2005, the total operating revenue, total operating expenses and farm type revenue from 2003 along with the GFI or sales from the universe files for 2004 and 2005 were used as matching variables.

Each imputation class must have had at least 10 donors and the minimum percentage of donors per imputation class was 30%. Records imputed during the first attempt at donor imputation for the longitudinal data were not included in the donor pools for subsequent runs of the donor imputation.

Once a donor was identified, a ratio adjustment based on the total operating revenues of the recipient and the donor was applied to the donor's data before it was copied to the recipient. This was done to bring the donor's data more in line with the existing recipient data.

3.2.2 Donor imputation strategy for 2006

Missing data for the 2006 tax year was imputed after the imputation for the 2001-2005 tax years had been completed. The imputation strategy was very similar to the strategy used to impute the 2001-2005 tax years, however there were some differences.

The most recent year that the recipient had unimputed data prior to 2006 was used to identify a donor whose data could be used to impute for 2006. The earliest year used to find a donor was tax year 2002, which was four years prior to the missing year (2006). This was consistent with the strategy used for the imputation of tax years 2001-2005. Cases where 2001 was the only year that could be used to find a donor were looked at manually to determine if they should be coded as non-farms.

The grouping of recipients and donors into response patter groups differs slightly from the groups created for the 2001to 2005 imputation. There were only 5 response pattern groups created based on the presence of unimputed data from 2001 to 2005. For example, recipients that had unimputed data in 2004, imputed data in 2005 and no data in 2006 were grouped with potential donors that had unimputed data in 2004 and 2006. See Blanchard (2008) for more detailed information on the response pattern groups used for the imputation of tax year 2006. For cases where the recipient's 2005 data was imputed, the donor used in 2005 was not automatically used to impute for 2006. This means that the rule of using the same donor for consecutive missing years only applies to years 2001 to 2005.

The matching variables used to identify a donor were the same as those used for the imputation of tax years 2001 to 2005. Once a donor was identified, a ratio adjustment based on total operating revenues was applied to the donor's data for 2006 before being copied to the recipient in order to bring the donor's data in line with the recipient data.

3.2.3 Historical imputation

Historical imputation with a trend was done for a small number of units that could not be imputed by donor imputation. A record requiring imputation would not have been imputed using donor imputation because there were not enough donors in the imputation class or because of invalid data in the recipient's matching variables that were used to find a donor. In these cases, data for a missing unit were imputed with data from the most recent year that the missing unit had unimputed data. This was

usually the same year that would have been used to find a donor in donor imputation. The historical data were adjusted based on gross farm income for T1 or sales for T2 for the missing year(s).

Donor imputation was performed before historical imputation in order to allow for changes to the farm type assigned to a unit from year to year. Every year, the farm type is assigned based on the commodity whose revenue represents more than 50% of the total revenue for the current year. This means that a unit can change their farm type from year to year if the majority of their revenue comes from different commodities one year to the next.

3.2.4 Imputation rates

In most cases, the imputation rates are quite low; typically below 2% for T1 filers most years and below 6% for T2 filers for most years. The imputation rates for the T1 and T2 filers of the 2001 cohort are shown in Table 3.2.4-1. Any anomalies in the imputation rates are due to data processing issues or changes that existed in that year only. These issues have been addressed and a decrease in the imputation rates where they are high is expected. The table also shows that there is an increase in the percentage of units coded as non-farms because they are not on the universe file. It should be noted that T1 filers may not appear on the universe file because they have incorporated and are now in the T2 universe. It is not possible to trace these cases so they are coded as non-farms for the T1 universe.

Table 3.2.4-1
Imputation rates for 2001 cohort

	T1 Filers (%)			T2 Filers (%)		
Tax Year	**Farms**	**Non-Farms**	**Imputation Rate**	**Farms**	**Non-Farms**	**Imputation Rate**
2001	100.0	…	…	100.0	…	…
2002	94.2	3.8	1.9	96.6	2.0	1.3
2003	86.0	7.5	6.6	91.3	3.3	5.4
2004	87.9	10.2	1.9	88.6	6.5	5.0
2005	84.8	13.5	1.7	81.1	5.9	13.0
2006	81.3	17.3	1.4	81.8	6.9	11.3

Note: … not applicable

4. Longitudinal weighting

The longitudinal weights for the 2001 tax year longitudinal cohort are intended to represent the population in tax year 2001. After the processing of the 2001 tax year data, the T1 and T2 longitudinal populations were created. Population and sample sizes were then determined for each stratum. The longitudinal estimation weight is calculated as the ratio of the population size divided by the sample size. It should be noted that the longitudinal sample contains only the randomly selected units and does not include the clean electronic units and clean AgriStability/AgriInvest units that are not randomly selected. For all 10 years of the 2001 tax year cohort, the estimation weights remain the same because the cohort represents the T1 and T2 populations of the initial tax year. For the 2006 and 2007 cohorts, their estimation weights can be calculated similarly and will remain constant for the subsequent 10 years.

5. Concluding remarks

The longitudinal component of the Agriculture Tax Data Program provides a wealth of information to analysts and researchers concerning changes in the agriculture sector over time. As the 2001 cohort nears completion, there will be 10 year of data available for analysis.

As the number of cohorts being followed increases, the number of longitudinal units being processed within the cross-sectional processing system also increases. In order to reduce the burden on the cross-sectional component, data validation and imputation for the longitudinal units will be done in a more automated fashion separate from the cross-sectional units. Using a more automated data processing system may also allow for the creation of a longitudinal file that contains data for the complete universe of unincorporated farm operators and incorporated farms instead of only a sample of these units.

The development of an automated data validation and imputation process, improvements to the imputation strategy and the creation of a longitudinal file containing data for the complete universe will be focus of future work related to the longitudinal component.

References

Blanchard, T. (2008), Imputation Strategy for the 2001 Cohort of the Agriculture Tax Data Program Longitudinal Data Project, Internal Statistics Canada document.

Statistics Canada (2007), BANFF User's Guide (for v2.01 and 2.02), Internal Statistics Canada document.

Statistics Canada (2006), Generalized Sampling System Version 2.3 User's Guide, Internal Statistics Canada document.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Entrepreneurship: Longitudinal Surveys on Hard-to-Trace Emerging Populations

E.J. Reedy[1]

Abstract

For the last decade, the Ewing Marion Kauffman Foundation has been a major funder of longitudinal research data sets on entrepreneurship. Entrepreneurship measurement spans household and establishment populations, making it an exciting area to explore methodologically. The focus of this paper is on what the Foundation has learned from experiences with the Kauffman Firm Survey (KFS) and the Panel Study on Entrepreneurial Dynamics (PSED) longitudinal surveys with young and emerging firms.

Key Words: Entrepreneurship, New Business, Establishment Survey, Household Survey, Longitudinal.

1. Introduction

1.1 Measuring entrepreneurship

Ewing Kauffman was the prototypical entrepreneur, who started with few resources and grew his firm into a multibillion-dollar company over four decades. He experienced the rewards of entrepreneurship for himself, his family, his work associates, and his community. He believed philanthropy could and should promote entrepreneurship as means to improve lives. The Ewing Marion Kauffman Foundation began its modern operations in 1992 by providing training to existing entrepreneurs and introducing entrepreneurship curricula to students. Beginning in the late 1990s some of our funding began to shift to data collection projects, along with other research, a trend which accelerated under the leadership of our current president in 2003. Today, the Kauffman Foundation supports numerous original data collection projects and efforts to improve statistical infrastructure globally.

But why is research and measurement important to advancing entrepreneurship? Promoting positive policies for the fostering of entrepreneurship remains a key priority in our work. But sound policies need quality research as a basis, and historically, entrepreneurship as an area of research has been largely vanquished to lower quality institutions. To improve research we recognized that a change in the data infrastructure was needed. Beginning in 2003, we increased our spending significantly on data and broadened our engagement with the national and international statistical communities on how best to measure entrepreneurship in a consistent manner.[2]

But measuring entrepreneurship means defining entrepreneurship, and this has been a nebulous area within research. A recent review of the literature characterized many common measurements: startups, age of firm/survival, self-employment, venture capital funding, academics holding a second job, firm size, high-tech firms, and owner characteristics, just to name a few (Reedy and Frazelle, 2009). For our purposes in measurement, rather than arrive at the correct definition of the term entrepreneurship, we focus on the different parts of the concept and encouraging better measurement of each.[3]

1.2 Panel Study on Entrepreneurial Dynamics (PSED)

Our first engagement with panel data was the PSED research program which was designed to enhance the scientific understanding of how people start businesses. The PSED provides valid and reliable data on the process of business formation based on nationally-representative samples of nascent entrepreneurs. PSED I began with screening in 1998-2000 to select a

[1]E.J. Reedy, Ewing Marion Kauffman Foundation, 4801 Rockhill Road, Kansas City, MO, USA, 64110, ereedy@kauffman.org and www.kauffman.org/datamaven.

[2] This paper is too short for a full treatment of our strategy. Interested readers can find more on our Web site and, specifically, in a series of annual publications called the *Kauffman Thoughtbooks* or at www.kauffman.org/research.

[3] Many improvements have been made in definitions. The Organization for Economic Co-operation and Development has led a multi-year process towards harmonization of national statistical office definitions, but that process has largely been focused around government administrative data and derivative statistics.

cohort of 830 with three follow-up interviews. PSED II began with screening in 2005-2006 to identify and follow 1,214 nascent entrepreneurs and is currently funded to continue for a total of six waves. The information obtained includes data on the individual nascent entrepreneurs, the activities undertaken during the startup process, and the characteristics of startup efforts that become new firms.

1.3 Kauffman Firm Survey (KFS)

The Kauffman Foundation created the KFS—the largest longitudinal survey of new businesses in the world—by selecting a cohort of 4,928 firms that began operations in 2004. This cohort is tracked annually and queried on the background of the founders, the sources and amounts of financing, firm strategies and innovations, and outcomes such as sales, profits, and survival. The KFS is in its fifth collection period with eight periods planned.

2. Hurdles faced and lessons learned

2.1 Building the sample

Most entrepreneurship surveys face difficulties in finding a high-quality, representative frame. Even for government institutions conducting such research, tax and other data which could be used for a frame are not typically available because of legal hurdles or timeliness issues. It is even harder for private institutions. For the PSED, the target population was the nascent entrepreneur, as an "individual" or team of individuals, and thus a household frame was chosen. However, given the relatively low incidence of active nascent entrepreneurship[4]—namely individuals showing recent startup behavior, expectations of ownership, and no evidence of a going business—screening households to identify people in the process of starting a business required screening more than 62 thousand households in PSED I and almost 32 thousand households in PSED II.[5] To reduce costs, a commercial survey research firm was used to screen a representative sample of adults to identify those active in creating a business. This same procedure has been used in most international iterations of the PSED.[6] Qualifying individuals were invited to complete a longer interview with about 87 percent of those identified in the screening as active nascent entrepreneurs agreeing to participate. About 60 percent of these consenting nascent entrepreneurs completed the initial hour-long phone interview for the first wave (Reynolds and Curtin, 2008). Reynolds and Curtin have given extensive thought to modifications to the screening protocol, which in hind-sight, could have eliminated cases from the panel who were ultimately determined not to be active nascent entrepreneurs, but suggest that at least a twenty-minute initial interview would be necessary in order to have eliminated 83 cases from PSED I and 277 cases from PSED II (Reynolds and Curtin, 2008). Indeed, for the PSED, it would be nearly impossible to use existing government data to build a survey frame unless questions were added to an existing monthly population survey, such as the Current Population Survey in the United States (Reynolds, 2008).

The Kauffman Firm Survey was designed as a longitudinal survey of new businesses (not individuals), and as such, we determined a commercial source of new business listings which could be further screened was prudent. Dun and Bradstreet (D&B) has the largest database of firm names and addresses of any private organization in the United States. And though we, like many, have mixed feelings on the quality of the raw data, D&B was determined to be the only nationally representative database from which to sample. All data in the final research data set was provided or confirmed by the participants following an extensive screening to determine that 2004 was the birth year (Ballou, Barton, DesRoches, Potter, Zhao, Santos and Sebastian, 2007).

2.2 Funding and administration

The PSED program was very innovative in its initial funding design, perhaps to the detriment of longer-term sustainability. The effort was self-financed initially, without major philanthropic or government support, through the creation of the Entrepreneurial Research Consortium, which consisted of 120 scholars from 34 research institutions who contributed financially to the creation of the data with each getting limited input into the questionnaires. In 1999, the Kauffman Foundation stepped in as a major funder of the PSED effort after the University of Wisconsin Survey Research Laboratory, the vendor used to collect the data, closed unexpectedly. Transitioning the panel data to a new vendor, the University of Michigan's Institute for Social Research, was an expensive process, and the consortium had no budget to continue collection (Curtin and Reynolds, 2007). With the PSED II, the Kauffman Foundation made an initial three-year commitment to funding the project at the University of Michigan. After three years, the researchers involved were successful in getting the Small Business

[4] Found to be about 6 percent of the adult population in the United States (Reynolds and Curtin, 2008).
[5] The significantly larger PSED I sample reflected a procedure designed to increase the number of women and minorities in the nascent entrepreneur cohort (Reynolds and Curtin, 2008).
[6] http://www.kauffman.org/Blogs/DataMaven/August-2009/Overview-of-International-Studies-on-Nascent-Entre.aspx

Administration and the National Science Foundation (NSF) to support a fourth year of data. With continued success in panel maintenance, the principals were awarded two additional years of NSF funding, lengthening the panel to six years in total.

With the Kauffman Firm Survey, project management and administration remained centered at the Foundation. While we received a significant amount of external coaching, inspiration, and practical suggestions for collection (Ballou et al., 2007), the KFS would not have happened without internal championing. From the PSED experience, we learned it was difficult to keep a large group of scholars engaged in the process in an agreeable way, although we too attempted design by committee for too long. We decided to take clear ownership of the project—one of our largest—putting the Kauffman brand on it and operating it directly. Through a competitive bidding process, Mathematica Policy Research was selected to execute the survey. Scott Shane of Case Western Reserve University served as the initial principal investigator, working closely with the Foundation, but Mathematica was funded directly by the Foundation, and leadership on the project has always been centered internally.

Operating the PSED and the KFS under these different models had, I believe, the following implications:

- Academic Ownership. The PSED model of a consortium engaged academics around the world in a rich theoretical conversation and indeed, I believe, along with the parallel but complementary Global Entrepreneurship Monitor[7] program, was key to early international replication.[8] However, perhaps because of the consortium complexities, its bootstrapped financing, and the external shock of a discontinued survey vendor, the avalanche of academic papers (particularly in top-tier publications) which would be expected has been slow to materialize.[9] The PSED I data set remains under-analyzed with many questions untouched by scholars. For the KFS, we realized during the second wave collection that the close Kauffman association could pose challenges, too. Without clear academic ownership and in committing to open data access, getting an initial set of academics to adopt the data for research was going to require significant outreach. We determined that an additional academic principal investigator would be necessary.
- Steep Learning Curve. While we had funded surveys, the Foundation had no staff with survey administration experience before the KFS. As such, many parts of the project had to be modified. For example, the initial concept approved was for a four-year panel and a second cohort of firms to be followed for two years. This design changed several times to allow for an extensive pilot and higher than expected initial costs. But ultimately, the testing time allowed us to achieve high response rates and low annual costs.
- Deepening Knowledge. With a deepening of internal knowledge from the KFS, it became clear that the KFS and PSED were not the only data projects which the Foundation was funding. As such, we centralized supervision of data projects rather than have them diffusely under different program officers. This has helped us to continue supervising investments in these two programs while also helping to leverage smaller projects and direct scholars to the most appropriate data available for their particular lines of research.

2.3 Expected vs. actual sample activities

One of the surprising finding from the PSED-line of research is that nascent entrepreneurship is extremely variable in duration. For many, the process may last less than a month, the smallest unit the survey was able to capture. But for approximately a third of the sample, nascence appears to be a process that can last more than five years, without ever having come to conclusion (Reynolds and Curtin, 2008). This was a longer period than conceptualized. It is also one reason the ideal longitudinal survey using the same sample to examine nascent entrepreneurship, actual entrepreneurs, operating businesses, and eventual business growth, has not been implemented.

[7] The Global Entrepreneurship Monitor (GEM) was developed by Paul Reynolds, one of the PSED principals, and the theoretical architect of both projects. It launched as an index concurrently and achieved quick replication. Additional information is available at http://www.gemconsortium.org/about.aspx?page=re_about_research.

[8] International replications are overviewed at http://www.kauffman.org/Blogs/DataMaven/August-2009/Overview-of-International-Studies-on-Nascent-Entre.aspx.

[9] The PSED principals would likely dispute this. A bibliography is available at http://www.psed.isr.umich.edu/psed/download_document/13.

Table 2.3-1
Status of startup efforts by survey wave for PSED II

	Wave A	Wave B	Wave C	Wave D
Establish new firm		128	164	214
Quit all efforts		231	422	490
Total final outcome		359	586	704
In startup process	1,214	613	472	381
Unknown status		242	156	129
Total cases		1,214	1,214	1,214
Percent final outcome		26.9%	48.3%	57.9%

Source: Curtin and Reynolds (2009)

For the KFS, firm exit after the first wave was smaller than we expected (see table 2.4-1). We suspect this is an artefact of our survey design interacting with two other factors—the D&B frame and short-lived businesses. With the D&B frame, we suspect a bias towards credit-seeking businesses, which might have a higher likelihood of survival, at least initially.[10] With many businesses entering and exiting relatively quickly, if we were able to contact a business in 2005 to talk about their 2004 activities, non-response could be elevated or perhaps our questions in the baseline biased towards populating with an active baseline panel. Differences in attrition, as compared to other government sources, narrowed significantly since the baseline.

2.4 Incentives, mode, and questionnaire design

When the PSED began, industry standards for high-quality collection other than in-person interviews were through mail and phone interviewing. In the PSED I, the principals found that much of the data from the mail questionnaire was less consistently completed, so the mail survey process was dropped for the PSED II, leaving a single mode of collection. For the KFS, a number of possible modes were considered but, ultimately, a dual-mode approach of CATI and Web was selected. By accepting responses from the fixed panel on the phone or online, we hoped to be convenient to entrepreneurs. It should be noted that the CATI interviewers are actually using the same online system to input data that users see on the secure Web site. Panel members were initially hesitant to provide information online, but over time, many panel members voluntarily transitioned to online (see table 2.4-1), a conceptual finding which would seem to indicate many cross-sectional surveys could potentially be turned into panel surveys at a reasonable cost once trust is established.

Table 2.4-1
Kauffman Firm Survey statistics by wave

	Baseline (2004)	First Follow-up (2005)	Second Follow-up (2006)	Third Follow-up (2007)
Survey Conducted	June 2005 – July 2005	June 2006 – January 2007	June 2007 – December 2007	June 2008 – December 2008
Completed Interviews	4,928	3,998	3,390	2,915
Verified Out-of-business	N/A	369	408	540
Phone Completes	77%	41%	37%	35%
Web Completes	23%	59%	63%	65%
Response Rate	43%	89%	82%	78%

Source: Robb, Ballou, DesRoches, Potter, Zhao and Reedy (2009)

In turning to questionnaire design, in the PSED I the consortium funding approach led to a questionnaire which was very long and cumbersome. And many of the questions submitted by academics initially involved in raising funds for the research have subsequently never been analyzed. Thus, for the PSED I, survey design and funding problems are integrated and led to the general conclusion that design by committee is difficult. For the PSED II a smaller group was queried while keeping tight limits on length of the survey.

For the KFS, we received questionnaire input from outside academics but after an extended pilot, approximately half of the questions were discarded because of lack of incidence, difficulty in cognition, or to decrease survey length (Ballou et al, 2007). An annual questionnaire review is undertaken by the principal investigators, and proposals are sought from researchers using

[10] At this point in time this is only a hypothesis, but additional research work is underway to compare the D&B frame to other businesses which showed up in governmental administrative records during the same time period.

the data for suggested question changes or additions. Additionally, the KFS did a large randomized test of several different incentive structures in the pilot. This research (see table 2.4-2) showed a significant increase in participation for a post-payment. Ultimately a $50 post-pay incentive was selected in the baseline and has been maintained for each subsequent interviews.

Table 2.4-2
Kauffman Firm Survey response rate incentive experiment

	Pre-paid Incentive	
Post-paid Incentive	Nothing	$1.00
Nothing	20%	22%
$50.00	28%	30%

Source: Ballou, Zhao, DesRoches and Potter (2007)

One small aside on the incentive and branding that is purely speculative, but in the KFS we have seen at least ten cases a year who contact the Kauffman Foundation through our Web site about the survey. In some cases, they have deleted their password for participation and, in others, they haven't received their incentive check. It appears the panel members are keeping not only the KFS materials which Mathematica provides but also that many respondents remember the Kauffman Foundation name and are coming to our Web properties, which are separate and distinct from the survey domain. To accommodate this, Kauffman associates have been trained to recognize contacts that should be redirected to Mathematica and have had good success in maintaining these cases as a result.

2.5 Beyond building the data

Kauffman had two goals in funding panel surveys: data creation to advance understanding and to encourage broader academic interest. We naively believed that academics needed only data access, and if we supported data, that any data would be adopted swiftly. But public-use data products are not always attractive to an academic audience where disciplinary norms promote proprietary data use in the publication and tenure process. We have concluded that just because the data is collected does not mean it will be used. Also, survey research firms give little emphasis to dissemination, leaving the process to academic principal investigators with little systematic knowledge accumulated.

With the PSED program, the principals on the project have taken the lead in dissemination, targeting the academics that provided input to the project as well as submitting papers to conferences and planning professional development workshops at existing conferences. However, no funds were built into the original budgets for the PSED for dissemination (conferences, research, or other outcomes). While some supplemental funds have been provided by Kauffman and other agencies, in our opinion, not having funding for these dissemination in the budget originally is regrettable (but all too common).

With the KFS, we included dissemination in the budget. We have experimented with a travel grants program and a small grants program, in addition to conference outreach, research webinars, general audience publications, and traveling presentations to selected universities. But in addition, there are nuanced differences in the PSED and KFS dissemination strategy which I believe are important:

- Web Site Registration. Both the PSED and KFS are publicly available for free online download. But with the KFS, registration is required, allowing us to communicate with people using the data and to support them with email updates, trainings, and outreach.
- Versioning. We have created multiple versions of the KFS data set with escalating security requirements and benefits for use. In the KFS Data Enclave version, richer data is available in a secure, remote environment. For example, zip codes in the public-use data file would be a disclosure risk, but in the Enclave we can make this information available. The Enclave also continues to be enhanced with added code contributed from more than 30 researchers.

Lastly, we built broader activities designed to supplement the PSED and KFS projects. The Kauffman Symposiums on Entrepreneurship and Innovation Data provide a venue for potential users of new data and producers of new data to connect. In its inaugural meeting, 150 people attended and 38 data sets were discussed, some of which Kauffman funded but many we had not. Reducing barriers for academics to "test drive" data sets proved very popular and efficient. Additionally, tracking new developments in entrepreneurship and innovation data and blogging on them has helped us to become more knowledgeable about our own and other projects.

References

Ballou, J., Barton, T., DesRoches, D., Potter, F., Zhao, Z., Santos, B. and Sebastian, J. (2007). Kauffman Firm Survey (KFS) Baseline Methodology Report. January 2007. Available at http://ssrn.com/abstract=1024045

Ballou, J., Zhao, Z., DesRoches, D. and Potter, F. (2007). Meeting the Challenges of Designing the Kauffman Firm Survey: Sampling Frame, Definitions, Questionnaire Development, and Respondent Burden (Presentation). June 2007. Available at http://ssrn.com/abstract=1026393

Curtin, R. and Reynolds, P. (2007). Panel Study of Entrepreneurial Dynamics: History of the Research Paradigm. October 2007. http://www.psed.isr.umich.edu/psed/download_document/1

Curtin, R. and Reynolds, P. (2009). U.S. Panel Study of Entrepreneurial Dynamics: PSED Overview. August 2009. Available at http://www.kauffman.org/Blogs/DataMaven/August-2009/Slides-and-Materials-from-the-Academy-of-Managemen.aspx

Reedy, E. and Frazelle, S. (2009). Entrepreneurship, Innovation, and Economic Growth: A Review of Recent Studies, unpublished report, Kansas City, Missouri: Kauffman Foundation. July 2009.

Reynolds, P. (208). Current Population Survey: Expanding Understanding of U.S. Business Creation. *Proceedings of the 2008 Kauffman Symposium on Entrepreneurship and Innovation Data*. November 2008. Available at http://www.kauffman.org/uploadedFiles/ResearchAndPolicy/EntrepreneurshipData/2008data/current-population-survey.pdf

Reynolds, P. and Curtin, R. (2008). Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment. *Foundations and Trends in Entrepreneurship*. Vol. 4, No. 3 (2008) 155–307.

Robb, A., Ballou, J., DesRoches, D., Potter, F., Zhao, Z. and Reedy, E. (2009) An Overview of the Kauffman Firm Survey: Results from the 2004–2007 Data. April 2009. Available at http://ssrn.com/abstract=1456380

Santos, B. and DesRoches, D. (2009). Exploring Mode Effects in a Panel Survey of New Businesses. May 2009. Available at http://ssrn.com/abstract=1409361

SYNTHETIC DATA APPROACHES TO CONFIDENTIALITY

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Analytical Validity and Confidentiality Protection in Longitudinally Integrated Statistical Data Systems

John M. Abowd [1]

Abstract

This paper summarizes the results of six different synthetic data projects conducted with the support of the National Science Foundation and using longitudinally integrated statistical data from censuses, surveys, and administrative record systems. The systems were all designed to produce statistically valid releasable micro-data protected by synthetic data techniques. The systems that were studied included longitudinal establishment data, longitudinally integrated administrative employer-employee data, geo-spatially integrated residence/workplace data, and household surveys integrated with longitudinal administrative data. Analytical validity and confidentiality protection results from these projects will be summarized.

[1] John M. Abowd, Cornell University, U.S.A. (john.abowd@cornell.edu)

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Summary of Methods and Preliminary Assessment of the SIPP Synthetic Beta, version 5.0

Gary Benedetto, Martha Stinson and Melissa Bjelland[1]

Abstract

This paper summarizes the methodology and quality assessment of the most recent version of the SIPP Synthetic Beta (SSB v5.0), a public use dataset that combines variables from the Census Bureau's Survey of Income and Program Participation (SIPP), the historical earnings data from Internal Revenue Service (IRS) tax forms, and the Social Security Administration's (SSA) individual retirement and disability benefit data. Multiple imputation and partial data synthesis were used to complete and perturb the data so that the final data product (multiple data sets, called implicates, which have the same structure as the underlying confidential data) would not compromise data confidentiality. The benefits of the methods used in this project are that the data users can run their analyses on each synthetic implicate exactly as they would have if they had access to the single, confidential data set. After getting results for each synthetic implicate, relatively simple formulae exist to combine these results to get proper point estimates and measures of variance that take into account the uncertainty introduced by the modeling. Moreover, since every value of the vast majority of variables on the file has been replaced by random draws from a probability distribution, the partially synthetic data offer a very high level of confidentiality protection. We also attempt to assess the analytic validity of the partially synthetic data and quantify the disclosure risk of making such data available to the public.

[1] Gary Benedetto (gary.linus.benedetto@census.gov) and Martha Stinson (martha.stinson@census.gov), U.S. Census Bureau, U.S.A.; Melissa Bjelland, Cornell University, U.S.A.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Synthetic Data Creation for the Cross National Equivalent File

Cynthia Bocci and Jean-François Beaumont[1]

Abstract

The creation of synthetic data as a method of disclosure avoidance has gained popularity over the last 15 years. Statistics Canada has recently started investigating techniques to create synthetic data for the Canadian portion of the Cross National Equivalent File (CNEF). The Canadian portion of the CNEF is derived from a subset of variables from Statistics Canada's Survey of Labour and Income Dynamics, a longitudinal survey. Due to confidentiality constraints, the Canadian data can only be accessed through special arrangements. The creation of synthetic data would allow easier access and hopefully increase its use. In this article, we describe the methodology used to create longitudinal synthetic data for the Canadian portion of CNEF and discuss the challenges of creating consistent households that preserve as much as possible the relationships in the original data while keeping the risk of divulging confidential information at a low level. We present selected results.

Key Words: Confidentiality, Household Surveys, Linear Regression, Logistic Regression, Parametric Bootstrap.

1. Introduction

The creation of synthetic data as a method of disclosure avoidance has gained popularity over the last 15 years. Statistics Canada has recently started investigating techniques to create synthetic data for the Canadian portion of the Cross National Equivalent File (CNEF). The CNEF involves panel surveys from six countries with comparably defined variables. The Canadian portion of CNEF is derived from a subset of demographic, employment and income variables from Statistics Canada's Survey of Labour and Income Dynamics (SLID), a six-year longitudinal survey with unique person and yearly household level identifiers. Due to confidentiality constraints, the Canadian data can only be accessed through special arrangements unlike data from other countries that is collected by universities or private institutes. As a result, the Canadian data is sometimes omitted from analyses. The creation of synthetic data would allow easier access for a larger number of researchers and hopefully increase its use.

In this article, we describe the methodology used to create longitudinal synthetic data specifically for the Canadian portion of CNEF and discuss the challenges of creating consistent households that preserve as much as possible the relationships in the original data while keeping the risk of divulging confidential information at a low level. The discussion is thus limited to this particular experience. A more ample development of the theory behind our approach as well as a more thorough discussion of the literature review in this area are addressed in an unpublished document (Beaumont & Bocci, 2010).

The layout of this article is as follows. Section 2 briefly introduces the underlying concepts and theory. The generation of synthetic data is described in Section 3. Validation techniques and selected results for the first year of data are presented in Section 4. Section 5 discusses longitudinal issues and section 6 concludes with a summary and possible future work.

2. Preliminaries

2.1 Main assumption

Suppose data from a finite population is generated according to some model. Let $\boldsymbol{\theta}$ be the vector of model parameters of interest and $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ be the estimator of $\boldsymbol{\theta}$ using sample data $\mathbf{Y}$. Due to confidentiality concerns, neither $\mathbf{Y}$ nor $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ can be released.

[1]Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, 150 Tunney's Pasture Driveway, R.H. Coats Building, 16th Floor, Ottawa, Ontario, Canada, K1A 0T6 (Jean-Francois.Beaumont@statcan.gc.ca); Cynthia Bocci, Statistics Canada, Business Survey Methods Division, 150 Tunney's Pasture Driveway, R.H. Coats Building, 11th Floor, Ottawa, Ontario, Canada, K1A 0T6.

The goal of synthetic data methods is to generate a sample matrix $\mathbf{Y}^*$ so as to obtain the synthetic estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)$ of $\boldsymbol{\theta}$. The matrix $\mathbf{Y}^*$ should be generated so as to preserve, as much as possible, the relationships between the original variables. We generate $\mathbf{Y}^*$ using a parametric multivariate model; this is called the parametric bootstrap. The general strategy for the generation of synthetic data is explained in section 3.1. We make the following main assumption on the synthetic data.

Main Assumption: The first two bootstrap moments of $\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})$ are asymptotically unbiased and consistent for the first two moments of $\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}$ under the model *m* (conditional on the sample *s*). More explicitly, we assume that

i) $E_*\left(\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})\middle| s,\mathbf{Y}\right)\approx E_m\left(\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}\middle| s\right)\approx \mathbf{0}$; and

ii) $E_m E_*\left\{\left(\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})\right)^2 \mid s,\mathbf{Y}\right\}\approx E_m\left\{\left(\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}\right)^2 \mid s\right\}$.

The subscript * refers to the bootstrap distribution. This assumption should hold reasonably well in practice if the model selected for the generation of $\mathbf{Y}^*$ is satisfactory. This is a standard assumption in (parametric) bootstrap theory. Part (i) of the assumption means that the generation of synthetic data brings no significant bias in the estimators. Protection of confidentiality is thus achieved through an increase in variance. Part (ii) is only required for the validity of our proposed variance estimators in (2).

2.2 Variance estimation

In the literature on synthetic data, multiple synthetic data files are usually created. Assume that *M* synthetic data matrices, $\mathbf{Y}^{*(1)},...,\mathbf{Y}^{*(M)}$, are generated independently and lead to *M* synthetic estimators $\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(1)}),...,\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(M)})$. The final synthetic estimator is the average over all *M* datasets

$$\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)=\frac{\sum_{i=1}^{M}\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(i)})}{M}.$$

Then an estimator of $\mathrm{var}_{mp*}\{\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)\}$, where the subscript *mp* refers to the joint distribution induced by the model *m* and the sampling design *p*, is given by

$$V_M=\tfrac{1}{M}\sum_{i=1}^{M}v(\mathbf{Y}^{*(i)})+\tfrac{1}{M(M-1)}\sum_{i=1}^{M}\left\{\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(i)})-\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)\right\}^2 \qquad (1)$$

where $v(\mathbf{Y}^{*(i)})$ is a standard variance estimator such as one obtained by linearization, bootstrap or any other replication method. Variance estimator (1) is exactly the same as the one proposed by Reiter (2003), with the difference that he obtains $\mathbf{Y}^{*(i)}$ using multiple imputation and not bootstrap. The advantage of using a bootstrap procedure is that it does not require reflecting the uncertainty in the estimators of model parameters, thus simplifying the synthetic data generation.

Alternatives to the variance estimator (1) are obtained using the main assumption and are given by

$$\tilde{V}_{1M}=\left(1+\tfrac{1}{M}\right)\frac{\sum_{i=1}^{M}v(\mathbf{Y}^{*(i)})}{M} \quad \text{and} \quad \tilde{V}_{2M}=\left(1+\tfrac{1}{M}\right)\frac{\sum_{i=1}^{M}\left\{\hat{\theta}(\mathbf{Y}^{*(i)})-\hat{\theta}_M(\mathbf{Y}^*)\right\}^2}{M-1}. \qquad (2)$$

A practical advantage of $\tilde{V}_{1M}$ over $\tilde{V}_{2M}$ and V_M is that it can be applied even when $M=1$. Typically, $\tilde{V}_{2M}$ and V_M will require a large number *M* of synthetic datasets to be reliable. Since it is often more convenient to produce and use a single synthetic data file, the variance estimator $\tilde{V}_{1M}$ becomes attractive. However, a disadvantage of using a small value of *M* is that it leads to a less efficient synthetic estimator $\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)$. This might be the necessary price to pay for protecting against disclosure of confidential data.

3. Synthetic data creation

In this section, we describe the methodology used to create the CNEF synthetic data. The methodology includes aspects of modelling, elements of randomness and a strategy to create synthetic values consistent within a household.

The first year, 1999, of CNEF real data has approximately 38,000 individuals comprising 15,000 households and contains demographic, employment and income variables. The initial task is to create a synthetic version of this data set with the same number of individuals and households. Of the nearly 40 variables on the file, only the marital status, the relationship to head indicator and a weight variable are not synthesized. That is, the real values for these three variables are retained for all individuals to form the basis of the synthetic data set. All the other variables must be synthesized for all the individuals on the synthetic data set. Once the first year of data is complete, the task is to produce longitudinal data sets for the 5 remaining years.

3.1 Methodology

Assuming that the observations are independent in the sample, the generation of synthetic data first requires choosing a parametric multivariate distribution of $\mathbf{y}_k' = (y_{1k},...,y_{qk})$ conditional on unit k belonging to the sample s using the sample data $\mathbf{Y}$, with q being number of variables of interest to be synthesized. Then, this multivariate distribution is estimated by estimating its unknown parameters and, finally, the estimated distribution is used to generate synthetic data. The multivariate modelling task is decomposed into q conditional models. First, y_{1k} given $k \in s$ is modelled and synthesized. Then, y_{2k} given y_{1k} and $k \in s$ is modelled and synthesized and so on until y_{qk} given $y_{1k},...,y_{q-1,k}$ and $k \in s$ is modelled and synthesized. Since each of these q models is conditional on the sample, it is not necessary to use survey weights when selecting models. However, it might be a good strategy to condition on the design information, or at least on the survey weights, when models are selected. This ensures that sampling is not informative; i.e., conditioning on the sample does not modify the unconditional multivariate model that holds in the population. To this end, the weight variable is categorized and this categorical weight is used in the modelling process.

The conditional multivariate task is repeated until all variables have been synthesized. The order in which the variables are synthesized is irrelevant if the modelling is adequate. Nevertheless, an ordering is required and may be determined by considering practical issues and anticipating possible analyses performed by data users. In the case of the Canadian portion of CNEF, the established ordering of variables to be synthesized is demographic, employment and then income. The actual multivariate models used in our application generally do not involve high order interaction terms. While the selection of the model is itself an unweighted process using real data, the subsequent use of these models to generate synthetic data is a weighted process.

3.2 Preserving household structure

The CNEF data contains both personal and household identifiers thus allowing for both person-level and household-level estimates. Therefore, it is necessary to create individual synthetic values which are consistent with the values of other members of the household. To preserve relationships among household members we typically synthesize values for a variable in question as follows. First, we synthesize a value for the head of the household. Next, we synthesize a value for the partner of the household given the value of the head of that household. We then synthesize the values for the children given the values of the head and the partner in the household. Finally, we synthesize a value for the remaining members of the household given the value of the household head. This structural household approach is similar to the one taken by Reiter (2005).

3.3 Categorical variables

Table 3.3-1 lists the categorical variables to be synthesized. In general, for a categorical variable, we use a logistic regression model to yield a probability of belonging to each category. For a given synthetic record, assignment to categories is done at random using these probabilities.

Table 3.3-1
Categorical variables to be synthesized

Variable type	Variable
Demographic	sex, race, education level, disability status, health status, province
Employment	annual work hours, employment status, employment level, primary activity, occupation, 1-digit industry code, 2-digit industry code

3.4 Quantitative variables

The quantitative variables consist of two demographic and all income variables. Table 3.4-1 lists the quantitative variables to be synthesized.

Table 3.4-1
Quantitative variables to be synthesized

Variable type	Variable
Demographic	age, number of years of education
Income	Incomes: pre-government, post-government, labour, asset, private retirement, lump sum, pension Transfers: private, public, government, Taxes: federal, provincial, other, total

For the demographic quantitative variables, in most instances we use a general linear model to yield predicted values for both real and synthetic units as well as residuals for the real units. The predicted values are grouped together to form classes. An element of randomness is then introduced and the value is further modified to yield a final synthetic value as follows. For each synthetic unit, we select a residual at random belonging to a real unit from within the given class. This residual is then added to the predicted synthetic value and the sum is rounded to form a final synthetic value. Edit rules are verified and if they fail for a particular synthetic unit, the random process is repeated. Certain modifications to the randomization procedure for the age of partners and children are necessary in order to produce synthetic data values with distributions that closely resemble those of the real data values.

The income variables present certain challenges. These variables consist of revenues, taxes and transfers whose correlations have to be maintained in order for the synthetic data to be useful. These continuous, generally skewed variables are also subject to outliers which make modelling more complicated. Furthermore, the variance of the data values is difficult to capture using the general method for continuous variables described above. For these reasons, income variables are created differently. Individual labour earnings are synthesized first and then all the other income variables are created simultaneously. Note that income variables apply only to individuals greater than 16 years of age.

The variable individual labour earnings requires a separate treatment for employed and unemployed individuals. For all employed individuals, a general linear model is used to obtain predicted values for both synthetic and real units. Predicted earnings for both real and synthetic units are grouped together to form classes. For each synthetic unit, a value of labour earnings from a real unit is selected at random within the class and is assigned to the synthetic unit. For the unemployed, we first model the probability of having zero or non-zero individual earnings. For those unemployed synthetic units who are randomly determined to have non-zero earnings based on these probabilities, a general linear model for earnings is then applied and a synthetic value is created similarly as in the employed case. Finally, the non-zero earnings of all synthetic units are modified by an adjustment described in section 3.4.1 which serves to not only help reduce the risk of disclosure but also force the first and second moments of the synthetic data values to more closely resemble those of the real data values.

Next, all the remaining income variables are synthesized simultaneously as follows. Within the same classes used to create non-zero labour earnings, a real nearest neighbour with respect to labour earnings, referred to as a donor, is identified for each synthetic unit. This donor's ratios of the income variables of interest to labour earnings are applied to the final synthetic value of labour earnings of the synthetic unit. For those unemployed with zero earnings, the classes are defined by grouping the probabilities of having zero labour earnings for both the real and synthetic units. A real unit within the class is then chosen at random and the real values of the remaining income variables are assigned to the synthetic unit.

This procedure applies to individual level income variables. With the exception of the individual labour earnings, all other income variables on the CNEF file are at the household level. Therefore, the household level value is created simply by summing of values of everyone in the household for a particular income variable.

3.4.1 Adjustment to individual labour earnings

The weighted univariate analysis of the synthetic values of individual labour earnings before an adjustment suggests that the second moment in particular is not adequately captured. To improve this and to further disguise the donor values, we modify the synthetic labour earnings by linearly transforming them subject to two constraints. The adjustments are made within a group g. Let y denote a real value and y^* denote a synthetic value before the adjustment. Define the final adjusted synthetic value, y^{**}, for unit $k \in g$ as

$$y_k^{**} = \begin{cases} b_{1g} y_k^* & \text{if } y_k^* \geq 0 \\ b_{2g} y_k^* & \text{if } y_k^* < 0 \end{cases} \qquad (3)$$

for real constants b_{1g} and b_{2g} such that the values y^{**} satisfy the constraints

$$\frac{\sum_{k \in s_{g,syn}} w_k y_k^{**}}{\sum_{k \in s_{g,syn}} w_k} = \frac{\sum_{k \in s_{g,real}} w_k y_k}{\sum_{k \in s_{g,real}} w_k} \quad \text{and} \quad \frac{\sum_{k \in s_{g,syn}} w_k y_k^{**2}}{\sum_{k \in s_{g,syn}} w_k} = \frac{\sum_{k \in s_{g,real}} w_k y_k^2}{\sum_{k \in s_{g,real}} w_k},$$

where $s_{g,syn}$ and $s_{g,real}$ denote the synthetic sample pertaining to group *g* and the real sample pertaining to group *g* respectively. The transformation (3) is such that the adjusted synthetic value keeps the same sign as the pre-adjusted synthetic value. A closed-form solution for b_{1g} and b_{2g} is easily derived. Calibration groups may need to be manipulated to ensure that the solutions of b_{1g} and b_{2g} have real values. This type of adjustment is similar to a calibration-type of adjustment where the weights are normally modified rather than the *y*-values.

3.5 Limitations and practical issues

There are a number of limitations and practical issues that arise from implementing the approach described above. The quality of the synthetic data depends on the validity of the model used. As the number of variables to consider in a model grows, it becomes more difficult and impractical to find all significant interaction terms. Also, rare events may prove difficult to model.

Another difficulty is encountered with extreme values in the real data set. While it is possible to generate extreme values for quantitative variables on the synthetic data set, they are not very likely to occur. Extreme values can result in differences between real and synthetic univariate distributions.

4. Empirical validation

4.1 Validation

Throughout the synthetic creation process, a certain amount of validation occurred after the creation of each synthetic variable. Various frequencies, cross-tabulations and univariate distributions were inspected visually for both the synthetic and real data. If synthetic results were unsatisfactory, the models were refined or the random mechanism was modified in an attempt to fix as many inconsistencies as possible.

4.2 Selected results

Comparisons between one year of synthetic data and the corresponding year of real data are presented in this section. The results shown are for illustration purposes rather than for commenting on the overall quality of the synthetic data.

Figure 4.2-1
Estimates of Average Labour Earnings by Age/Sex/Occupation Groups

It appears that many of the correlations between the household level income variables are well preserved on the synthetic data set. For example, the correlations between total household earnings and total household taxes for the synthetic and real data, with one real unit outlier removed, are 0.825 and 0.829 respectively. Similarly, the correlations between household private retirement income and household social security pensions for the synthetic and real data are 0.489 and 0.448 respectively. Next, adults aged 16 to 69 with positive individual labour earnings are assigned to groups created by crossing 6 age categories with sex and with 25 occupations. Figure 4.2-1 graphs synthetic estimates against real estimates of the average earnings by group for group sizes greater than 30. The size of the circle indicates the synthetic sample size of the group while the straight line represents equality between the real and synthetic estimates. While not all points lie close to the line, the median of the relative difference is 7%. The coefficient of correlation between these real and synthetic estimates is 0.86.

For these same data points, we calculate the variance using $\tilde{V}_{1M}$ in equation (2) and bootstrap weights. We then calculate the percentage of the confidence interval overlap; that is, the overlap between the synthetic confidence interval and the real confidence interval relative to the length of the real confidence interval. Theoretically, the probability of complete overlap under our main assumption, normality assumptions and using only one synthetic data set is calculated to be 58% (Beaumont and Bocci, 2010). In this example, the proportion of complete overlap is 41% which is not too far from the theoretical value.

5. Longitudinal considerations

The ultimate goal is to create a longitudinal synthetic file for the Canadian portion of the CNEF. The process of creating a synthetic file for any of the years t after the first wave is threefold. First, there are some peculiarities with the source of the Canadian portion of CNEF which complicate identifying exactly which real individuals need to be on the real file at time $t+1$ for the purposes of generating the synthetic data set for the year $t+1$. The second step involves applying changes in marital status and household structure from one year to the next, similar to those observed on the real data set, to the synthetic data set. Finally, once the synthetic individuals and households have been determined, the lengthy process of synthesizing all the variables can begin. Note that there is no non-response on the synthetic data set since synthetic values are created for everyone on the file. A brief description of each of these three steps is given in the subsections 5.1 to 5.3.

5.1 Establishing a real data set in a longitudinal year

Recall that some variables on the CNEF data set are at the person level while others are at the household level. Our first step is to carefully create a person-level real working file with household identifiers to reflect i) deaths of longitudinal individuals ii) household births arising from longitudinal units who split from a household in the previous year iii) individual births comprising of new entrants in the household of the longitudinal unit and iv) returning cohabitants living in the household of a longitudinal unit. Although we follow and eventually retain only the longitudinal unit, we need to keep track of everyone in the household including those in groups iii) and iv) in order to properly calculate household-level variables in a given year. A file containing longitudinal units as well as cohabitants is referred to here as a working file.

5.2 Generating changes in households

The creation of a synthetic database for year $t+1$, containing the same number of individuals and households as in the real data set of the same year, begins by applying changes in status and structure to the synthetic data set of year t. Changes in status and structure are brought about in part by births, deaths, household splits or combinations thereof. The synthetic working file must experience the same type of changes in approximately the same frequencies as on the real working file. In order to achieve this, classes are carefully constructed by crossing certain household-level variables for households of a given size. Classes are assigned to both real and synthetic households with values of variables from year t and must be fine enough to generate appropriate changes. For each synthetic household in a given class in year t, a real household is selected at random within that class. Whatever structural changes, if any, experienced by the real donor household are forcibly applied to the synthetic household thus creating the correct number of synthetic individuals and households. In addition, the individual donor values of marital status and relationship to head indicator for the year $t+1$ are retained to form the basis of the synthetic working data file.

5.3 Creating longitudinal data values

The longitudinal data set contains the same variables for each of the six years of data and attempts to capture relationships over time. With the same variable ordering as in the previous year, each variable is synthesized in much the same manner as in the first year except that, for longitudinal units, variables from the previous year are now available for use as potential explanatory

variables in the modelling. Specifically, we chose a model where the j^{th} synthetic variable , $j = 1,...,q$, for the year $t+1$ is a function of the first j-1 synthetic variables on the data set of year $t+1$ and the last q-(j-1) synthetic variables from the data set of the previous year t. This approach means that we consider only two consecutive years in the creation of any data set other that the first one and we limit the potential explanatory variables to a manageable number.

For new cohabitants appearing on the synthetic working file, no previous year information is available and synthetic data values are created similarly to the way in which synthetic values were created in the first year.

6. Conclusions

To date, only one year of synthetic data has been created. Preliminary investigations suggest that the synthetic data created thus far has some usefulness and hence, a longitudinal strategy has been put into place with the aim of completing a six year longitudinal data set for the Canadian portion of the CNEF. Through sequential modelling and randomization, it is hoped that the individual synthetic values are sufficiently different from the real values so that the risk of disclosure is minimal while the multivariate distributions among variables are satisfactorily preserved.

Future work includes the completion of the longitudinal data set and a more thorough analysis of the synthetic data quality including recently developed global measures of data utility for synthetic data by Woo et al. (2009). A greater use of non-parametric methods to better capture the real data relationships could also be investigated, although these may lead to a higher risk of disclosure.

References

Beaumont, J.-F. and Bocci, C. (2010). Some Theory on Synthetic Data Generation and Its Application to the Canadian Portion of the Cross National Equivalent File, unpublished document, Ottawa, Canada: Statistics Canada.

Reiter, Jerome P. (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: an Illustration and Empirical Study, *Journal of the Royal Statistical Society A*, 168, Part 1, pp. 185-205.

Reiter, Jerome P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets, *Survey Methodology*, 29, No. 2, pp. 181-188.

Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A.F. (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation, *The Journal of Privacy and Confidentiality*, 1, Number 1, pp.111-124.

WAKSBERG AWARD WINNER ADDRESS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Methods for Oversampling Rare Subpopulations in Social Surveys

Graham Kalton[1]

Abstract

Increasingly social surveys are required to produce estimates for subpopulations, often rare subpopulations. Sometimes a survey focuses on a single subpopulation but often the survey is required to produce estimates for several subpopulations and also for the total population. When membership of a rare subpopulation can be determined from the sampling frame, selecting a sample of the required size for the subpopulation is relatively straightforward. In this case the main issue is the extent of oversampling to employ when the survey aims to produce estimates for several subpopulations and the total population. Oversampling a rare subpopulation that cannot be identified from the sampling frame presents a major challenge. Methods to perform this oversampling include disproportionate stratified sampling, two-phase sampling, the use of multiple frames, multiplicity sampling, panel surveys, and the use of multi-purpose surveys. This paper will describe these methods and illustrate their application in a range of surveys.

Professor Kalton's complete paper can be found in Survey Methodology December 2009.
http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11036-eng.pdf

[1] Graham kalton, Westat, U.S.A. (grahamkalton@westat.com)

LONGITUDINAL HEALTH DATA: ISSUES AND CHALLENGES

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Establishing a Longitudinal Community Health Research Methodology: Issues and Challenges

David Marshall PhD[1]

Abstract

Longitudinal methods have enormous potential for better understanding of community health and wellbeing. In particular, awareness of local social, economic and environmental conditions and their influence on the burden of chronic disease can be enhanced by place-based studies. However, to date they have been relatively under-utilised in this field. This paper outlines some methodological issues being considered in the establishment of 'The Ipswich Study', a proposed place-based longitudinal study in in Queensland, Australia. Given the rapid changes and forecast growth anticipated for the region, a unique opportunity to analyse impacts of change on community health has emerged. In particular, the paper focuses on potential problems in studying a region which is undergoing enormous social, economic and environmental change.

Key Words: Community Health, Place-Based, Chronic Disease, Socio-Economic Change.

1. Introduction

This paper discusses issues around a proposed longitudinal study entitled 'The Ipswich Study'. In general terms 'The Ipswich Study' will be a place-based longitudinal study of health and wellbeing in the community of Ipswich, an outer suburban region in Queensland, Australia. Ipswich is about 40km south west of Brisbane, the capital of Queensland. Both Brisbane and Ipswich are located in South East Queensland, a region which has experienced extremely high rates of population growth over recent decades, particularly in the Ipswich area. The purpose of this paper is to identify some of the logistical and methodological challenges in establishing the study.

2. Background contextual issues

The contextual circumstances are critical in determining the approach taken in establishing The Ipswich Study. Not only do they provide the wider background in which the study is occurring, but they also have a profound influence on the logistical and methodological development of the study. While some of these issues are quite specific to The Ipswich Study, they are likely to be similar to challenges encountered in other longitudinal methodologies.

2.1 Stakeholders and funding

The Ipswich Study emerged from the establishment of the Healthy Communities Research Centre (HCRC) at the University of Queensland's Ipswich campus. The HCRC was established as a joint initiative between the Ipswich Hospital Foundation (IHF) and the University. The Study and its objectives are thus heavily influenced by the interests of both organisations, which although similar, are not perfectly aligned. The IHF has a charter to promote healthy lifestyles, support health professionals and to build a healthier community in Ipswich. As such, research in and with the Ipswich community which produces results which have application at a practical, local level is of great importance. Contrasting this, the University demands internationally relevant research which will raise the profile of both the HCRC and the University overall. There is thus a need to design a study which addresses specific issues of local interest but which engages with and advances understanding of the field in an international forum.

In addition to these two stakeholders, there is also a need to consider the interests of potential stakeholders. In large part this is because The Ipswich Study is currently unfunded. Current stakeholder funding is sufficient to develop the study but will not

[1]David, Marshall, Healthy Communities Research Centre, University of Queensland, Salisbury Road, Ipswich Australia 4305.

extend to implementation. While there are numerous potential sources to fund the study (e.g. various levels of government), until funds have been committed, there is a need to develop the study with a wide degree of flexibility in terms of subject matter, timeframes and budgets, all of which will impact upon methodology.

2.2 Research agendas, hypotheses and questions

While issues surrounding stakeholders and funding present both methodological and logistical challenges for the project, they also have an influence on the development of hypotheses and research questions. Specifically, because of the need to keep the study flexible to enable adaptation, the development of The Ipswich Study has not been guided by specific research hypotheses or questions but is driven by broader agendas of the stakeholders and HCRC. While this has its positives, it also poses challenges in terms of the methodological development of the Study. The challenge is being able to situate and focus the study in a niche with clearly evident topical boundaries, but which simultaneously retains flexibility to allow the study to address specific research questions as they emerge or are requested by stakeholders. There is thus a balancing act between being broad and being specific.

The broad agenda for the HCRC is to further understanding of social and contextual determinants of health at both the individual and community level. Such determinants essentially encompass the non-medical and non-individual determinants of health, specifically around chronic preventable conditions. They include the built and social environment, policies and programs which influence the community, local economic resources and general environmental conditions, among others. From this set of objectives and the need to generate a program of research which addresses the locally focussed needs of the IHF as well as the University's international charter, a longitudinal study of health and wellbeing in Ipswich was proposed, with a vision that the study would:

- examine the interface between various determinants of health;
- integrate understanding of contextual and environmental influences on the health of communities;
- give insights into health differences across Ipswich and between communities reflecting different characteristics and contextual environments, over time; and
- assess how interventions and changes to the wider community manifest, and are experienced locally.

3. Major methodological issues

In addition to the background issues which have a major bearing on the development of The Ipswich Study, there are numerous challenges which have emerged around 'longitudinal' issues related to community health and wellbeing. Such challenges encompass subject matter as well as logistical and definitional problems which ultimately have a major bearing on the overall methodological design. These are outlined here.

3.1 Ipswich: A rapidly changing community

Given that The Ipswich Study may run for a long period of time, like all longitudinal studies a critical issue is ensuring that the study retains relevance over time. Whilst this will always be a challenge for a place-based longitudinal study, in Ipswich it may be more difficult due to rapid changes in the area. South East Queensland has been one of the fastest growing regions in the country for many decades and this is forecast to continue. At present the population of the Ipswich area is around 150,000 but is forecast to triple within 25 years (State of Queensland 2009). The Queensland Government has responded with a regional plan and placed Ipswich firmly in the centre of that plan. As such a number of major master-planned communities are either under construction or slated for development. The Ipswich Study must therefore be in a position to encompass this growth which through new population groups, changing demographics and new infrastructure will have a major impact on the region. Designing a study which can encompass new groups and changing contexts is thus critical. However, such changes also present an opportunity to understand their impact over time. To our knowledge no such study has occurred internationally which has been positioned to understand the impact of major environmental, social and economic change on the health and wellbeing of individuals and communities. This will be a key goal of The Ipswich Study, but ensuring a capacity to incorporate the changes occurring will present numerous methodological challenges.

3.2 Individual and ecological variables

A further challenge will be to ensure the study can encompass a wide range of issues which have been identified in the literature as relevant to community and individual health at a local level. In reviewing the social determinants of health

literature, two streams of contemporary research approaches were identified as central to our agendas – the life-course and ecological approaches. The life-course approach to health outcomes focuses on the pathways individuals take and their exposure to risk and protective factors over time. The ecological approach is based upon the well evidenced understanding that where one lives can make a difference to a person's health and wellbeing over and above their individual characteristics, experiences, intentions and behaviours. The Ipswich Study will seek to unify these two research themes into a single study.

Essentially these two themes emerged because there has been a distinct lack of research examining the extent to which the area of residence at different stages of the life-course may influence health (Leyland and Næss 2009). As Berkman (2009) observes, the fact that these two perspectives have operated separately from one another helps explain why studies often come to different conclusions about social and behavioural risks to health. To fully understand area effects on health, there is thus a need to implicate both the individual and ecological factors (Moon et al. 2005). There have thus been many authors calling for methodological approaches which combine both life-course and ecological approaches to the understanding of health and wellbeing. Such research will have a greater chance to further understanding of causality (Bauman 2005). The Ipswich Study will adopt such an approach and:

- will incorporate information on health characteristics of both individuals and communities;
- will collect evidence from the places, circumstances and environments in which people live their lives;
- understand adaptations and decisions people make under diverse and changing circumstances; and
- pay attention to transitions across the life-span and the changing nature and shape of communities.

As such, the research agenda will involve data collection which encompasses both individuals and their environments. This poses many challenges for the study. For example, studies looking to assess ecological models in a local area must be able to ensure areal variation in the independent variables of interest. Additionally, in a longitudinal study it will be necessary to encompass variables which also change over time – and for which change is actually detectable. This is however one of the major benefits of The Ipswich Study because it is known in advance that Ipswich will be changing dramatically and quickly. There is thus an ideal opportunity to establish a cohort study designed to understand the impact of contextual changes on health as they occur.

3.3 Compositional, contextual and collective factors

Furthermore, whilst the decision to encompass both life-course and ecological approaches will definitely place the study at the cutting edge of community health research, enormous logistical challenges are immediately raised, primarily being the vast numbers of life-course and ecological variables identified in the literature which are thought to have relationships with health and wellbeing outcomes. Although the study will be looking to encompass as many critical evidence-based variables as possible, it is also crucial to retain a manageable study which does not over-burden the participants through too many questionnaire items and complex methodological protocols.

There are thus a vast range of factors which encompass both physical and social attributes of neighbourhoods which could plausibly affect the health of individuals (Diez-Roux 2007). However, given the study's interest in place and changing places, variable selection can be narrowed to be drawn from three categorisations of environmental variables often utilised in the ecological literature. These are:

- Compositional factors – referring to the make-up of the place as determined by the characteristics of the people living in it (e.g. socio-economic and demographic characteristics);
- Contextual factors – related to the physical and non-physical variables of the environment (e.g. built, natural and social environments which includes buildings, policies and programs among other things); and
- Collective factors – which recognise characteristics which exist above and beyond the individuals, but which are not tangible contextual factors either (e.g. social norms, social capital and collective efficacy).

From the above three categories of variables, there will be a need to encompass indicators from each, as well as a need to recognise that not all variables within each category will be straightforward to collect. For example, while some variables can be measured in an objective sense through observation, interviews and existing data sources, many may be highly subjective and thus can differ from place to place, from person to person and over time. In such cases it is often the perception rather than the reality which is most important in influencing behaviour and thus health outcomes. In some cases and for some types of variables it may be prudent or even necessary to collect both objective and subjective data related to the same variable. For example, both actual crime rates and fear of crime as perceived by residents could each have independent effects on behaviour and thus health related outcomes. The Ipswich Study must identify the most important variables which can legitimately be measured and changes detected across space and over time.

3.4 Dynamic and reflexive relationships

The task of identifying a key set of variables and indicators is further complicated by increasing acceptance that reflexive relationships between people and their environments need to be scrutinised to understand how contexts affect health. This is because relationships between individuals and the circumstances in which their lives are lived are both two way and dynamic. That is, every action and experience by an individual is in some way influenced by the environment in which the action takes place, but the action also has an effect in the other direction. Therefore relationships between people and their contexts and environments should not be treated as passive or static if the fullest understanding is to be achieved. Unlike cross-sectional research which operates as a point-in-time snapshot, longitudinal research which is looking at relationships between individuals, their behaviours and their environments over time must accommodate understanding of the reflexive and dynamic nature of these relationships.

The challenge is to design a study which can collect and analyse data which reflects the dynamic relationships described above. To better understand such relationships, it is likely that both qualitative and quantitative data will be needed to understand decision-making processes which influence these relationships. This may help provide insights to the mechanisms which influence the relationships between places and the health of people living in them.

3.5 Defining communities and neighbourhoods

The last major issue being raised in this paper is the ever-present problem of defining what is meant by a community or neighbourhood. This is because what constitutes a neighbourhood or a community can vary depending on a variety of factors including the nature of the place, the person/group in question or the issue of interest. For example, younger people may have very different understandings of local community than do older people. Similarly, what constitutes a community might vary across neighbourhoods with some areas having tighter and more rigid understandings of community than other areas. When considered from the perspective of a service or facility, definitions will be influenced by the client base, the charter of the organisation or what else exists in the area.

Also critical is the recognition that not only might these issues vary from place to place and variable to variable, but they can also change over time. This is critical because different variables may have an effect on health at different scales, but may also vary for individuals and communities over time. For example, if participants are asked about issues in the 'local area' we need to be mindful that what constitutes 'local area' can change during the study.

The straightforward approach in cross-sectional ecological research is to utilise pre-existing definitions and associated data which have been compiled for administrative or other reasons. However, such data whilst convenient to access and use, does not fully recognise the diversity of communities and neighbourhoods as experienced by those who live in them. Furthermore, such data may not be fine-scaled enough to detect variations if they differ across small areas, and certainly may not detect changes over time at a small area level. Our challenge is to accommodate definitions of community and neighbourhood which reflect this diversity.

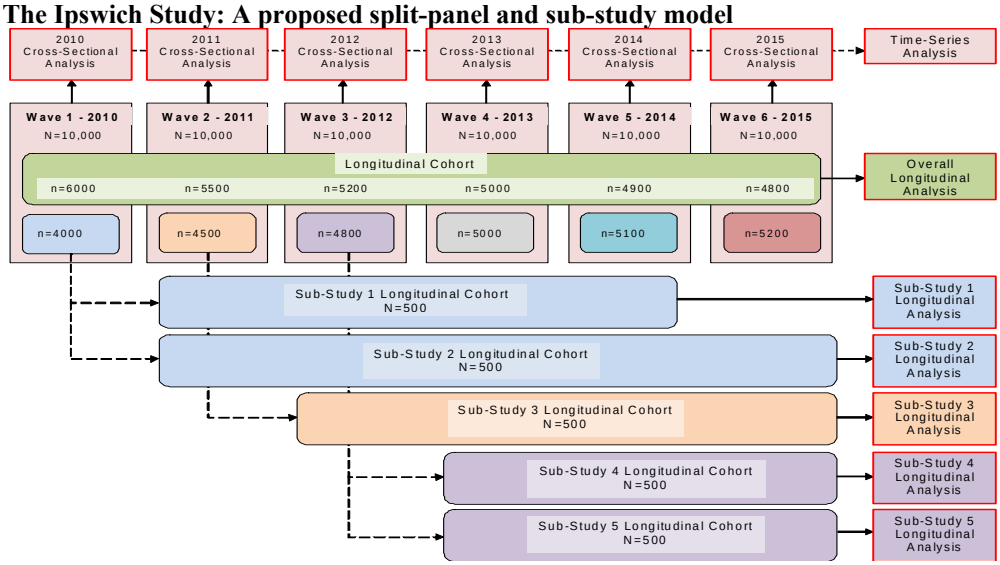
4. Discussion and conclusion

Despite all the uncertainties and unknown parameters within which The Ipswich Study is being developed, there remains a need to develop the project to an advanced level in readiness for when funding becomes available. Therefore, a broad methodological structure has been developed which is proposed to assist the study meet its many objectives, satisfy the different stakeholders and accommodate a wide range of subject matters related to health and the influence of environmental contexts upon health-related behaviour over time. Such a model needs to be scalable as well as retain the capacity for relevance both locally and internationally over time.

4.1 A Proposed methodological model

Figure 4.1-1 depicts a model being considered to carry The Ipswich Study agenda. The model is based on a split panel design and for the purposes of illustration has been designed with six annual waves of data collection and a 10,000 person annual cohort. An important component of the model are the sub-studies which rather than being additional after-thoughts, are built into the framework as critical components designed to help address many of the challenges raised in this paper.

Figure 4.1-1
The Ipswich Study: A proposed split-panel and sub-study model



At the heart of the split panel is a longitudinal cohort which will be recruited and retained over a long period of time. Initially, the data collected from this group will be wide-ranging but less detailed - around a broad range of health-related issues and agendas but which does not delve into great depth on many of them. This cohort of participants would not be topped-up over time and nor would they be asked to participate in any sub-studies. In time however it is acknowledged that it may be prudent to recruit a new longitudinal cohort to run in parallel.

The other component of the split panel is an annual recruitment of new participants. This group would supplement the core cohort in a number of ways and ultimately help the study to meet its multiple objectives and stakeholder needs. In particular, because the region is changing rapidly, the annual cross-sectional recruitment will provide a chance to recruit participants who for a variety of reasons are under-represented in the main cohort. This would then enable more accurate cross-sectional estimates for the purposes of the IHF's policy and program needs at each wave. Additionally, the persons recruited to the waves for the cross-sectional sample each year would be utilised to recruit participants for numerous sub-studies. Sub-studies could be either cross-sectional or longitudinal in nature, and may be qualitative or quantitative. They would all however be used to look at specific issues in far greater detail than can be done in the broader focussed longitudinal cohort. For example, a sub-study may focus on a specific demographic group (say young people) whereas another sub-study might look at a specific health issue. A major benefit of this approach is that the number of sub-studies will not be limited by a need to minimise burden on the longitudinal cohort. At every wave a new sample will be available from which to recruit sub-study participants.

Ultimately the different aspects of the study – namely the longitudinal, cross-sectional and sub-studies – would work together by feeding information to and from each other. Specifically it is envisaged that the sub-studies will help to inform the main cohort by helping refine the indicators used, whilst the main cohort will identify specific sub-studies which need conducting.

4.2 Summary and conclusion

What the Ipswich Study is seeking to achieve is an understanding of the how the health of individuals and communities is affected over the life-course and by the places in which people live as they change over time. Based on an extensive literature review, the social ecological and the life-course models, which are usually utilised independently of each other, should be incorporated for a better methodological approach. Such approaches can further understanding of the effects of neighbourhood environments on health over the life-course; the impact of moving from one neighbourhood to another; and changes over time to the neighbourhoods themselves. Such studies will require designs that follow both neighbourhoods and individuals over time (Diez-Roux 2003,). In a region such as Ipswich an excellent opportunity exists to establish such a study due to the rapid changes forecast.

However, what this paper is also seeking to highlight is that while the potential for understanding social determinants of health through longitudinal place-based cohort studies is great, there are major challenges to overcome in designing a study which can encompass the wide range of nuances and issues warranting inclusion. Additionally, in the case of The Ipswich Study, there are extra challenges resulting from the need to meet diverse stakeholder interests and to retain short-term flexibility. With the multiple challenges presented in establishing such a study, coupled with the specific parameters within which The Ipswich

Study is being established, there is a need to follow the advice of (Byles, et al, 2007) who note that, "like the making of fine wine, longitudinal studies need careful planning and management in the groundwork stages and will not produce their best results for many years". With careful planning coupled with innovative and flexible designs such as that proposed here, the issues and challenges raised can be overcome and ultimately utilised to the advantage of The Ipswich Study to produce locally useful but internationally significant results for many years to come.

Acknowledgements

This paper is based upon material prepared by Marshall and Bush (2009a; 2009b). Dr David Marshall's attendance at the XXVth International Symposium on Methodological Issues was partly funded by the Ian Potter Foundation.

References

Bauman, A. (2005). The physical environment and physical activity: moving from ecological associations to intervention evidence, *Journal of Epidemiology & Community Health,* 59, pp. 535-536.

Berkman, L. F. (2009). Social epidemiology: Social determinants of health in the United States: Are we losing ground?, *Annual Review of Public Health,* 30(1), pp. 27-41.

Byles, J., Dobson, A., Bryson, L. and Brown, W. (2007). Getting started: 'Preparing the ground' and 'planting the vines' for longitudinal research, *International Journal of Multiple Research Approaches,* 7(2), pp. 80-91.

Diez-Roux, A. V. (2003). The examination of neighbourhood effects on health: Conceptual and methodological issues related to the presence of multiple levels of organization, in I. Kawachi and L. Berkman (eds.) *Neighborhoods and Health*, New York: Oxford University Press, pp. 45-64.

Diez-Roux, A. V. (2007). Neighborhoods and health: where are we and where do we go from here?, *Rev Epidemiol Sante Publique,* 55(1), pp. 13-21.

Leyland, A. H. and Næss, Ø. (2009). The effect of area of residence over the life-course on subsequent mortality, *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172(3), pp. 555-578.

Marshall, D. and Bush, R. (2009a). The Ipswich Study: A Review of Longitudinal Methodology, unpublished monograph, Healthy Communities Research Centre. http://www.uq.edu.au/health/healthycommunities

Marshall, D. and Bush, R. (2009b). The Ipswich Study: Guiding Approaches and New Agendas, unpublished monograph, Healthy Communities Research Centre. http://www.uq.edu.au/health/healthycommunities

Moon, G., Subramanian, S., Jones, K., Duncan, C. and Twigg, L. (2005). Area-based studies and the evaluation of multilevel influences on health outcomes, in A. Bowling and S. Ebrahim (eds.) *Handbook of Health Research Methods: Investigation, Measurement and Analysis*, Maidenhead: Open University Press, pp. 266-292.

State of Queensland (2009). *South East Queensland Regional Plan 2009-2031*, Brisbane: Department of Infrastructure and Planning.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Analysis of the Longitudinal Health Approach Implemented in Belgium

Ann Ingenbleek, Yves Coppieters, Lies Lammens, Patrick Deboosere,
Florence Cols, William D'hoore and Alain Levêque[1]

Abstract

The implementation of "electronic government" in Belgium is an opportunity to modernize the health and social security sectors. The national health information system should be matched by a dynamic vision. Some recent Belgian initiatives are conducive to a longitudinal approach. These include computerized medical records, clinical-administrative registers, and the Permanent Sample, which is based on health insurance data. Legislative amendments, use of the personal ID number and the availability of substantial means are factors that favour a forward-looking vision of health. The related challenges entail social choices, the propagation of a comprehensive vision of public health and the assent of health professionals to the changes introduced.

Key words: Health information system, Belgium, longitudinal, indicators, computerization of data.

1. Introduction: research framework

The current Belgian health information system (HIS) is capable of giving a fairly complete picture of the consumption of health care, but it has not yet been adapted to describe the social and health dynamics—individual and environmental—that sooner or later lead people toward good or poor health statuses. Accordingly, the Service Publique Fédéral Sécurité Sociale has sought to expand its knowledge of the health of the population and has assigned the BeLHIS[2] project the two-fold task of taking stock of the issue and proposing strategies suited to the Belgian context, strategies that would provide for interaction and coordination of the different initiatives for future follow-up of health. Establishing a longitudinal vision of health implies that the information generated should shed light on health statuses and their determinants and on the use of the health system. In the end, this expanded capacity to observe should bring about a shift in health care management policies, toward approaches focusing on health promotion and disease prevention. The BeLHIS project offers the institutions and actors concerned a frame of reference, a place and an opportunity for debate for organizing a national longitudinal health information system. In the short term, the objective is also to show that the longitudinal approach to health is achievable when based on data linkage.

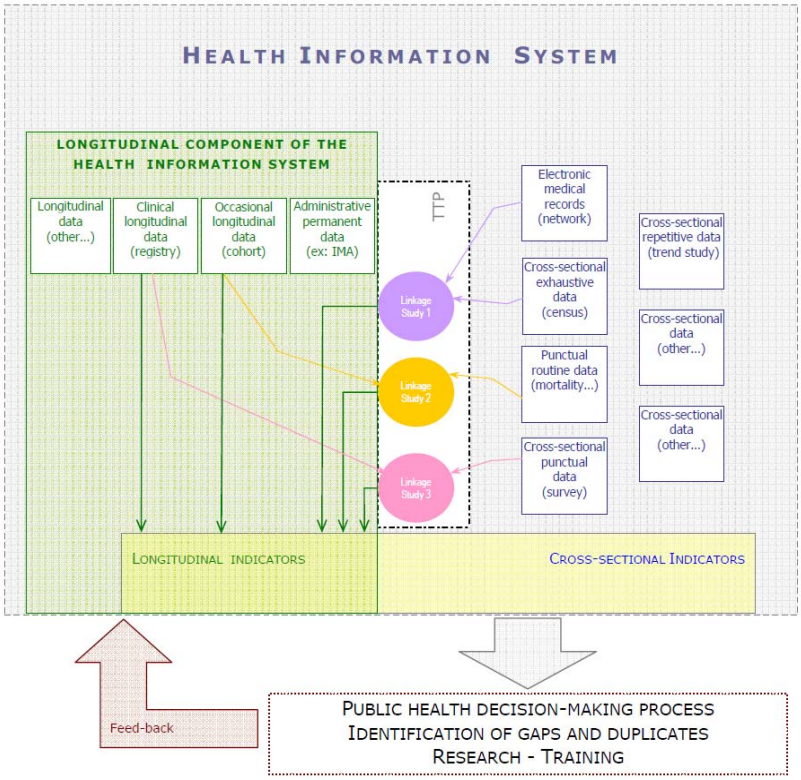
Belgium has few cohorts and few means of creating them. The officials concerned are therefore seeking to supplement the usual methods of follow-up of individuals by re-using data, and potentially by one-time linkages of data, for purposes of secondary analyses.

2. Proposed model for the Belgian longitudinal health information system

In the proposed conceptual model, the vast HIS is endowed with a longitudinal component (Coppieters, Bazelmans and, Levêque). One-time studies or collections of cross-sectional data (such as death records) provide the usual health indicators. For their part, longitudinal follow-ups summarize phenomena that take place over time, showing an evolution. They result in indicators that are presented in **Figure 2-1** under the term "longitudinal." Examples of such indicators are the rate of resumption of smoking or the survival rate five years after treatment of a cancer. Together, cross-sectional and "longitudinal" indicators make it possible to evaluate the health system, orient its objectives and identify priorities and appropriate approaches.

[1] Ingenbleek Ann, Ecole de Santé Publique, Université Libre de Bruxelles, 808 Route de Lennik, B-1070 Bruxelles, Belgique (aingenbl@ulb.ac.be) ; Coppieters Yves, Université Libre de Bruxelles ; Lammens Lies, Vrije Universiteit Brussel, Belgium ; Deboosere Patrick, Vrije Universiteit Brussel, Belgium ; Cols Florence, Université Catholique de Louvain, Belgique ; D'hoore William, Université Catholique de Louvain, Belgique ; Levêque Alain, Université Libre de Bruxelles.
[2] **Be**lgian **L**ongitudinal **H**ealth **I**nformation **S**ystem

Figure 2-1
Conceptual model of the health information system proposed for Belgium



Data linkage is an operation that is limited both in time and in its purposes. Linkage studies are carried out under law and under the watch of an independent monitoring agency. The trusted third party (TTP) certified by the government irreversibly encodes the identity of persons participating in the study. Its involvement guarantees that individual freedoms are respected. The linkage enables researchers to access fragments of individual health histories, based on data that are collected either on a one-time basis or routinely, without infringing the principle of privacy.

3. Recent changes in the situation of personal information in Belgium

3.1 E-government in general

New information and communication technologies (ICT) are transforming the relationship between citizens and their government. To achieve greater efficiency and budget control, public services are being reorganized and are using new electronic tools and developing the delivery of online services in order to get closer to the public. Since the mid-1980s, the country's different levels of government have become broadly committed to this reform, and this commitment has been ratified by a number of federal laws and legislative amendments.

The main strategy applied in Belgium is based on the sole collection of information and its subsequent validation. Then, the agency that collected the information can use it and share it with other government agencies. If necessary, the data thus validated will be reused instead of being collected again (Robben, Desterbecq and Maes, 2005). Meanwhile, the protection of the information—i.e., its security, integrity and confidentiality—is assured by structural and organizational measures as well as by appropriate technologies.

3.2 E-government applied to the health sector

The health sector, or more exactly the social security sector, is the first to benefit from this e-government strategy [2]. The process of updating the health system has accelerated, reflecting the fact that the issue of management of health systems has risen in public priorities, both in Belgium and internationally. This acceleration is mainly driven by financial considerations, based on observations of demographic and epidemiological trends.

As noted above, this process is accompanied by legislative changes that specify the authorized means and ends, in the fields of patient rights, privacy, data protection, hardware and software certifications, etc. (Devlies, Thienpont and De Moor, 2006). Complementing this, the modernization of the sector also consists in better describing and better managing the health system, in accordance with international recommendations. Along these lines, a list of performance indicators for the health system is currently being developed (Vlayen et al 2009).

As a concrete expression of this modernization, Belgium has since July 2008 had a new institution called eHealth,[3] which is both a platform for the electronic exchange of data made available to players in the health sector and a tool for simplifying the administrative relationship between health care recipients and social security managers. The eHealth platform provides a technical infrastructure which facilitates the flow of some health information and which fits into a legal environment that conforms to national decisions and European regulations. Its implementation is giving rise to a number of other initiatives at the national scale, designed to promote a dynamic approach to public health in the broad sense.

4. Examples of health data sources with a longitudinal objective or potential

The three projects discussed below may be described with reference to the eHealth platform. They provide health information that is of good quality, objective, up to date and centred on individuals. They therefore offer a potential that is conducive to the longitudinal approach.

4.1 Electronic medical record (EMR)

This is the digital version of the comprehensive medical file that a general practitioner creates and maintains on a patient in order to record important medical information regarding him or her. The data may be shared with or obtained from other health care professionals. The EMR is not only an archiving system; it is also a tool for improving the quality of the care provided, especially in cases of chronic illness. The EMR is a patient-centred compendium of information, a tool for improving decision-making and potentially a partial source of data for conducting evaluations in health economics.

The comprehensive medical record was launched in Belgium in 1999 and its electronic version a few years later. It is mainly used for primary health care, but the potential exists for supplementing it with sub-files according to medical or paramedical specialty. The major problem that it poses is that it requires ethical standards of use and the interoperability of the systems that use it. For this reason, the EMR is subject to a certification procedure, carried out under the authority of the Ministère de la Santé publique.

It should also be noted that its primary aim is clinical rather than epidemiological. A recent study conducted by a Belgian network of general practitioners reveals that at this point, the data registered in the EMR do not lend themselves to use in epidemiological analysis or in the development of government policies (De Clercq et al, 2009). However, this situation does not appear to be inevitable; if it is resolved, EMRs will become a crucial source of depersonalized clinical and personal information, potentially covering the entire population.

4.2 Clinical registries

Our country has different types of registries. The collection of targeted information must be systematic when its goal is to develop epidemiological indicators covering the population. A legal obligation to record data tends to make such a register comprehensive. It was in this context that the Registre du Cancer[4] was organized. It records the initial data and the therapeutic follow-up of new cancerous tumours, after encoding the identity of the patient affected. The role of this database is to maintain counts and to perform general and statistical control of the quality of oncological care management.

If participation in a registry is not compulsory, its focus is primarily clinical. The new Belgian registry of arthroplasties will enable orthopedists to participate in the initiative of monitoring the effectiveness of the techniques used and the quality of the prostheses that they have implanted (INAMI, 2009). The electronic application SAFE[5] concerns the management of arthritis; in particular, it can be used to monitor the patient's evolving condition, access the registry of juvenile polyarthritis and report the effectiveness and side effects of medications. Along with these clinical aspects, it also has an administrative function, since it verifies the prescription conditions for some medications and facilitates access to refunds where applicable.

[3] https://www.ehealth.fgov.be
[4] http://www.registreducancer.be
[5] https://www.ehealth.fgov.be/fr/application/applications/SAFE.html

Registries can also be constructed in connection with the EMR. When data are encoded, they are simultaneously entered in a specialized section of the EMR and in the registry. This avoids duplication in data entry and the attendant risks of error, and it lends itself to a clinical assessment that takes account of each patient's complications and co-morbidities. Beyond their immediate value for the individuals concerned, registries in the long run provide detailed and specific information constituting a macroscopic view of pathologies or events, some of them rare.

4.3 The permanent sample (PS)

This is a database that consists of an anonymized and representative sample of the Belgian population. It is made up of a few pieces of demographic information, a personal identifier and administrative data collected in connection with compulsory health/disability insurance, which covers 98% of residents. The basic sample consists of 1/40 of beneficiaries, who are followed for ten years. For the population over 65 years of age, there is a sub-sample: one person in twenty in this age range participates in the PS. The total sample is estimated at 300,000 persons, or 3% of the national population. The level of observation is the individual beneficiary. The data are updated annually and are made available to a few institutions in charge of public health and social security (INAMI, 2008).

This information source is therefore oriented toward follow-up of care consumption; it does not offer the potential for monitoring morbidity or individual habits. However, it could become a working base for linkages with data from other health-related sources.

5. Opportunities for implementing a longitudinal approach to health in Belgium

5.1 Implementing a technological infrastructure

Legislative changes have been made to allow the use of ITC for reorganizing information flows and managing personal health data. They do so by setting out a framework for use and regulation, in that they establish terms and conditions for use, along with security and protection standards and possible remedies. This legislative framework is complemented by an independent monitoring agency, the Commission de la Protection de la Vie Privée,[6] which reports to the federal parliament and is composed of experts from civil society.

The technological infrastructure that is thus beginning to unfold must be incorporated into local and regional electronic architectures for health information, such as those of hospitals, which, at the behest of federal authorities, have already developed computer-based environments to manage the records of their patients and carry on their activities. Thus on the one hand, the primary production of data is stimulated, and on the other hand, owing to the interconnection of networks and with the agreement of the various players, data can transmitted from one end of the network to the other, following a virtual path that is uninterrupted and secure.

5.2 Existence of a personal identification number

In 1983, a Registre National des Personnes Physiques (national register of natural persons) was officially created. It maintains and indexes each Belgian (as well as foreign residents) unequivocally. Even though Belgium is oriented toward decentralized management of information (in principle, each structure that generates data maintains control over those data), it is possible to link information owing to the existence of the personal identification number. Of course, the linkage of such information is carried out according to the law, and this can entail a step in which personal identity is anonymized or encoded. The different aspects of a citizen's life are no longer compartmentalized; for example, it is theoretically possible to examine health data in relation to social or environmental data.

5.3 A favourable international context

Finally, the international context offers many valuable examples of the longitudinal approach, particularly in the fields of economics and health. International institutions favour this methodology, since it can be used to verify at the individual level the appropriateness and adequacy of actions taken by governments.

[6] http://www.privacycommission.be/en/

6. Obstacles to implementation

6.1 A fragmented decision-making process

Belgium's federal structure is constructed in such a way that, like in all areas of civil life, both health care and health economics are segmented among several levels of jurisdiction. Powers are divided among the different federal entities, both exclusively (there is no provision for coordination among regional, local and federal jurisdictions) and totally (i.e., each entity alone assumes the normative role, the executive function and budgetary responsibilities for its particular sphere). Because the standards established by these different entities are equal in force, each level is entirely autonomous in the decisions that it makes and their applications. Therefore, most policy resolutions, once arrived at, collide with other, divergent priorities.

6.2 Lack of overall strategy

Consequently, the country does not have an overall health strategy whose implementation would be closely coordinated through the different echelons of the pyramid of state. Instead, Belgium moves forward by taking opportunities as they arise, with no fixed timetable but in a consensual direction established by experts in the field [3]. First, the necessary applications are created through efforts of groundbreakers within government and a few clinicians. Second, these applications are improved on a pragmatic basis as resources permit. The appropriation of these concrete results by health professionals is sometimes less than complete, and it must be must be stimulated by financial incentives or by regulation.

6.3 Different ways of thinking

The major efforts made by administrative bodies in charge of social security and public health must be matched by corresponding efforts by players in the field. To build confidence, it is necessary to inform and educate the professionals. And yet securing the support of the professionals for the changes brought down is not yet clearly included in federal priorities. On the contrary, it seems regrettable that professionals and decision-makers have different ways of thinking: the former are oriented toward the private practice of medicine, while the latter are endeavouring to get budgets under control, and the reservations expressed by medical professionals are not brought up to the extent they should be in, say, dialogues and preliminary exchanges.

Finally, in the sphere of privacy and use of personal data, the choices facing society await crucial decisions. The delays in engaging the public to participate in these debates will only put off acceptance of the system as a whole[7].

7. Conclusion

While scientists now agree on the value of developing a longitudinal vision, both in health and in other fields, it is not obvious at this point that this conviction is being translated into a political priority in Belgium. For a longitudinal component to be added to the national health information system, it will still be necessary to persuade a number of partners.

The BeLHIS project recommends that our country's health information architecture be strengthened by means of enhanced links and coordination among data-producing and data-using structures, and that greater data compatibility be developed. The project calls for prospective follow-ups within recurring surveys and recommends better access to social and health data.

The project also seeks to assert the value of the longitudinal approach in all discussions concerning the performance of the health care system and the indicators used for it, so that "longitudinal" indicators will be integrated into the evaluation grid for the health sector. In particular, it stress the need to take into consideration, via the indicators adopted, measures other than those relating strictly to economics and the hospital sphere, such as the sociodemographic and environmental aspects of the themes described.

Within the known methodological limitations, the reuse of health information for prospective analytical purposes offers the possibility of obtaining "longitudinal" indicators at less cost than by conducting traditional follow-ups of cohorts, and this reduces the obstacles to implementing a longitudinal HIS.

[7] Also see Lammens, L *et al*. Ethical implications of longitudinal data collection on both the individual and the societal level. Proceedings of Statistics Canada Symposium 2009.

References

Coppieters, Y., Bazelmans, C. and Levêque, A. (year not specified). Etude préparatoire en vue de compléter le système d'information sur la santé au moyen de données provenant d'une perspective longitudinale dynamique. Rapport de recherche. Projet AGORA AG/II/131.

De Clercq, E., Van Casteren, V., Jonckheer, P., Burggraeve, P., Lafontaine, M.F., Artoisenet, C. and Lorant, V. (2009). Are GPs' Electronic Health Records suitable for use in Public Health Research ? Scientific Institute of Public Health, April 2009, Brussels, Belgium. Available at: http://www.iph.fgov.be/epidemio/epien/re_pr_en/D_2009_2505_20.pdf

Devlies, J., Thienpont, G. and De Moor, G. (2006). eHealth strategy and implementation activities in Belgium. Report in the framework of the eHealth ERA project. December 2006.

INAMI (2008). Echantillon Permanent. Rapport des travaux de la Commission technique de l'échantillon permanent au Conseil général de l'INAMI. Note CTPS 2008/26. INAMI, Bruxelles, Belgique. Available at: http://www.inami.fgov.be/information/fr/sampling/pdf/report.pdf

INAMI (2009). Communiqué de Presse 06/05/2009. INAMI, Bruxelles, Belgique. Available at: http://www.inami.fgov.be/news/fr/press/pdf/press20090506.pdf

Robben F., Desterbecq T. et Maes P. (2005). Les apports de l'e-sécurité sociale à la simplification et à l'harmonisation de l'application de la sécurité sociale. *Actes des Journées Juridiques Jean Dabin - Colloque 60ème anniversaire de la sécurité sociale*, 2005. To be published. Available at: http://www.law.kuleuven.be/icri/frobben/publication%20list.htm

Vlayen J., Leonard C., et al. (2009). Belgian Health System Performance : How do we do? Projet N° 2008-50 (KCE-ISP-INAMI). Preliminary report. May 2009. Not published.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Ethical Implications of Longitudinal Data Collection on Both the Individual and the Societal Level

Lammens Lies, Deboosere Patrick, Cols Florence, D'hoore William, Ingenbleek Ann, Coppieters 't Wallant Yves and Levêque Alain[1]

Abstract

Technological innovation confronts researchers in industrialized societies with a fundamental ethical dilemma: the 'knowledge' versus 'privacy' dilemma. Longitudinal data offer particular interesting advantages for research, but they contain elements which could be threatening (or perceived as being threatening) to individuals' privacy. In an earlier paper[2] we developed a conceptual framework on ethical implications of (health) data collecting projects. We reflected on the content of policy goals developed within the context of an ever more detailed data collection, and pointed to the potentially abusive use of collected data in the short (threatening people's privacy) and in the long run (threatening democracy). Here we confront this framework with the two differently organized statistical systems of the UK and Denmark.

Key Words: Data Collection, Privacy, Democracy, Longitudinal, Statistical System.

1. Introduction: research in modern societies

1.1 Technological and societal context

Technological developments have altered the traditional research environment considerably. More and more information is acquired, analysed and stored digitally. The Internet enables information to be transferred globally and more rapidly than ever before. Increasingly sophisticated statistical techniques are used to manage information.

Consequently, the costs of data collection, storage, analysis, integration and dispersion have decreased and the population can be monitored in a more systematic and detailed way. At the same time, these technological innovations pose a challenge to our societies: people's privacy rights might get endangered. Privacy concerns have been around at least for 50 years, i.e. ever since the development of large electronic databases. As an increasing amount of detailed data on individuals' lives and behaviour have been collected and electronic information can be more easily copied, transported and spread compared to paper records, consequences of misuse may be much bigger now than in the past (Myers et al, 2008, p.794).

Furthermore, the societal context of data collection has changed. Since the September 11 attacks governmental control of potential national security threats seems to have become more important than the protection of people's privacy. 9/11 has changed the role of information and the way people perceive privacy rights remarkably (Duncan, 2004, pp.4-5).

1.2 Data collection

Simultaneously, the way data are collected has changed considerably. More and more databases cover the whole population. Most information within these databases is obtained through systems of administrative records. A growing amount of records gets linked across databases. And there is an increasing interest in analyzing individual life histories by exploiting longitudinal data.

The various benefits offered by these recent research techniques represent a challenge for data confidentiality. For example, the use of administrative data and linking processes - requiring access to micro data - pose a great challenge in obtaining 'informed

[1]Lies Lammens, Vrije Universiteit Brussel, Pleinlaan 2 - 1050 Brussels, Belgium (Lies.Lammens@vub.ac.be); Patrick Deboosere, Vrije Universiteit Brussel, Belgium; Florence Cols, Université Catholique de Louvain, Belgium; William D'hoore, Université Catholique de Louvain, Belgium; Ann Ingenbleek, Université Libre de Bruxelles, Belgium; Yves Coppieters 't Wallant, Université Libre de Bruxelles, Belgium; Alain Levêque, Université Libre de Bruxelles, Belgium.

[2] Policy versus privacy: an analysis of ethical issues in a health monitoring project (XXVI IUSSP International Population Conference, Marrakech, 2009)

consent', i.e. the right to give or deny consent for the use of information about oneself, the classical solution to overcome the 'knowledge versus privacy'-dilemma. Moreover, longitudinal data contain much more detailed information on the characteristics and behaviour of individuals than cross-sectional data. Hence, the risk of disclosure is higher.

1.3 Different societal approaches

Policies are shaped by their historical background. Confronted with a similar technological evolution, governments are developing specific and at times very different policies in accordance with the existing institutions, statistical system and prevalent customs concerning privacy.

The Danish system is internationally recognized for its highly centralized production of statistics. Its main producer of statistics is Statistics Denmark (°1850) which is responsible for the coordination of all official statistics in Denmark, supervises central public registers and must be consulted by all public bodies with respect to design and content of new registers (Møller et al, 2001, p.2). Since the set up of the system, the high degree of centralization was compensated for by very strict rules of data access. During a long period of time this resulted in a highly restricted data access for scientific research, at least for researchers outside the statistical office. Gradually, Statistics Denmark has widened the access to its information. Nowadays, the rules of access to data are still very strict but are accompanied by an active policy of releasing data for research. The highly centralized Danish system allows a coherent strategy in this regard.

The UK system is decentralized, reflecting the structure and evolution of the UK government, and perhaps also reflecting the historical assumption that official statistics mainly exist to meet the needs of government. Gradually, this view has moved over to a wide acceptance that statistics must serve society as a whole. Today, the Office for National Statistics (°1996) represents the main statistical institute in the UK, but some 80% of the UK statistics are under the responsibility of organizations other than the ONS, such as the devolved administrations and government departments (Dunnell, 2007, p.1).

These historical backgrounds have shaped the actual statistical systems of Denmark and the UK as well as the general attitude and sensibility towards data collection, disclosure and protection in both countries.

Denmark has developed a statistical system which is almost entirely based on administrative registers and is considered a pioneer country worldwide in taking register-based censuses. This use of administrative data is facilitated by the introduction of unique identifiers (for persons, real estates and businesses) used in all public administrations. The uniqueness of these identifiers allows to directly link information across statistical units and/or subject matters.

In the UK, the importance of the protection of people's privacy has historically been estimated more important than the mass collection of information on the population. Nevertheless, the number of administrative databases has grown gradually, though focusing on service provision rather than citizen identity, neither being publicly accessible nor offering complete coverage of the population. Most of the UK data provision today comes from traditional decennial censuses and punctual, at times very large-scale, surveys on a large range of subjects (Dunnell, 2007, p.2). As a result of this policy the UK probably has one of the largest and best collections of population surveys in Europe.

2. Framework on ethical implications of health monitoring

2.1 What are the implications of these different strategies of data collecting in the field of health monitoring?

Assessing population health has become a major concern of modern governments. The reason for this is not only budgetary, although ageing populations are evidently adding pressure to the health care expenditures. In a modern democracy, the well-being of the population is the main political goal and the ultimate justification of governmental policies, and health is undoubtedly the cornerstone of this well-being. Public expectations, budgetary constraints and the complexity of an evidence-based modern health policy have propelled the monitoring of population health on the statistical agenda.

Individual longitudinal follow-ups are of particular interest to health monitoring. They are not only crucial to correctly assess the evolution of population health and to avoid fallacies due to the use of aggregated population data, but they also belong to the few methods available to establish causal relationships. It is evident that individual data and longitudinal research, be it through periodical follow-up or through post factum linkage of data, are particularly challenging to privacy and data confidentiality, and concerns about that have rightly been raised within the scientific community and the public at large.

However, in evaluating the implications of data collection it is not only important to consider the actual tensions. A proper evaluation should also include a long-term view. Hence we argue that it is important to add a time dimension in the evaluation

of every extensive data collection process. We can summarize our approach in a two by two table where time is introducing an uncertainty, illustrated by the question-marks.

	Health monitoring	**Societal implications**
Short-term dilemma	1. Health policy	2. Privacy deficit
Long-term risks	↓ 3. 'Health police'?	↓ 4. Democracy deficit?

2.2 Short-term dilemma: policy versus privacy

Looking at the short term, health monitoring projects might bring forth the dilemma between blocks 1 and 2. On the one hand, the accumulation of personal data within an increasingly sophisticated and automated health information infrastructure has led to a better monitoring and knowledge of the population. Consequently, more accurate policies can be implemented, eventually bringing along significant health benefits and a higher quality of life. On the other hand, the recent technological opportunities and increasingly detailed monitoring of people's lives and health behaviour contain elements that could threaten - or be perceived as being threatening to - private life in society. Citizens might start to worry about being controlled or 'big brother'-wise being 'watched'.

The fact that privacy is a fundamental right of every citizen seems to be a widespread and universally accepted idea. But the implementation of this right brings along considerable debate. The balance between the public's right to privacy and society's need to know entails an important ethical dilemma in research nowadays. The health sector specifically is faced with this dilemma, since research in this sector contains data reflecting some of the most personal and sensitive aspects of individuals' lives. Although press articles concerning data losses and security breaches may create the impression that privacy is more often sacrificed for the benefit of health policy, overall authorities seem to be very aware of this tension and more often have the tendency to decide in favor of privacy concerns. Strange enough, in the aftermath of 9/11 and being confronted with the reality of terrorism, governments and the public opinion have accepted large breaches in the protection of privacy in the name of security.

2.2 Long-term <u>risks for policy and privacy concerns</u>

Thus far, we considered the debate concerning data protection from a limited point of view, by mainly focusing on the privacy right of individuals, as 'the right to be protected from anyone else being informed about their personal life'. However, long-term implications of the way in which statistics, the collection of personal data and administrative systems are organized, should receive even more attention (block 4). For example, it is often overlooked that information processes might in the long run contribute to the introduction of practices or policies inconsistent with our traditional understanding of democracy, as for example discrimination or stigmatization of individuals or specific subpopulations. In the context of health research, sensitive personal data such as someone's sexual orientation, ethnicity or cultural background might be particularly relevant, but it is exactly this kind of personal information that is often used for discrimination.

We cannot be certain that data collected and protected properly nowadays will not be misused in the future. We distinguish at least three different dangers. Firstly, possessing information about persons entails power which can be used to control people. Secondly, the greater the importance of data to the securing of power, the stronger the incentives to those in power to ensure that the collected data present a favourable picture (Prewitt, 1985, p.116). Thirdly, it might become tempting to transform statistical records into administrative and surveillance records. Administrative use of statistical information is not only disastrous for the public's perception of the impartial collection of data and production of statistics. Also, blurring the borders creates a concentration of power and might facilitate potential abuse. Especially in times of crisis or political instability it might be tempting for people possessing information to use that information for undemocratic purposes, e.g. to harm or favour certain groups (Seltzer and Anderson, 2001).

Implementing measures to prevent potential undemocratic abuse of information is essential in the short and the long run. Crucial safeguards involve legal regulations, technical barriers, organizational measures and ethical guidelines. It is clear that none of these measures offer an absolute guarantee for a democratic use of data. However, all these measures together can help to suppress data abuses by raising the financial, personal or political costs of such misuse.

Historically, the concern of an undemocratic abuse of data has often been at the forefront of parliamentary interventions in almost each country where information technology and data collection have been discussed. Strange enough, not only did this concern fade as we all became more familiar and acquainted to the daily use of information technology. Gradually, it also moved towards a discussion on privacy instead of democracy. Long-term concerns on the democratic impact of IT tend to disappear, one reason being that it is often presumed that protecting privacy is at the same time also protecting democracy and democratic values. However, although many privacy protection measures may help to ensure democracy, they are definitely conceived from a different perspective and not necessarily most adapted to protect democracy in the long run. It is necessary to make a distinction between these two objectives or risks and to discuss the implementation of statistical systems with both in mind.

3. How are the statistical systems of Denmark and the UK dealing with these conflicting aspirations?

3.1 Short-term dilemma: policy versus privacy

Although the aforementioned concerns have been expressed both in the UK and in Denmark, these two countries have developed different implementations of their statistical systems. By analyzing the two systems, is it possible to conclude that one model is better suited than the other to meet the competing aims and concerns of modern data processing?

In the short run, the Danish statistical system is known to offer many advantages for research as well as policy purposes, such as the possibility to collect a great range of detailed data on various domains covering the whole population, and the fact that the system is relatively low-cost and highly efficient.
On the other hand, the system is often perceived as privacy-threatening. The main reasons therefore are the centralized organization - information is concentrated in one place, under one authority -, the use of unique identifiers and the possibility to link information across databases. Apart from that, the perception of the privacy-threatening character of this system has formed the basis for many protective measures and a difficult access to Danish micro data for research during many years. The efficient statistical system has for a long time not been a very fertile ground for research of the scientific community at large.

The UK system is less efficient and more expensive than the Danish one. But it offers some particular advantages such as the extensive use of survey data containing much more in-depth information on specific topics. As far as privacy is concerned, the decentralized system is perceived as safer. The UK system offers fewer possibilities to link information, excluding some important privacy issues.

Strange enough, the public trust in the UK system is low. Recent series of data or information losses released in the UK press have incited this. Strict rules and a transparent policy on the other hand have resulted in a high public trust in the Danish system.

3.2 Long-term risks for policy and privacy concerns

In the long run, it is difficult to estimate how both systems will evolve. Technology is progressing with an enormous velocity, making it always easier to collect data at a large scale and mapping the conditions and characteristics of a population. Therefore, future research potential as well as its implications on data confidentiality are difficult to foresee.

The fundamentals of the actual Danish statistical system were set when democracy was introduced in Denmark at the end of the 19th century, and seem to have been profoundly influenced by this ideology. The necessity of transparency, openness, reliability, user-friendliness and professional competence of the statistical authorities seemed recommended in response to the ancient regime in which statistics at times had been kept secret. Yet, what might happen to the system, containing a huge amount of personal information on the Danish population, if the societal conditions or political context become instable, as for example in times of war or changing power relations?

The UK statistical system has been gradually constructed, rather guided by successive reforms consistent with the predominated political context than based on a strong ideology or a radical change at a certain time in history. Since 9/11 the focus on national security has increased considerably in the UK, while decreasing the importance given to the protection of people's privacy. Also, recent steps towards a more integrated information system seem to be taken gradually.

Denmark and the UK approach the organization of their statistical system and the matter of data confidentiality in a very different, at times opposite, way. No system can totally prevent potential data abuse, and the systems of Denmark and the UK both offer particular advantages. But on the whole, for both research and policy purposes as well as for the protection of the data being collected, the Danish system seem to offer more advantages than the UK system, even if the Danish one at first sight

appears more dangerous. By publicly and explicitly debating the issue of data protection as well as by implementing accurate safeguards, Denmark has until now succeeded in preventing important data abuses or losses to happen, even if its statistical system is highly centralized and contains more detailed information on its citizens than the UK system. As such, the Danish example seems to show that an extensive personal data collection can be compatible in the short and in the long run with a society in which the protection of people's privacy and democratic values are perceived as fundamental.

4. Conclusion

The success of our industrialized societies, illustrated by the continuous growth in average life expectancy, and more importantly in healthy life expectancy, is primarily related to the organization and functioning of our societies itself. But the organization of knowledge in our societies, i.e. the way in which knowledge is spread, education is given and research is organized, has also considerably contributed to this success.
Due to technological advances in modern societies, it has and will become easier to gather health data on a population level. That is why it is now more important than ever to consider the consequences of the mass collection of, at times very personal, information on individuals. Longitudinal data offer particularly interesting research advantages in trying to relate social outcomes to underlying causes, but they pose a great challenge on data confidentiality. They contain detailed information on individuals' lives and behavior, and might reveal patterns disclosing a personal identity. Besides, to construct longitudinal data individual data need often to be linked repeatedly, increasing the risk of identification as well.

Given the growing ability of our information systems to capture reality into data, it is of the utmost importance that the evaluation of data collecting projects includes potential evolutions in the long run. Finding the right balance between the demand for knowledge in our societies and the protection of citizens' privacy and a democratic use of data seems to be difficult if not impossible. There is certainly no ideal solution to overcome this dilemma.

Besides up-to-date technical, legal and organizational measures, we perceive a democratic control of data as crucial to protect against a potential abuse of data, especially since technological possibilities are progressing with an unpredictable velocity. After all, not the information itself is dangerous for people's privacy concerns, nor for the democratic system. Moreover, a highly and accurately informed public is a fundamental condition of democracy. The problem arises when only a select group of people gets access to the collected information and uses this information for undemocratic purposes. If, on the contrary, the collected information is shared by many, and the analysis and use of that information is dealt with in a transparent and open way, information itself is not incompatible with privacy and democracy values. Furthermore, in search of a balance between data collection or protection against potential data abuse, it is important to discuss this ethical issue explicitly and to stimulate the awareness concerning potential problems.

References

Col,s F., Lammens, L., Ingenbleek, A., D'hoore, W., Deboosere, P, Coppieters, Y. and Levêque, A. (2009). *Gestion des données à caractère personnel. Réflexions sur le système d'information sanitaire belge, à partir d'une analyse comparative des systèmes britannique et danois.* Working paper N° 2, Projet BeLHIS, AGORA AG/00/139, Bruxelles, 2009.

Duncan, G.T. (2004). *Exploring the Tension Between Privacy and the Social Benefits of Governmental Databases*, paper presented at the Georgetown University Law Center Conference on Security, Technology, and Privacy: Shaping a 21st Century Public Information Policy, Washington DC.

Dunnell, K. (2007). *Evolution of the United Kingdom statistical system*, paper presented at the UN Seminar on Evolution of National Statistical Systems, New York, USA.

Myers, J., Frieden, T.R., Bherwani, K.M. and Henning, K.J. (2008). Privacy and Public Health at Risk: Public Health Confidentiality in the Digital Age, *American Journal of Public Health,* 98, 5, pp. 793-801.

Prewitt, K. (1985). Public statistics and democratic politics, in J.J. Smelser and D.R. Gerstein (eds.), *Behavioral and Social Science: Fifty years of Discovery*, Washington, D.C.: Nat. Academic Press.

Seltzer, W. and Anderson, M. (2001). The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses, *Social research*, 68, 2, pp. 481-513.

Statistics Denmark (2008). *Numbers on time. Introduction to Statistics Denmark*, [http://www.dst.dk/HomeUK/About/introSD.aspx, consulted 12/04/2009].

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Contribution of administrative and medical administrative databases to the Constances cohort

Alice Guéguen, Rémi Sitta, Laetitia Bénézet, Gaëlle Santin, Jean-Louis Lanoë,
Marcel Goldberg , Marie Zins and les Centres d'examen de santé de la Sécurité sociale[1]

Abstract

Constances is a French cohort that will include 200,000 people. It has a two-fold objective: public health and epidemiological research. Information of interest will not only be collected from individuals, but will also be drawn from administrative and medical-administrative databases. As regards data collection from individuals, the expected response rate is low (approximately 15%), and corrections for non-response will be made using auxiliary information available on respondents and non-respondents. This information, which will predict non-response, will be extracted from the above-mentioned data bases.

Key words: administrative data, non-response, epidemiology, longitudinal survey, cohort, weighting.

1. Introduction

Epidemiology is the science that studies the frequency and the temporal and spatial distribution of health conditions in populations (descriptive epidemiology), as well as the role of factors that determine those conditions (etiological or analytical epidemiology). Currently, research questions in epidemiology mainly centre on identifying low risks concerning infrequent or low-intensity exposures for which the latency times between exposure and illness can be quite long. It is therefore necessary to form very large cohorts (or panels), with individual follow-up that can extend over decades.

Constances is an epidemiological cohort formed by the Institut national de la santé et de la recherche médicale (Inserm). It will consist of 200,000 persons who are aged 18 to 69 at the time of their inclusion and are covered by the general social security system. Constances is designed to provide information for public health purposes, and it will contribute to the development of epidemiological research, mainly concerning the social and occupational determinants of health. A pilot phase including approximately 3,000 subjects began in September 2009 and will end in March 2010. Formation of the cohort will begin in late 2010. The expected non-response rate is high at around 85%. This is not uncommon in epidemiological studies, since they are usually volunteer-based; therefore, the terms participation/non-participation will be used instead of response/non-response. Biases due to non-participation will be dealt with by using available auxiliary data on participants and non-participants, drawn from French national administrative and medical-administrative databases.

The second part examines biases due to selection effects in epidemiology. The French databases used to correct non-participation biases will be described in the third part. The fourth part will be devoted to describing the Constances cohort. The last part will contain points for discussion.

2. Bias due to selection effects in epidemiology

2.1 Description of biases

In epidemiological surveys, the data are most often collected directly from the subjects themselves, owing to the nature of the information collected (alcohol and tobacco use, food consumption, etc.). For ethical and legal reasons, active participation in the survey cannot be made compulsory; as a result, the surveyed population differs from the target population owing to phenomena related to non-participation, inclusion and attrition during follow-up.

[1] Alice Guéguen, Rémi Sitta, Jean-Louis Lanoë, Marcel Goldberg, Marie Zins, Unité mixte Inserm-Cnamts 687 / Centre de recherche en Epidémiologie et Santé des Populations, Hôpital Paul Brousse, Bâtiment 15/16, 16 avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France ; Laetitia Bénézet, Gaëlle Santin, Marcel Goldberg, Institut de veille sanitaire, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex, France.

When the objective is descriptive in nature, the biases introduced by non-participation may have to do with estimating the frequency of the illness or of exposure to a risk factor. When the objective is analytical in nature, these biases may influence the estimate of the association between exposure and illness. There is a bias in the estimate of the frequency of the illness or of exposure, or the exposure-illness association, if the probability of being ill (or exposed) is not independent of the probability of being included in the study, or if the exposure-illness relationship differs between subjects included and those not included. The bias is generally greater when it is a matter of estimating a frequency than in the case of estimating an association.

2. 2 Solutions

Whatever solution is implemented, it requires having a thorough knowledge of the factors that predict participation in epidemiological surveys. The studies of Goldberg *et al* (2001) and Zins *et al* (2008) have shown that the factors related to participation are essentially age, social category and marital status, whereas the factors related to attrition are primarily health-related behaviours (alcohol and tobacco consumption) and the onset of serious health events.

In analytical epidemiology, the method that is usually used consists in including participation-related factors in the regression model explaining illness by exposure. The advantage of this method is that it makes it possible to dispense with measuring these factors on non-participants. However, it should be noted that in some cases, this approach can pose problems (Hernan, 2004) that prevent its use.

In random sample surveys, the method used in both descriptive and analytical studies consists in reweighting sampling weights to compensate for biases due to non-participation. To do this, however, it must be possible to measure these factors in both participants and non-participants. This need is met by the administrative and medical-administrative databases described below, since they are comprehensive and contain many pieces of information that serve as proxies for the factors influencing participation in epidemiological surveys.

3. French administrative and medical-administrative databases

In France there are several comprehensive national databases, containing a range of information that can be linked to factors influencing participation. These include the socio-occupational databases of the Caisse Nationale d'Assurance Vieillesse (Cnav), health data from the Système national d'information inter-régimes de l'assurance maladie (SNIIR-AM) and the database of Inserm's Centre d'épidémiologie des Causes de Décès (CépiDc). All these information sources can be matched using an encrypted form of the Numéro d'identification au Répertoire (NIR), a unique national identifier used in all these databases (Goldberg, 2008).

3.1 Socio-occupational data of the Caisse nationale d'assurance vieillesse

The role of the Cnav is to ensure entitlement to retirement benefits for any individual who, at least at one point in his/her life, has belonged to the general social security system, which covers approximately 85% of the French population. For this, the Cnav has developed a system for collecting and processing social data from the different plans and agencies that manage social benefits. To build and enrich its databases, the Cnav regularly receives data transmitted by employers, including employers of domestic personnel. It also receives information from various agencies on periods not in the labour force (owing to unemployment, illness or maternity). These data, registered prospectively, serve as a basis for calculating retirements and are therefore complete and especially well validated, particularly for the most recent periods, and their quality (completeness and accuracy) is steadily improving over the years with the automation of data collection at source.

3.2 Data from the Système national d'information inter-régimes de l'assurance maladie

The SNIIR-AM contains individual data that are medicalized, structured and coded in a standardized manner:

- Health care consumption data: these data come from the health insurance payments database; they contain no information on the nature of the illnesses treated, and by definition they exclude self-medication and services for which no claim is filed.
- Hospitalization data: the database of the PMSI (Programme de médicalisation des systèmes d'information des hôpitaux) brings together, for each hospital stay, the main diagnosis, any associated diagnoses and the most costly diagnostic and medical procedures. Diagnoses and medical procedures are coded in a standardized fashion.
- Data concerning industrial accidents, occupational injuries and chronic conditions, with the latter representing the 30 most serious illnesses.

3.3 Mortality data from Inserm's Centre d'épidémiologie des causes de décès

Mortality data—vital status and cause of death—can be obtained from Inserm's Centre d'épidémiologie des causes de décès.

4. The Constances cohort

4.1 Objectives

Constances (www.constances.fr) is designed as a general cohort serving a very broad purpose. In other words, it has no specific objectives in terms of hypotheses concerning pathologies and/or specific risk factors, but it is instead intended to serve as a platform for epidemiological research. Furthermore, the duration of the project has not been determined, and Constances will be the object of a longitudinal follow-up with no time limit, so as to be able to study the effects of very long-term risk factors and to take account of evolving knowledge and techniques, which are constantly raising new scientific questions.

Constances has a twofold objective of public health and research, and it will be open to the scientific community. The research objectives of Constances are largely centred on studying the occupational and social determinants of health across four main themes: occupational risks, social inequalities in health, aging and women's health. Constances also has public health objectives; as regards the objectives of epidemiological monitoring, a collaborative relationship will be established with the Institut de veille sanitaire and more especially with the Département santé travail to monitor occupational risks via the Coset program (Santin, 2009).

4.2 Network of Centres d'examen de santé

The cohort of Constances (CONSultants des Centres d'Examen de Santé) will be based on the network of Centres d'examen de santé (CES), which are open to participants in the general social security system. The latter can, every five years, receive a thorough health examination by going to their local centre. Seventeen centres throughout France have agreed to participate in Constances. The target population, living in the seventeen departments in which those centres are located, reflects the makeup of the French population with respect to age, sex and social category.

4.3 Sampling plan and data collection

A sample of two million subjects, belonging to the population covered by the general social security system and living in the seventeen departments in the Constances plan, will be drawn in a random sampling with unequal probabilities and stratified by age, sex, labour force status, social category and department The persons randomly drawn will be invited to participate in Constances by undergoing a medical checkup in one of the seventeen CESs. On the basis of earlier surveys, in which subjects were invited to visit a health examination centre, it was estimated that participation in Constances would be approximately 10%-15%, and estimates are available on the probability of participating based on the subject's age, sex and social category. Consequently, the probabilities of being drawn will be unequal, and therefore subjects who are expected to have a lower participation rate will have a proportionally higher sampling weight. The objective is for the expected numbers of participants to be proportional to those of the target population according to age, sex and social category.

It is expected that in total, there will be approximately 200,000 participants in Constances, for whom there will be data from the medical checkup as well as data collected by regular and self-administered questionnaires. A sample of 400,000 persons will be drawn from among the 1,800,000 non-participants. For the 600,000 persons in the two samples (participants and non-participants), the data will be available from the Cnav going back indefinitely and from the SNIIR-AM going back two years. For the latter data, since the information is stored over rolling three-year periods, these data will be available for the current year and for the two years preceding inclusion (or non-inclusion) in Constances. During the longitudinal follow-up, the two samples (of participants and non-participants) will be followed passively in the Cnav and SNIIR-AM databases. For participants, active follow-up will be carried out via annual self-administered mail questionnaires. It is also expected that participants will be re-invited to undergo a health examination every five years. The vital status of participants and non-participants, as well as causes of death, will be sought each year from the Cépidc.

4.4 Adjustments for non-participation

Because of the wealth of information available in the Cnav and SNIIR-AM databases, it will be possible to model the probability of participation acceptably. Homogeneous response groups (Eltinge, 1997) will be created, so as to calculate corrections for non-participation. This operation may be followed by a calibration, after the margins have been specified.

Probabilities of participation will first be estimated using the information available at the time of inclusion, but they may also be updated using information collected during the longitudinal follow-up. For example, certain health behaviours, such as high alcohol and/or tobacco consumption, are known to be predictors of non-participation in epidemiological surveys. However, data relating to the consumption of these substances will be available only for participants in Constances, since the data will be collected by the self-administered questionnaire; also, no proxy exists for these variables, at least at the time of inclusion, in the databases of the SNIIR-AM. On the other hand, we know that these behaviours can generate health events over the longer term, such as cancers of the lung or upper digestive track. These events, occurring after inclusion or non-inclusion in Constances, will be collected for all subjects in the SNIIR-AM databases. It will therefore be possible to refine the probabilities of participation as the follow-up continues.
Corrections for attrition will be calculated by similar methods, and more data will be available on participants, both at the time of their inclusion and during follow-up.

5. Discussion

The administrative and medical-administrative databases of the Cnav, the SNIIR-AM and the CépiDc are an integral part of the Constances system; not only will they shed light on the selection effects due to the use of volunteers, but they will also provide data of interest that are useful in themselves.

However, they have limitations: for example, the medical diagnoses contained in the SNIIR-AM databases are not systematically validated, since those databases are designed for management purposes. While this could pose a problem if one were interested in medical diagnoses as such, that will not be the case if these diagnoses are used to estimate probabilities of participation, since it can be hypothesized that the distribution of the errors associated with these diagnoses will not differ between participants and non-participants.

The administrative and medical-administrative databases used in Constances offer several advantages:

- the databases are national in scope and comprehensive;
- the information is available going back two years before inclusion in the SNIIR-AM data and going back indefinitely for the Cnav data;
- the information, collected prospectively, will be available at the time of inclusion and during the follow-up;
- the information is coded in a standardized fashion that is independent of the subject.

Apart from the strategy described above, designed to compensate for the selection biases due to the use of volunteers, supplementary surveys of non-participants will be conducted on a parallel track. The intention is to collect information on the reason for non-participation, certain occupational exposures and illnesses, and a few behaviours known to be related to participation. These studies will add other elements to the discussion of biases due to the use of volunteers.

References

Eltinge, J. and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells, with an application to income nonresponse in the US Consumer Expenditure Survey, *Survey Methodology*, 23, pp. 33-40.

Goldberg, M., Chastang, J.F., Leclerc, A., Zins, M., Bonenfant S., Bugel, I., Kaniewski, N., Schmaus, A., Niedhammer, I., Piciotti, M., Chevalier, A., Godard, C. and Imbernon, E. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population, *Am J Epidemiol,* 154, p. 373-84.

Goldberg, M. and Luce, D. (2001). Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et consequences, *Rev Epidemiol Santé Publ*, 49, p. 477-492.

Hernan, M.A., Hernandez-Diaz, S. and Robins, J.M. (2004). A structural approach to selection bias, *Epidemiology,* 15, p 615-625.

Goldberg, M., Quantin, C., Guéguen, A. and Zins, M. (2008). Bases de données médico-administratives en épidémiologie: intérêts et limites, Courrier des statistiques, 124, p59-70.

Santin, G., Bénézet, L., Guéguen, A., Sitta, R., Gauvin, S., Sarter, H., Razafindratsima, N., Zins, M., Geoffroy-Perez, B. and Goldberg, M. (2009). Strategies for studying non-response bias in the Coset and Constances cohort, *Proceedings of Statistics Canada Symposium 2009, Longitudinal Surveys: from Design to Analysis*.

Zins, M., Chastang, J.F., Coeuret-Pellicer, M., Leclerc, A., Bonenfant, S., Guéguen, A., Ozguler, A. and Goldberg, M. (2009). Factors associated with participation in the GAZEL cohort, *Proceedings of Statistics Canada Symposium 2009, Longitudinal Surveys: from Design to Analysis*.

DATA COLLECTIONS ISSUES IN LONGITUDINAL SURVEY

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Keeping in touch with mobile families in the UK Millennium Cohort Study

Lisa Calderwood[1]

Abstract

This paper focuses on the problem of locating mobile families in Millennium Cohort Study (MCS), a large-scale longitudinal study in the UK, and examines what proportion of families who move between waves are successfully located through the study's tracking procedures. It also examines the effectiveness of techniques designed to pick up address changes prior to the start of fieldwork for a particular wave compared with interviewer tracking in the field and investigates some of the factors associated with success or failure to locate mobile families. It shows that over 9 in 10 mobile families were successfully located between wave 2 and wave 3 of the study with the majority (55%) located before the start of fieldwork for second wave. Although some differences are found in the observable demographic and socio-economic characteristics of mobile and non-mobile families, very few of these characteristics are associated with the success or failure to locate families.

Key Words: Tracking, Attrition, Mobility, Non-Response, Millennium Cohort Study.

1. Introduction

One of the main analytic benefits of longitudinal surveys is that they offer researchers the opportunity to study change over time. Attrition from longitudinal surveys can lead to bias in the findings from the study if sample members who drop out over time are systematically different to those who remain in the study. A particular concern is that if the factors associated with sample loss are themselves associated with the substantive processes which the study is aiming to measure over time, this can lead to biased estimates of change. Lepkowski and Couper (2002) distinguish between three different sources of attrition: failure to locate, failure to make contact having located and failure to co-operate having contacted. This paper focuses on sample attrition due to failure to locate. One of the main reasons that longitudinal studies aim to track sample members who move is that the dynamics of residential mobility, and the processes related to it such as relationship and employment change, are of substantive interest and failure to locate sample members who move may lead to biased estimates of change in these and other important domains.

Information from the Millennium Cohort Study (MCS), the fourth in the series of internationally renowned birth cohort studies in the UK, will be used to examine the location problem. Previous research on the MCS reported in Plewis (2007a) and Plewis et al. (2008) showed that residential mobility between wave 1 and wave 2 was a predictor of non-response at wave 2 but conditional on successful location, mobile families were no more likely to refuse to be interviewed at wave 2 than non-mobile families. The research also showed that although residential mobility was associated with non-response at wave 2, it was not a predictor of permanent drop out from the study.

This paper will examine what proportion of families who move between wave 2 and wave 3 are successfully located through the study's tracking procedures and, in particular, examine the effectiveness of techniques designed to pick up address changes prior to the start of fieldwork for a wave compared with interviewer tracing in the field. It will also examine some of the factors associated with success or failure to locate mobile families.

2. Locating sample members in longitudinal surveys

The problem of locating sample members in longitudinal surveys is related to individual's propensity to move and, conditional on moving, to be located. Couper and Ofstedal (2009) offer a general model to help understand the location process which hypothesises that the main factors affecting the propensity to move are person-level factors such as age, family circumstances, employment and housing situation, and societal-level factors such as the general level of mobility and degree of urbanisation. The propensity to be located, on the other hand, is influenced by survey design factors, such as the interval between waves and tracking procedures and structural factors, such as the availability of population registers, mail forwarding rules and the portability of phone numbers. Couper and Ofstedal provide a review of the literature in relation to the likelihood of moving

[1]Lisa Calderwood, Institute of Education, 20 Bedford Way, London, United Kingdom, WC1H 0AL (l.calderwood@ioe.ac.uk)

showing that mobility rates vary both within and between countries and that a variety of demographic and socio-economic factors are associated with mobility. They also discuss the structural factors and survey design factors which are likely to associated with the ability to locate sample members who move. This includes a useful review of tracking procedures which are commonly employed on longitudinal surveys which distinguishes between retrospective tracking, designed to find sample members with whom contact has been lost and prospective tracking, designed to prevent the loss of contact by keeping details up to date and between office and field based tracking. The authors note that although most longitudinal surveys devote considerable resources to tracking mobile sample members and have developed highly successful procedures for minimising attrition through failure to locate[2], there is very little methodological evidence on the relative success, and cost-effectiveness, of different tracking procedures.

As well as reporting on the overall effectiveness of the tracking procedures on the MCS, this paper aims to advance the literature in this area in two ways. Firstly, by examining what proportion of mobile sample members are located prior to the start of fieldwork for a wave compared with during fieldwork and secondly, by evaluating whether individual-level demographic and socio-economic variables are associated with the propensity to be located. It is well-established that these types of variables are related to the propensity to move but there is very little theory or evidence about whether they are related to the propensity to be located. Tracking is sometimes characterised as something that is 'done to' sample members but they are, of course, active agents in this process and so it is reasonable to hypothesise that tracking procedures may be more effective for certain types of people, distinguishable by their observable demographic and socio-economic factors, than others.

3. The Millennium Cohort Study

The Millennium Cohort Study is a longitudinal birth cohort study following the lives of over 19,000 children in the UK who were born in 2000 and 2001. The sample was drawn from the Child Benefit register and was initially geographically clustered by electoral ward with an over-representation of areas with high proportions of Black or Asian families, disadvantaged areas and areas in the three smaller UK countries. Child Benefit is a universal benefit payable to families with children and payments begin from the time of the child's birth. There have been four waves of the study so far, when the cohort member was aged 9 months, 3, 5 and 7. At all waves, interviews were conducted with both resident parents and from the second wave onwards data has been collected directly from the cohort member. The study has also collected data from siblings and teachers as well as consents to link to administrative data for cohort member, parents and siblings. More information about the design of the study can be found in Plewis (2007b).

The MCS employs a variety of both prospective and retrospective tracking procedures. The study provides a Freephone number, email address and a website through which cohort families can inform us if they change their address or contact details. Contact details for study members are updated annually. In survey years, this is done during the interview. In non-survey years, this is achieved through the mailing of a reply-slip which is pre-printed with all of the families' contact details i.e. address, names, phone numbers, email address and stable contact details. They are asked to return the form, either with corrections and/or additions or to confirm that the information is correct and complete. The forms are returned by around 75% of families after two reminders. Undelivered mail, usually indicating that the family has moved, is returned to the study by the post office which triggers retrospective office-based tracking. Multiple attempts are made to contact sample members, their nominated 'stable' contact person and the current occupiers of the address previously occupied by sample members through telephone, mail, email and text messaging. We also use publicly available Post Office, electoral and phone records which are available on the internet or through specialist software and through other administrative data sources such as the National Health Service Central Register and Child Benefit Records. During the fieldwork for the study, interviewers also attempt to track families who have moved. Interviewers in the field are able to make personal visits to the last known addresses of cohort members and, if local, their stable contacts in addition to attempting contact through phone and mail. Interviewers can also attempt to trace through neighbours, follow visual clues at the property e.g. 'for sale' signs which can lead to tracking through estate agents and use other sources of information which are available locally.

From a survey management and budgetary perspective, it is much more desirable to find out that a sample member has moved and ideally, find a new address for them in advance of fieldwork for a wave than during fieldwork because field-based tracking by interviewers is generally more expensive than office-based tracking and can lead to delays in fieldwork due to the extra time needed for locating. There will always be a residual of movers who it would not be possible to locate before the start of

[2] For example, the Panel Survey of Income Dynamics and the Health and Retirement Study successfully located 97%-98% of sample members who moved between the 2003-5 and 2002-4 waves of these studies and the German Socio-Economic Panel and the British Household Panel had tracking rates of 96% between 2003-5 and 94% between 2003-4 respectively (Couper and Ofstedal, 2009).

fieldwork, either because the move does not take place until fieldwork has commenced or because the move is not discovered until the interviewer attempts to make contact.

4. Results

This paper examines mobility between wave 2 (age 3) and wave 3 (age 5) of the study and uses survey process data to identify whether a family has moved between wave 2 and wave 3, whether they were located if they have moved and if they have moved and were located, whether they were located prior to or after the start of fieldwork. The first part of this section presents these descriptive results and the second part presents results from statistical models which use substantive variables from the survey data collected at wave 2 to predict propensity to move, be located and be located before the start of fieldwork. For this reason, the analytical sample is restricted to families who took part in wave 2 (15,590).

4.1 Descriptive results

Table 4.1-1 shows that 21 per cent of co-operating families at wave 2 (MCS2) moved by wave 3 (MCS3). For a very small number of families (169), it is not possible to know with certainty whether or not they moved. These are a combination of ineligible cases and refusal and sensitive cases which were not issued to the field. For all other cases, it is possible to know with a very high degree of certainty whether or not they moved because, even if they didn't participate in the survey, an interviewer visited their address and established whether or not they were still resident.

Table 4.1-1
Mobility between MCS2 and MCS3 for families who co-operated at MCS2

	Co-operating families at MCS2
Moved between MCS2 and MCS3	3,278 (21%)
Not moved between MCS2 and MCS3	12,143 (78%)
Unknown if moved between MCS2 and MCS3	169 (1%)
Base	15,590

The first column of Table 4.1-2 shows that an extremely high proportion of mobile families (93%) were located. It also shows that mobile families were much less likely than non-mobile families to be located and co-operate at wave 3: 84 per cent compared with 91 per cent.

Table 4.1-2
Location and co-operation at MCS3 for families who co-operated at MCS2 by whether moved since MCS2

	Mobile families i.e. moved since MCS2	Non-mobile families i.e. not moved since MCS2
Located and co-operated	2,766 (84%)	11,036 (91%)
Located and did not co-operate	284 (9%)	1,107 (9%)
Not located	228 (7%)	0 (0%)
Base	3,278	12,143

Table 4.1-3 shows that conditional on location, mobile families were no less likely than non-mobile families to co-operate at wave 3: 91 per cent for both groups.

Table 4.1-3
Co-operation at MCS3 for families who co-operated at MCS2 and were located at MCS3 by whether moved since MCS2

	Located mobile families i.e. moved since MCS2 and located at MCS3	Non-mobile families i.e. not moved since MCS2
Co-operated	2,766 (91%)	11,036 (91%)
Did not co-operate	284 (9%)	1,107 (9%)
Base	3,050	12,143

Overall, 55 per cent of all mobile families were located prior to the start of fieldwork with 38 per cent located during fieldwork (and 7 per cent not located). Table 4.1-4 shows that mobile families who were located prior to the start of fieldwork were just as likely as mobile families who were located during fieldwork to take part in an interview: 90 per cent and 91 per cent respectively.

Table 4.1-4
Co-operation at MCS3 for families who co-operated at MCS2 and were located at MCS3 by whether moved since MCS2 and when located

	Located mobile families i.e. moved since MCS2 and located prior to the start of fieldwork for MCS3	Located mobile families i.e. moved since MCS2 and located during fieldwork for MCS3	Non-mobile families i.e. not moved since MCS2
Co-operated	1,635 (90%)	1,131 (91%)	11,036 (91%)
Did not co-operate	175 (10%)	109 (9%)	1,107 (9%)
Base	1,810	1,240	12,143

Overall, these descriptive results show that the MCS has tracking rates which are high and has tracking procedures which locate a high proportion of mobile families in between waves of fieldwork.

4.2 Statistical modelling

This section reports results from logistic regression models which were used to predict propensity to move, be located and be located before the start of fieldwork. All variables shown in tables were statistically significant (Wald test; $p<0.05$) and 95% confidence intervals are shown for all categories.

4.2.1 Predictors of mobility between wave 2 and wave 3

This section examines how the characteristics of mobile families differ from non-mobile families. A variety of geographic, demographic, socio-economic and attitudinal factors were examined and both unadjusted and adjusted results in the form of odds ratios are presented in Table 4.2.1-1.

Overall, the results did not tell an entirely consistent story. Some of the results indicate that less advantaged families were more likely to move than more advantaged families. The families who were most likely to move were those with younger mothers (under 25), those with another child younger than the (3-year old) cohort child and those who were living in a rented flat which they were dissatisfied with at wave 2. However, other indicators of socio-economic status showed that more advantaged families i.e. those above the poverty line and those with higher numbers of vehicles were more likely to move[3]. Also, although having another child younger than the cohort child was associated with a higher propensity to move, having other children in addition to the cohort child (more than one child in the family) was associated with a lower propensity to move.

The regression model also showed that families in Scotland were slightly more likely to move than families in England and families who were dissatisfied with the area in which they lived were more likely to move than those who were satisfied. Families with mothers in all minority ethnic groups, except mixed, were less likely to move than those with white mothers in

[3] Interestingly, the direction of the relationship between mobility and both of these variables was reversed in the statistical model compared with the univariate analysis.

the unadjusted statistics though only those with black mothers were significantly less likely to do so in the model. Families who had changed from having two parents at wave 1 to one parent at wave 2 were more likely to move than families who had remained as one parent families at both waves, perhaps reflecting a delayed impact of relationship breakdown.

Other variables which were included in the model but did not show a significant relationship with mobility were family type, mother's education and household employment status.

Table 4.2.1-1
Percentage of co-operating families at MCS2 who moved between MCS2 and MCS3 and odds ratios of moving from a logistic regression model, by MCS2 variables

MCS2 Variable	Unadjusted % moved	Odds ratios (OR)	95% Confidence interval for OR
Country			
England	21.0	1	Fixed
Wales	17.5	0.87	(0.73,1.02)
Scotland	24.0	1.16	(1.01,1.35)
Northern Ireland	19.5	1.14	(0.95,1.37)
Age of mother			
16-24	33.3	1	Fixed
25-29	26.2	0.89	(0.75,1.05)
30-34	20.6	0.76	(0.64,0.91)
35-39	15.3	0.58	(0.46,0.72)
40+	14.1	0.51	(0.40,0.68)
Ethnic group of mother			
White	21.0	1	Fixed
Mixed	27.9	1.14	(0.68,1.91)
Indian	16.6	0.82	(0.57,1.16)
Pakistani and Bangladeshi	18.6	0.80	(0.63,1.02)
Black or Black British	19.7	0.62	(0.46,0.83)
Other	21.7	0.93	(0.55,1.57)
Number of children in household (including cohort member)			
One	24.2	1	Fixed
Two	20.5	0.85	(0.74,0.99)
Three	18.7	0.79	(0.66,0.95)
Four or more	18.9	0.71	(0.56,0.91)
Whether cohort member has younger siblings			
Younger siblings	24	1.34	(1.17,1.53)
No younger siblings	19.8	1	Fixed
Family change since MCS1			
Same two parent family	18.4	0.29	(0.08,1.11)
Two parent to one parent	34.9	1	Fixed
One parent to two parent	28.9	0.28	(0.07,1.06)
Same one parent family	28.1	0.67	(0.52,0.86)
Other	39.7	0.45	(0.12,1.71)
Family poverty			
Unknown	22	1.17	(0.97,1.42)
Above 60% median	19.6	1.29	(1.10,1.51)
Below 60% median	24.6	1	Fixed
Tenure			
Own	16.3	1	Fixed
Rent	29.9	1.49	(1.23,1.79)
Other	38.9	2.53	(1.92,3.35)
Accommodation type			
House	19.1	1	Fixed
Flat	39.6	1.87	(1.58,2.22)
Car ownership			

None	28.1	1	Fixed
One	22.2	1.24	(1.06,1.46)
Two	17.7	1.39	(1.14,1.70)
Three or more	20.5	1.53	(1.12,2.10)
Satisfaction with home			
Very satisfied	14.0	1	Fixed
Fairly satisfied	21.6	1.51	(1.33,1.70)
Neither satisfied or dissatisfied	34.0	2.30	(1.89,2.79)
Fairly dissatisfied	33.0	2.25	(1.78,2.85)
Very dissatisfied	46.7	3.62	(2.82,4.64)
Satisfaction with area			
Very satisfied	17.2	1	Fixed
Fairly satisfied	21.5	1.10	(0.99,1.23)
Neither satisfied or dissatisfied	30.7	1.43	(1.01,1.86)
Fairly dissatisfied	32.0	1.53	(1.23,1.90)
Very dissatisfied	38.4	1.48	(1.16,1.90)

4.2.2 Predictors of being located at wave 3, conditional on mobility

This section examines how the characteristics of mobile families who are located differ from mobile families who are not located. The same set of geographic, demographic, socio-economic and attitudinal factors looked at in the previous section are examined here and results in the form of odds ratios are presented in Table 4.2.2-1. As discussed in section 2, it was hypothesised that tracking procedures may be more effective for certain types of families than others.

Overall, the results give little support to this hypothesis, as the only variables which are significant predictors of being located are mother's ethnicity and accommodation type. Families in which the mother is in any non-white ethnic group, except mixed, are much less likely to be located than those with white mothers and those who were living in a flat are less likely to be successfully located than those who were living in a house. The study's tracking procedures are clearly working less effectively for minority ethnic groups which may be related to language barriers and for those living in flats which may be related lower residential stability in this part of the housing sector and barriers to contact such as entry phones.

Table 4.2.2-1
Odds ratios of being located, conditional on moving, from a logistic regression model, by MCS2 variables

MCS2 Variable	Odds ratios (OR)	95% Confidence interval for OR
Ethnic group of mother		
White	1	Fixed
Mixed	0.41	(0.14,1.24)
Indian	0.19	(0.07,0.51)
Pakistani and Bangladeshi	0.29	(0.13,0.63)
Black or Black British	0.21	(0.10,0.44)
Other	0.08	(0.03,0.19)
Accommodation type		
House	1	Fixed
Flat	0.60	(0.36,0.99)

4.2.3 Predictors of being located before start of fieldwork for wave 3, conditional on mobility and being located

This section examines how the characteristics of mobile families who are located prior to the start of fieldwork for wave 3 differ from mobile families who are located during fieldwork for wave 3. The same set of geographic, demographic, socio-economic and attitudinal factors looked at in the previous sections are examined here and results in the form of odds ratios are presented in Table 4.2.3-1. As discussed in section 2, it was hypothesised that tracking procedures which result in the location of families prior to the start of fieldwork may be more effective for certain types of families than others.

Table 4.2.3-1
Odds ratios being located before the start of fieldwork at MCS3, conditional on moving and being located, from a logistic regression model, by MCS2 variables

MCS2 Variable	Odds ratios (OR)	95% Confidence interval for OR
Age of mother		
16-24	1	Fixed
25-29	1.44	(1.05,1.98)
30-34	1.59	(1.12,2.26)
35-39	2.01	(1.42,3.01)
40+	2.35	(1.39,3.98)
Number of children in household (including cohort member)		
One	1	Fixed
Two	0.82	(0.64,1.05)
Three	0.62	(0.46,0.82)
Four or more	0.72	(0.48,1.09)
Whether cohort member has younger siblings		
Younger siblings	1.39	(1.10,1.76)
No younger siblings	1	Fixed

Overall, the results give little support to this hypothesis, as the only variables which are significant predictors of location prior to the start of fieldwork are mother's age, number of children and whether the cohort child has a younger sibling. Families with older mothers and younger siblings were more likely to be located prior to the start of fieldwork and families with more children were less likely to be located prior to the start of fieldwork.

5. Discussion

This paper, motivated by Couper and Ofstedal (2009), has shown that, in common with many other major longitudinal surveys, the Millennium Cohort Study has highly effective procedures for keeping in touch with mobile families. Over 9 in 10 (93%) families who moved between wave 2 and wave 3 were located with over half (55%) located prior to the start of fieldwork. It also showed that, conditional on successful location, mobile families were no less likely to co-operate than non-mobile families i.e. tracking efforts do lead to interviews. This evidence, along with the finding that mobile families have different characteristics from non-mobile families, provides scientific justification for the resources the study devotes to tracking.

As expected, several demographic and socio-economic characteristics were related to residential mobility. However, very few of these factors were related to the successful location of mobile families, either overall or before the start of fieldwork. This is reassuring as it shows that the study's tracking procedures are not systematically failing to reach certain types of respondents, with the exception of families in minority ethnic groups.

The planned next steps for this research are to use information on the timing of the move from wave 3 (available for responding families only) to estimate what proportion of mobile families who move prior to the start of fieldwork are located by study's tracking procedures prior to the start of fieldwork and to use information about the distance of the move (available for located families only) to examine whether mobile families who are located prior to the start of fieldwork are different from mobile families who are located during fieldwork in relation to the distance that they have moved. We would also like to use survey process data e.g. about how mobile families are located to try to evaluate the relative effectiveness and efficiency of different tracking procedures.

References

Couper, M.P. and Ofstedal, M.B. (2009). Keeping in Contact with Mobile Sample Members, in P.Lynn (ed.) *Methodology of Longitudinal Surveys,* Chichester: John Wiley & Sons, Inc, pp. 183-203.

Lepkowski, J.M. and Couper, M.P. (2002). Nonresponse in longitudinal household surveys, in R.M.Groves et al. (eds.) *Survey Nonresponse,* New York: John Wiley & Sons, Inc, pp. 259-272.

Plewis, I. (2007a). Non-response in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology*, 10, pp.325-334.

Plewis, I. (ed.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling (4th. ed.),* London: Institute of Education, University of London.

Plewis, I., Ketende, S.C., Joshi, H. and Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a British Cohort Study: Evidence from the First Two Waves of the Millennium Cohort Study, *Journal of Official Statistics,* 24, pp. 365-385.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Organization and monitoring of the survey area: Impact on estimator quality for a rotating household panel

Thomas Christin, Stéphane Fleury and Johan Pea[1]

Abstract

Where total nonresponse is non-negligible and imperfectly modelled, steps taken to maximize response rates are likely to reduce the variance and the bias of our estimates. In light of this, we propose to evaluate the impact that organizing the survey sphere can have on response rates, more especially focusing on monitoring the survey sphere on a daily basis and anticipating how cooperative households will be. Our work context is the SILC survey in Switzerland.

Key words: Survey sphere, survey monitoring, cooperativeness, nonresponse.

1. Introduction

A major and recurring challenge in statistical surveys is to get a good grasp of the parameters that explain the response process. For longitudinal surveys, the cumulative nature of total nonresponse makes this challenge all the greater. The treatment of nonresponse is primarily oriented toward correcting the nonresponse bias via weighting. We propose to supplement this by introducing measures to maximize response probabilities. Assuming that total nonresponse is non-negligible and imperfectly modelled, it increases the variance of our estimators and creates bias despite the adjustments made by weighting. Usually, the probability of response is explained and then corrected on the basis of the respondent's individual characteristics. While this approach is legitimate from a weighting perspective, it hardly serves to identify the factors that influence the probability of response. Those factors would seem to be primarily within the statistician's realm, and we could act on them. With this in mind, we propose to assess the extent to which the organization and monitoring of the survey sphere can significantly influence response rates.

We will begin by describing the context in which this approach was undertaken, by briefly examining the characteristics of the SILC (Statistics on Income and Living Conditions) survey in Switzerland. We also describe the scope and the characteristics of the total nonresponse process in the SILC survey. In Section 3, we present the principles and hypotheses that governed the development of instruments for monitoring the SILC survey sphere on a daily basis. After describing these instruments more specifically, we evaluate their effectiveness in order to understand the extent to which good monitoring of the survey sphere is likely to increase response probabilities. In Section 4, we propose to assess how information from previous waves can be used to estimate how cooperative households will be. We then examine whether it may be effective to use this estimation to increase the response probabilities for uncooperative households. To conclude, we evaluate the extent to which, by optimizing the relationship between respondents' characteristics and their probability of response, we may be replacing one problem with another. Indeed, what we gain in the quality of the estimators by minimizing total nonresponse may be partly offset by losses in the form of a higher partial nonresponse rate or greater response error among the units concerned.

2. The SILC survey

2.1 A European project

The SILC survey is a European project coordinated by Eurostat under the name EU-SILC. Launched in 2003, EU-SILC collects information on more than 250,000 individuals in more than 25 countries. Its main objective is to annually produce community-wide statistics on income distribution, living conditions, poverty and social exclusion, with cross-sectional and

[1]Thomas Christin, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Thomas.Christin@bfs.admin.ch); Stéphane Fleury, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Stephane.fleury@bfs.admin.ch); Johan Pea, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Johan.Pea@bfs.admin.ch).

longitudinal data that are comparable and updated. The SILC survey was launched in Switzerland in 2007 in the form of a rotating four-year panel.

2.2 Characteristics of the SILC sphere in Switzerland

In Switzerland, we are working with a proportional stratified sample for each major region, consisting of approximately 6,000 households. Thus, more than 11,000 persons are interviewed by telephone each year for SILC, in wave 1, 2, 3 or 4. The use of administrative data for some income sources is in the testing and development stage. A feature of the SILC survey is that it interviews all persons aged 16 and over in eligible households.

Addresses in the raw sample are divided into four packages, activated successively in the survey so as to minimize the time interval between receipt of the notification letter and the first telephone contact. A private polling institute is commissioned by the Office fédéral de la Statistique (OFS) to conduct the interviews. In accordance with Switzerland's multilingual makeup, the interviews are conducted in French, German or Italian out of two call centres located in the two largest linguistic regions, one in German-speaking Switzerland (Berne), which manages interviews in German, and the other in French-speaking Switzerland (Lausanne) for French and Italian. This situation raises a question as to the unity of our survey sphere, especially since the two call centres are considerably different in size and survey capability. More than 70% of the addresses are processed by the Berne call centre (German), which has a larger number of interviewers and more modern infrastructure than the Lausanne centre. There are also differences with respect to work culture and recruitment procedures between these two call centres.

The SILC survey in Switzerland is structured around several questionnaires (household composition, referred to as the matrix questionnaire; household questionnaire; individual questionnaire; child/proxy questionnaire) with a wide range of themes examined. The interview burden for SILC in Switzerland is considered sizable. In Wave 1, interviews last approximately 40 minutes for a one-person household and almost 1 hour 30 minutes for a household composed of three eligible persons (more than 16 years of age).

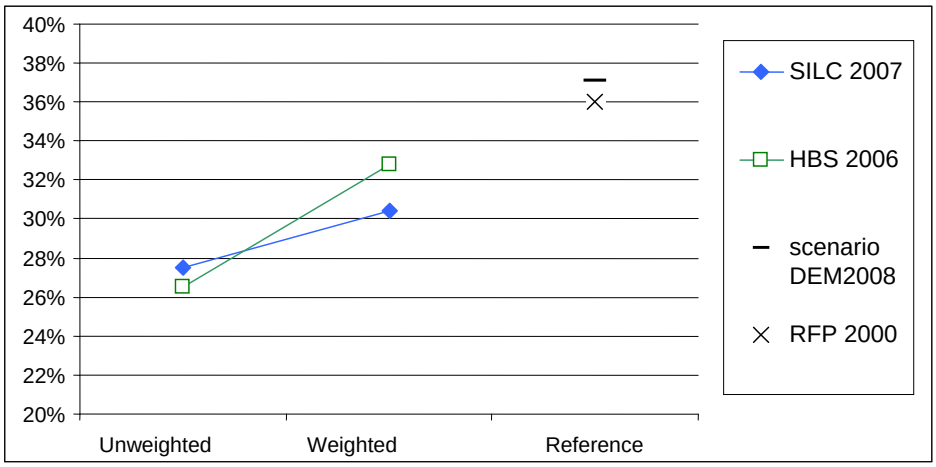
2.3 Total nonresponse for the SILC in Switzerland

Despite a sizable response burden, the response rate in Wave 1 of SILC in Switzerland in 2007 was 68%,[2] which is considered a good result in the Swiss context. In a Europe-wide comparison, Switzerland is in the middle of the distribution of response rates for EU-SILC. The nonresponse process is modelled by segmentation to adjust the weights. The main characteristics lessening the probabilities of response in Wave 1 are a low education level, no member of the household with Swiss nationality, and living in a one-person household. Since the main indicators estimated by SILC have to do with income aspects, it seems clear that the response probabilities are related to our variables of interest, thus leading to non-negligible nonresponse.

In addition to nonresponse being non-negligible, our working hypothesis is that the probability of responding is imperfectly modelled. Because of the complexity of the response process, some of the auxiliary variables explaining total nonresponse are unknown. Moreover, there is a high probability that a portion of the response process is not related to structural variables, but rather to subjective or hard-to-observe factors such as the respondent's mood at the time of the telephone call or the nature of an event preceding the call. We will therefore assume that nonresponse bias cannot be entirely corrected by weighting. Supporting this hypothesis, Figure 2.3-1 shows for SILC and the household budget survey (HBS) the estimates of the percentage of one-person households in Switzerland before and after weighting.

[2] This response rate is calculated as the ratio of the number of usable (valid) households to the total number of eligible households. Addresses not contacted for which the eligibility status is unknown are considered to be eligible.

Figure 2.3-1
Differing estimates of the percentage of one-person households in Switzerland



Sources:
OFS, SILC 2007, HBS 2006, demographic scenarios for 2008, Recensement Fédéral de la Population (RFP) 2000.

This figure shows that the weighted estimates of a parameter—even a fairly basic one such as the percentage of one-person households—are significantly different depending on the data source used. Also, we know that the risk of poverty, which is the main dimension of interest in SILC, depends among other things on the size of the household.

These findings encourage us to do our utmost to optimize, insofar as possible, the relationship between individuals' cooperativeness toward statistics and their probability of response to the SILC survey. Individuals' cooperativeness toward statistics is considered to be relatively stable in the short term. It depends on their life history and their socioeconomic characteristics. While the relationship between cooperativeness and the probability of response is generally considered to be strong, it is nevertheless variable. For example, it may be influenced by the organization of the survey sphere. It was precisely this consideration that led us to invest in monitoring the survey sphere; the objective was to minimize total nonresponse attributable to organizational defects.

3. Monitoring of the survey sphere and implications for total nonresponse

3.1 Objectives and basic principles

The initial observation that led us to set up monitoring of the SILC sphere in Switzerland is that the survey sphere may be seen as a large, complex machine for producing interviews. Like any sophisticated production machinery, our survey sphere is likely to have, unseen to us, faulty settings which reduce its performance level and which in this case are measured by an increase in nonresponse. Our survey sphere thus offers a potential for optimizing the settings, e.g., in terms of striking a balance between the amount of inputs (households to interview) and the amount of manpower needed (i.e., interviewers), or in terms of timing the activation of address packages. Thus, the effort to optimize the settings by appropriate and timely oiling of the machinery is designed to reduce total nonresponse due to defective management of our survey sphere.

The objective of survey monitoring is to enable us to obtain a daily picture of how the survey is progressing and to determine both the right time to activate address packages and the optimal length of the survey. Monitoring must also serve to quickly identify performance disparities from one call centre to the other or from one address package to the other and must make it possible to react quickly by alerting the call centres of the polling institute. It also serves not only to show problems concretely but also to justify our actions to the institute when, for example, we request an increase in the number of interviewers in a call centre. By displaying the key stages of the survey sphere, it provides an important database for improving its management from one year to another. By refining our picture monitors, it enables us to empirically determine the performance objectives required of the polling institute commissioned for the survey.

The monitoring tools are developed on the basis of daily statistics supplied by the call centres. These statistics cover the processing status of addresses activated in the survey at the household and individual levels, regarding initial contacts, eligibility status, the making of appointments, call-backs in the event of refusal and the conducting of interviews. This information is delivered in Excel format with subtotals by contact and response status; it is delivered in SAS format at the

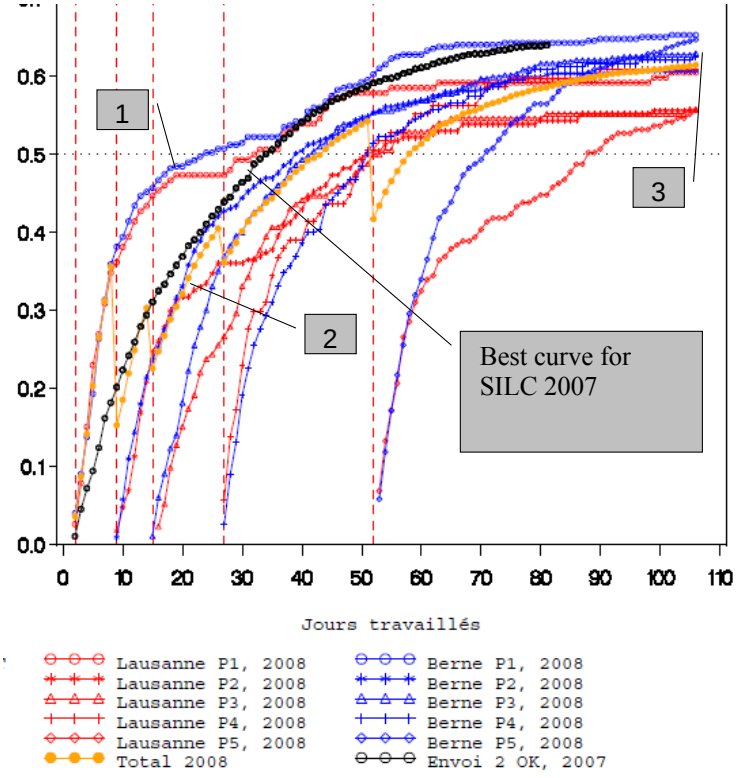
household and individual level, with contact and response status for each unit. This constitutes a sizable mass of data, from which it is not easy to extract the most relevant information. The challenge therefore becomes one of developing an instrument panel for piloting the machinery as effectively as possible without requiring excessive analytical resources on our part. For this, we work with SAS macros that daily produce charts showing the critical points of progress in the survey sphere. In these charts, we look for signs that the survey machinery is out of adjustment, such as performance differences between call centres or between address activation packages.

3.2 Information drawn from SILC monitoring for 2008 and 2009

In this section, we present a few examples of information drawn from the monitoring of the 2008 and 2009 SILC surveys in Switzerland. Chart 3.2.1 shows the daily rates for valid households[3] by activation package and by call centre for the cohort of households introduced into the SILC panel in 2008. The 2008 raw sample for Wave 1 was activated in four successive packages. Because SILC began in 2007 in Switzerland, the 2008 survey was the first to combine households interviewed in Wave 1 and Wave 2. A backup package had to be activated because the 2008 response rate was slightly lower than the one estimated on the basis of the response rate recorded in 2007.

Figure 3.2-1 shows an excellent advance in the first package, activated in the first two weeks of 2008, with a sizable increase in the number of households interviewed. Point 1 identifies the problem of a slackening when packages 2 and 3 were activated. This indicates that the call centres had difficulty simultaneously managing the addresses yet to be processed in the old packages and the new addresses activated. This problem of a "jamming of the survey machinery" was greater in the small Lausanne call centre (red curve). The same problem may be seen at point 2, which marks a substantial slowing in progress on the second package when the third was activated, but this time only at the Lausanne call centre. At the end, a systematic disparity in performance between the Berne and Lausanne call centres is observed, with the rates of households interviewed being lower in Lausanne in Wave 1 in 2008 (point 3). It may also be seen that the late activation of the backup package had no consequence for the Berne call centre, but caused a major performance shortcoming in Lausanne.

Figure 3.2-1
Daily increase in complete households per activation package, Wave 1, SILC 2008



Visualization of the problems encountered in 2008 (Figure 3.2-1 and discussions with the institute responsible for the interviews in 2008 led to correction of those problems in 2009. As may be seen in Figure 3.2-2, slowdowns in the progress on

[3] For SILC, a household is considered valid if the matrix and household questionnaires and at least one individual questionnaire were answered.

the old address packages when the new ones were activated were less pronounced, and above all the differences between the Lausanne and Berne call centres were partly erased. On the other hand, the institute responsible for the interviews performed much better on the first activation package than on the following ones. Since addresses were randomly distributed between the activation packages, these disparities are primarily due to problems in the organization of the survey sphere.

Figure 3.2-3 provides an example of good progress in a survey sphere with no difference in performance based on the call centre or the activation package.

Figure 3.2-2

Daily progress of completed matrixes for each activation package, Wave 1, SILC 2009

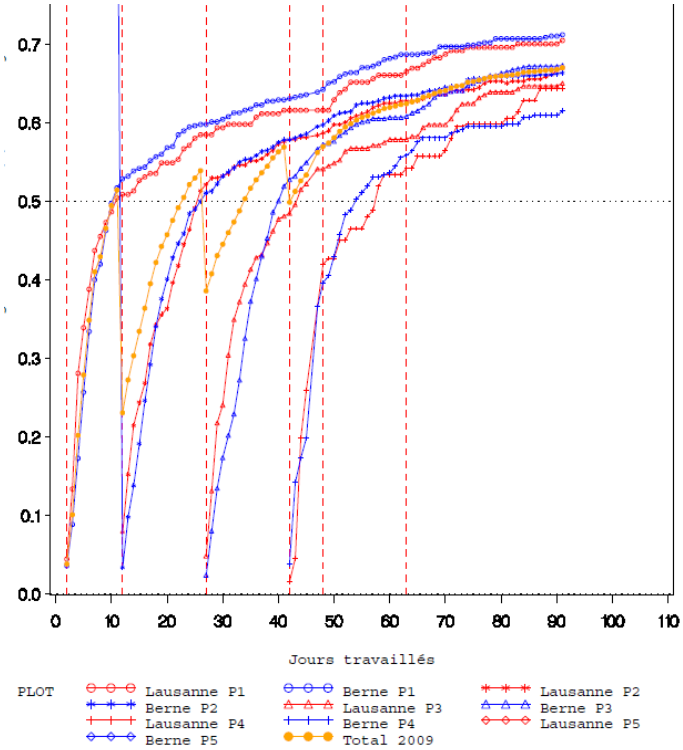
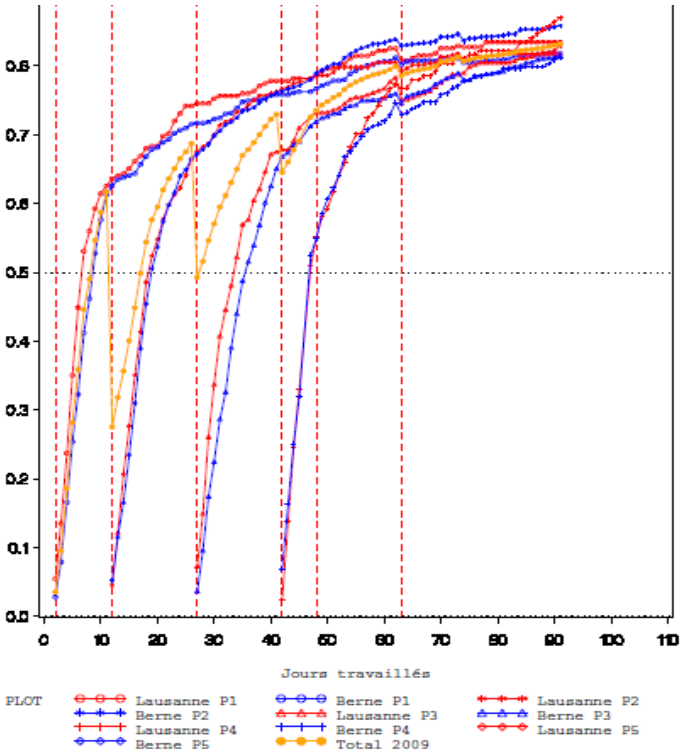


Figure 3.2-3
Daily progress of matrixes completed for each activation package, Wave 3, SILC 2009



The first information drawn from this exercise of monitoring the survey sphere is as follows:

- It seems essential for an address package to be properly launched. When a new address package is processed too slowly, it is quite difficult to catch up and its final response rate suffers. We explain this by the fact that it is preferable for interviews to be conducted soon after the subject receives the official notification of the survey, sent by mail to the households selected, and that subsequently it is difficult for the institute to manage simultaneously the addresses not yet processed in older packages and the fresh addresses in the new packages.
- The size of the address packages and the speed of activation are important challenges to be met in order to avoid the risk of overload at the call centres and enable the latter to draw as much as possible from the raw samples and thus maximize response rates.
- It is more difficult in small call centres to achieve an optimum balance between, on the one hand, the resources invested, and on the other hand, the number of addresses to be contacted, the interviews to be conducted and the refusal call-backs to be made.
- In a rotating panel, optimization of the survey sphere seems more difficult for households in the first wave. In subsequent waves, many of the least cooperative households have been lost in earlier waves. While performance in the first wave is especially important, it is the most difficult to optimize. This is a major challenge, since total nonresponse in Wave 1 owing to defects in the survey sphere adversely affects the statistical quality of the cohort throughout its existence in the panel.
- In attempting to explain performance differences between the call centres, it is important not to neglect the human factor as regards either the call centre managers and supervisors or the interviewers.

3.3 Anticipating how cooperative a household will be toward statistics

There are various ways to maximize response probabilities in the first wave of a rotating panel, including optimizing the survey sphere, offering compensation to respondents or targeting the mailings to the selected households. If, as a result of these different strategies, these uncooperative households participate in the survey, this means that the actions taken have helped to obtain their participation. For a longitudinal survey, this optimization in the first wave increases the risk of attrition in subsequent waves as households uncooperative toward statistics are included in the longitudinal phase of the survey. To minimize this problem, we sought to estimate how cooperative households were, and we tried to put in place special procedures for households that seemed to pose especially high risks of refusal in the next wave.

There are several possible options for estimating this cooperativeness starting in the second wave:

1. Use the cumulative number of calls needed in the previous wave to obtain interviews with the household. A high number of calls may be synonymous with a disguised refusal, but it may also indicate limited availability of the household or poor call management on the part of the call centre. Since this information was not sufficiently unambiguous, it was not used.
2. Use the interviewers' subjective evaluations as to the interviewee's cooperativeness.
3. Use the percentage of partial nonresponse in the previous wave. Households with the highest percentages of partial nonresponse would then be considered to be uncooperative toward statistics and hence more likely to refuse to respond in the following wave. Using the partial nonresponse criterion, we identified two types of uncooperative households. The first consisted of valid households in which more than 50% of the persons who were eligible for the individual questionnaire refused to respond to it (criterion of unit or total nonresponse). The second type consisted of households with a high level of partial nonresponse.

Comparing options 2 and 3, we found that there was little variation in interviewers' subjective evaluations. Interviewers are generally overly positive in their subjective assessment of the cooperativeness of the individual whom they have just interviewed. We therefore chose the third option for estimating cooperativeness. We did not prioritize the variables according to their importance or their degree of difficulty or intimacy. The total number of partial nonresponses on the household questionnaire and the individual questionnaires was examined in relation to the total number of questions asked in these questionnaires. This total number varies depending on the household's situation, more concretely according to the filters that control the logical sequencing of the questions.

The households that accounted for more than 50% of total or unit nonresponse and those that accounted for the largest percentages of partial nonresponse in 2007 were transmitted to the call centres marked with a red flag for the next wave. Those households, assumed to be uncooperative, were managed separately by a team of specialized interviewers, starting with the first attempts at contact in 2008. If during the first contact these households refused to participate, no refusal call-back was organized.

Table 3.3-1
Participation rate of households in 2008 according to estimated degree of participation in 2007

Estimated cooperation level in 2007 based on partial NR (W1)	Participation rate in 2008 (W2)
Cooperative households	84%
Uncooperative households/ partial NR	74%
Uncooperative households/total NR	56%

These results show that the cooperation variable constructed on the basis of partial nonresponse in Wave T-1 is effective for estimating the risks of refusal in Wave T. To our surprise, the sizable presence of persons who refused the individual questionnaire in valid households increased longitudinal attrition more than the strong presence of partial nonresponse. Contamination of cooperative members of the household by uncooperative members appears to be a major factor in explaining longitudinal attrition. However, with no control group with similar management of cooperative and uncooperative households, it is very hard to judge the performance of the actions taken in 2008 for non-cooperators.

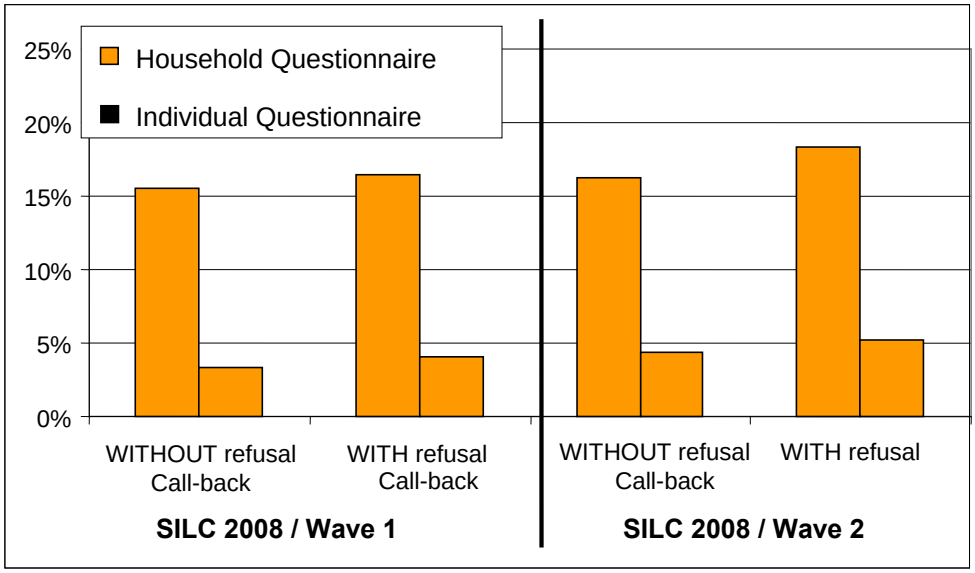
3.4 Cost of minimizing total nonresponse in terms of partial nonresponse

Investments made to improve the organization of the survey sphere via survey monitoring serve to reduce total nonresponse. In our view, this may be explained by optimization of the relationship between respondents' characteristics and their probability of response. In a context where total nonresponse is non-negligible and imperfectly modelled, any reduction in total nonresponse translates into a drop in the variance and the bias of our estimators. Nevertheless, this gain could be partly offset if households that agree to participate owing to these actions were to accumulate more partial nonresponse than the others, thus increasing the number of values to be imputed. If we were to evaluate this issue using the level of cooperativeness based on the percentage of partial nonresponse in the previous wave, this would not have been sufficient and it would have in part been tautological. It would have instead served to test whether the process of partial nonresponse remains stable from one wave to another. We therefore propose to use the presence of refusal call-backs as an indicator of non-cooperation to evaluate this issue. Figure 3.3-1 shows, for valid households, the average rate of unanswered questions in relation to the number of questions asked, depending on whether a refusal call-back was made. This figure shows that on average, partial nonresponse is slightly

larger in cases where the interview was conducted following a refusal call-back, especially for the household questionnaire. However, the difference is not sizable.

Figure 3.3-1

Percentage of partial nonresponse depending on whether or not a refusal call-back was made, SILC 2009



4. Conclusion

The machinery for producing interviews is difficult to optimize. Possible faulty settings in the machinery are many, and they are likely to introduce new structural dimensions to total nonresponse that are unrelated to the respondents' characteristics. Daily monitoring of the survey sphere, by call centre and by address activation package, is useful both for visualizing problems and for trying to change the organization of the survey sphere to maximize its overall performance.

For longitudinal surveys, it is useful to estimate households' cooperativeness using the partial nonresponse rate observed in the previous cycle. Identifying less cooperative households is highly appreciated by interviewers, who can better prepare themselves before contacting these households in the next wave. However, in the absence of a test group, it is difficult to assess the impact of the actions taken to maximize the response rates of households assumed to be uncooperative. It is also possible that if these households are managed separately by specialized interviewers, this will adversely affect the participation rate. If interviewers know in advance that a household has a high risk of refusing, they may be overly influenced by the idea of encountering a refusal and thus may unconsciously guide the household toward such a refusal. Other ways to minimize the high attrition for such households are currently being evaluated, starting with no longer identifying them or managing them separately in the call centres. Focusing solely on the least cooperative households may not be sufficient. It is possible that actions targeting households that are only somewhat cooperative may be more likely to influence the probability of response.

While longitudinal attrition is influenced by the sizable presence of partial nonresponse in the previous wave, the effect of contamination of the cooperative members of a household by uncooperative members is a much more important factor in explaining longitudinal attrition.

Since the hypothesis that total nonresponse is non-negligible and imperfectly modelled is in all likelihood a valid one, efforts to optimize the organization of the survey sphere are legitimate, especially since minimizing total nonresponse does not significantly increase the problem of partial nonresponse and imputations. Thus, reducing bias and variance by minimizing total nonresponse does not exact a price in terms of imputations.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Responsive Design for the Survey of Labour and Income Dynamics

Tracy Tabuchi, François Laflamme, Owen Phillips, Milana Karaganis, and Amélie Villeneuve[1]

Abstract

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey used to measure changes in the economic wellbeing of Canadians and the factors that may influence these changes. Interviews are conducted by means of a computer-assisted telephone interview (CATI). A number of initiatives have been put in place over the years to better manage collection resources and effort. Despite efforts to better manage collection, SLID has seen a consistent drop in response rates over recent years.

The Households and the Environment Survey (HES) piloted a Responsive Design (RD) strategy for the 2009 collection, wherein the collection strategy was modified based on analysis of the in-coming paradata. SLID will implement a RD strategy for the 2010 collection. This paper describes how SLID can implement its RD strategy within the framework of what has already been developed.

Key Words: Responsive Design, Longitudinal Survey, Paradata, Propensity Model.

1. Introduction

1.1 SLID overview

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey. The data are used to measure changes in the economic wellbeing of Canadians and determine factors that may influence these changes. To this end, SLID collects information relating to employment and family status, jobless spells and wages and income.

Interviews are conducted annually by means of a computer-assisted telephone interview (CATI). During each wave of collection, data are collected for two longitudinal panels simultaneously. To determine who is selected to become a longitudinal panel member of SLID, first a subsample of approximately 17 000 responding Labour Force Survey (LFS) households is selected. Then, all individuals that are members of those households become members of a SLID longitudinal panel. Members of a longitudinal panel are followed for six waves of collection. They are interviewed even if they are no longer members of the originally selected LFS household. Panels have staggered start dates; every three years a new panel is selected. When this new panel is introduced, the older panel is rotated out.

1.2 Survey response and collection effort

The response rates for SLID have been steadily declining since the first collection in 1994. Not only is there a downward trend in response rates for a given panel over its lifetime, there is a 'new panel' effect. The response rate for the first wave of collection for a panel is lower than that of the previous panel (see Table 1.2-1).

[1]Tracy Tabuchi, François Laflamme, Owen Phillips, Milana Karaganis, and Amélie Villeneuve, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6 (email contact: tracy.tabuchi@statcan.gc.ca)

Table 1.2-1
Longitudinal response rates by panel and wave of collection

	Wave of collection					
Panel (year panel began)	1	2	3	4	5	6
Panel 1 (1994)	93.3	89.6	86.5	83.9	82.6	81.5
Panel 2 (1997)	89.5	86.8	85.2	82.7	78.5	77.4
Panel 3 (2000)	83.9	83.0	83.0	79.6	76.4	73.7
Panel 4 (2003)	81.2	83.2	78.3	75.0	71.6	68.9
Panel 5 (2006)	78.8	80.6	77.3	72.8	…	…
Panel 6 (2009)	71.0	...	...	...	...	...

... not applicable

Statistics Canada's longitudinal surveys have seen increases in the effort required to resolve cases. For SLID, a case is considered resolved if the interview is completed or partially completed, if there is a hard
refusal or if the household is identified as out-of-scope for the wave of collection. The increase in effort required corresponds to an increase in the average time spent per case. Dufour (2009) found that this is attributable to increases in three areas: the number of cases where there was no contact with a household; the number of cases requiring refusal conversion; and the number of call attempts required to resolve a case. Since survey budgets are limited and there has been an increase in the amount of effort required to obtain targeted response rates, it is clear that ways of collecting data in a more efficient manner, while maintaining data quality, must be found.

Currently, for SLID and other surveys at Statistics Canada, the collection strategy is set out prior to collection. Each data collection office agrees to provide a certain level of collection effort, in terms of interviewer hours and ancillary expenses. Then, they are tasked with obtaining targeted response rates by apportioning this effort as efficiently as possible. Groves and Heeringa (2006) showed that it is possible to reduce the per unit costs associated with collection for computer-assisted personal interviews by using an adaptive approach to data collection. An adaptive approach uses paradata, that is, data related to the survey process, to modify the collection strategy. The impact of such an approach is greatest during the later stages of collection where more effort is required to finalize cases.

Mohl and Laflamme (2007) proposed several options for incorporating adaptive approaches to data collection at Statistics Canada. These were incorporated into the Responsive Design (RD) strategy that was piloted with the 2009 data collection of another CATI survey, the Households and the Environment Survey (HES) (Laflamme and Karaganis, 2010). The SLID RD strategy is based heavily on what was done with HES.

In this paper, Section 2 summarizes the current collection management tools that are in place for SLID. Then, Section 3 gives an overview of the proposed RD strategy. Finally, Section 4 explains how the RD strategy will work for SLID and what has been done so far to prepare for its implementation during the 2010 data collection.

2. Current collection management tools

2.1 Cap on calls and time slices

Several practices and initiatives have been put in place to better manage resources in the data collection offices. At Statistics Canada, there is a cap on the maximum number of call attempts allowed for sampled units. For SLID and other longitudinal surveys, this "cap on calls" is 40 call attempts. Each time a phone number is called, regardless of the outcome, it is counted as a call attempt. For example, if a house is called and the interviewer gets a busy signal, this is counted a call attempt. However, tracing attempts, that is call attempts that are made in order to trace a household, are not counted towards this cap on calls. This practice helps to manage interviewing effort more efficiently and ensures that respondents are not inundated with calls from the data collection offices.

Time slices and time slice groups are used to optimize the chances of contacting household members. Time slices partition the week into blocks of time, and call attempts are distributed among these blocks based on household demographic characteristics, which define the time slice groups. There are four time slice groups used for SLID: (1) households composed of elderly respondents; (2) households composed only of young single respondents; (3) households that were unresolved or refusals at the end of the last collection; and (4) all other households. The targeted distribution of the 40 call attempts for each time slice group is presented in Table 2.1-1. If a call attempt is a non-contact, the time slices will affect the next call attempt.

Table 2.1-1
Distribution of call attempts by time slice group

	Monday – Friday				Saturday		Sunday	
Time slice group	Start - 12:00	12:00 - 16:00	16:00 - 19:00	19:00 - End	Start - 12:00	12:00 - End	Start - 16:00	16:00 - End
Elderly	14	6	12	2	2		2	2
Young singles	5	2	10	13	2	3	2	3
Tracing / refusals	-	-	-	-	-	-	-	-
Others	4	5	13	11	1	2	2	2

2.2 Priority Groups

The CATI application allows survey managers to identify Priority Groups within the in-progress cases, in order to ensure that adequate response rates are obtained for pre-identified domains of interest. If the response rate for a particular domain is lower than expected, effort can be concentrated on households within that domain. Priority Groups are used and managed in conjunction with the cap on calls. Domains of interest for SLID include the panel, the time slice group and the province of residence.

2.3 Call scheduler

A call scheduler automates the way cases are assigned to interviewers based on many paramaters including the interviewer's profile. Factors that determine which cases will be sent to the next available interviewer include: the current status of the case, the number of calls already attempted, the time slice group, the date and time of the last call, time slice distribution status, etc.

3. Responsive Design overview

The proposed Responsive Design (RD) strategy breaks down the survey data collection process into four phases: planning, initial collection, RD phase-in 1 and RD phase-in 2. The first phase (planning) occurs before data collection starts. Prior to the beginning of collection, data collection activities and strategies are planned out, developed and tested for the other three data collection phases. The main activities include the analysis of previous data collection cycle(s) to identify improvement opportunities, frame and sample assessment and validation, the development of active management tools and the establishment of staffing plans.

The second phase (initial collection) includes the first portion of the data collection process; from the collection start date up until it is determined that RD phase-in 1 needs to be initiated. During the second phase, new features can be introduced into the collection process (for example, an intermediate cap on calls to avoid cases capping out before the last data collection phases). During this initial collection phase, many key indicators of quality, productivity, cost and responding potential of in-progress cases are closely monitored.

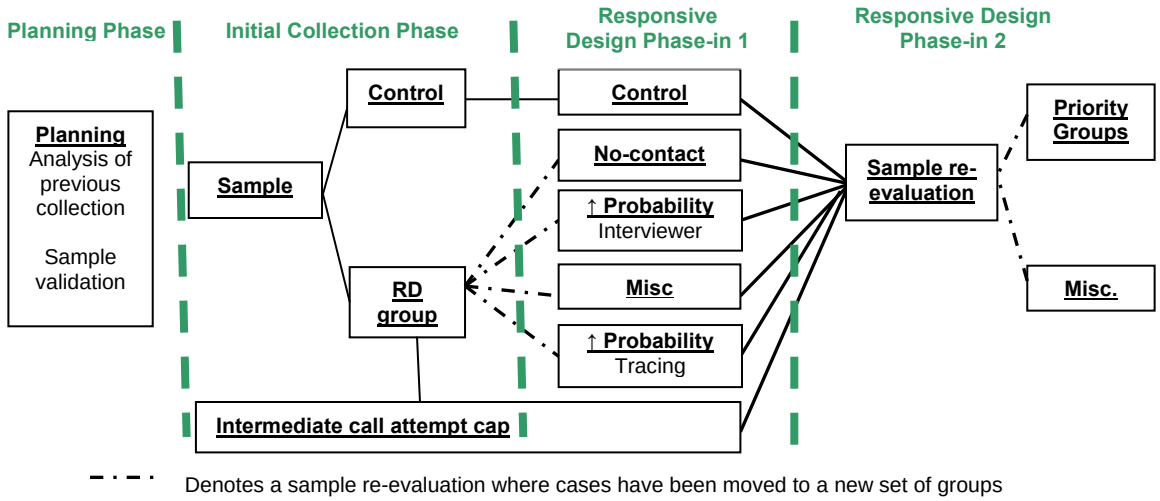
The third phase (RD phase-in 1) categorizes and prioritizes in-progress cases using information available prior to the beginning of collection and paradata, accumulated during collection with the objective of improving overall response rates. In particular, a propensity model (logistic regression) is used to evaluate each unit's likelihood of being interviewed and to categorize and prioritize each in-progress case. During this phase, key indicators continue to be monitored. Schouten, Cobben and Bethlehem (2009) identified the need for quality indicators that monitor not only response rates, but also sample representativity. Called R-indicators, they will be used to provide information on the variability of response rates between domains of interest to determine when the last phase should begin.

The last phase (RD phase-in 2) aims to reduce the variance of response rates between the domains of interest by targeting cases that belongs to the domains with lower response rates, thus improving the sample representativity.

4. Responsive Design for SLID

Figure 4-1 presents a summary of the RD strategy for SLID. Further details about the RD phases are provided in the following subsections.

Figure 4-1
Responsive Design phases for SLID



In order to allow for an evaluation of the effectiveness of the RD strategy, the sample is split into a control group and an RD group prior to the start of collection. The control group and RD group will be roughly the same size. To ensure representativity within each of these groups, a systematic sample will be drawn after the sample is stratified by panel, data collection office, time slice group, language of interview, previous wave case outcome, province of residence and total household income.

4.1 Planning phase

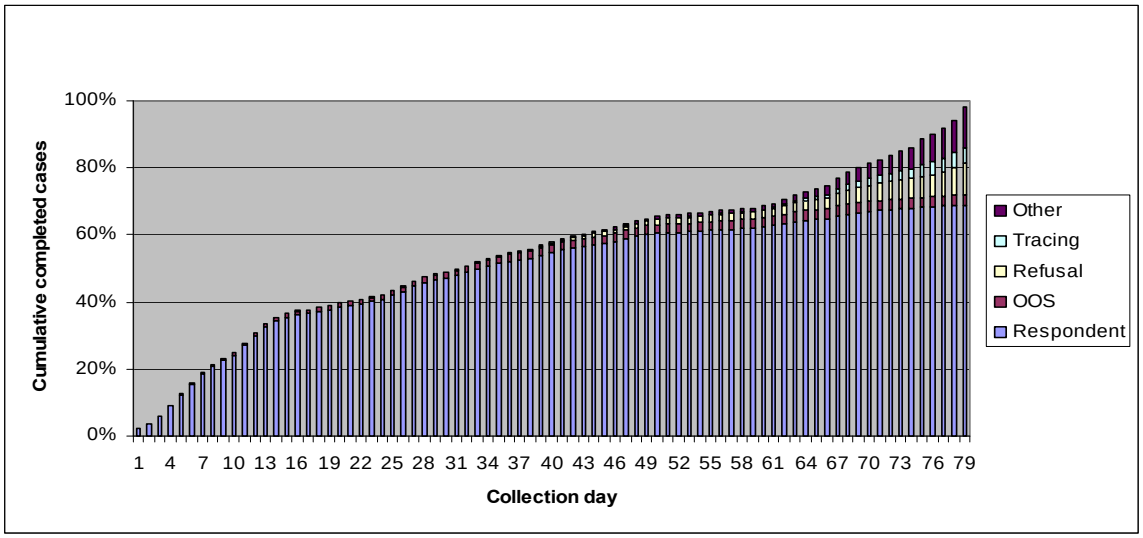
The data collection for the fifth wave of Panel 5 and the second wave of Panel 6 is scheduled to occur from January to March 2010. The first step in implementing the RD strategy for this collection is to examine the 2009 collection paradata. By understanding the way that collection progressed previously, the parameters for the RD strategy can be set.

The paradata comes from the Blaise Transaction History (BTH) file. This file contains a record for each time a case is accessed through the CATI application. Examples of paradata variables available on the BTH file include the date, start time and end time of call attempts and the result of the call attempt, recorded as a numeric outcome code.

4.1.1 Implementation date

By examining daily completion rates from the 2009 collection, (see Figure 4.1.1-1), there are two places where these rates plateau – collection days 17 through 23 and 52 through 58. These plateaus correspond to weeks where the data collection offices are conducting the January and February data collection for the LFS. The LFS is an important and large survey; often data collection for non-LFS surveys is minimal or suspended during the first three or four collection days of LFS. Considering the reduction in time that will be spent conducting SLID interviews, the February LFS week is an appropriate time to analyse the collection paradata, calculate the response probabilities, partition the in-progress sample into groups and begin the RD phase-in 1. The February LFS week for 2009 began on the 22nd.

Figure 4.1.1-1
SLID 2009 - Cumulative completed cases by collection day



4.1.2 Intermediate call attempt cap

In order to ensure that cases from the RD group will not be capped too early, an intermediate call attempt cap will be introduced. Any cases that reach this intermediate cap will be set aside until the start of the RD phase-in 1. An important consideration for determining the intermediate cap is how removing cases from the regular caseload will impact operations in the data collection offices. If too many cases are inaccessible during the initial collection phase, operations in the data collection offices could be disrupted. However, if not enough call attempts are reserved for the RD phases, it could be difficult to see the benefits of the RD strategy. There must be a balance between these two considerations.

Given that the cap on calls for SLID is 40, intermediate caps of 25 and 30 were examined. Table 4.1.2-1 presents some statistics about how intermediate caps would have impacted the 2009 collection, had they been in place. For example, almost 20% of the call attempts made between the date of the first capped case and the start of the February LFS corresponded to cases that reached 25 call attempts.

Previously, SLID defined high effort cases as those requiring 25 or more call attempts. Initially it was felt that setting the intermediate cap to 25 to correspond with the definition of a high effort case would be appropriate. However, seeing the high number of cases that would be removed from the regular caseload, it was felt this would be disruptive for the data collection offices. Considering this, the intermediate cap was set at 30 call attempts.

Table 4.1.2-1
SLID 2009 - Statistics comparing intermediate caps of 25 and 30

	Cap of 25	Cap of 30
Collection day of 1st case reaching the intermediate cap	day 24	day 26
Number of cases reaching cap by the start of February LFS	3 443	1 681
Percentage of total call attempts for capped cases*	19.6%	11.3%
Percentage of total time spent on capped cases*	14.0%	7.9%
Number of resolved capped cases*	256	144

* between the date of the first capped case and the start of LFS week, February 22, 2009.

4.1.3 Propensity model

A propensity model will be used to evaluate a household's likelihood of being interviewed. Using a logistic model, the cases with the highest predicted probabilities will be given priority during the RD phase-in 1. Several models were compared in order to choose one best suited for use during this phase.

As with most surveys, SLID is interested in the response status of households selected in its sample. So naturally, one of the response variables considered was whether or not the household was a responding household as of the February LFS cut-off

date. SLID is also concerned with whether or not the household is resolved during the collection period. So another response variable considered was the resolved status of the household as of the LFS cutoff date.

The explanatory variables considered in the models came from three different sources. For both panels the 2009 BTH data and the demographic information from the sample file were available. Since 2009 was the first wave of collection for Panel 6, only Panel 5 had BTH data from the 2008 collection. From the collection paradata, categorical variables were created from discrete variables (e.g. number of call attempts, number of contacts) and indicator variables were created. A full list of variables that were considered can be found in Table 4.1.3-1.

Table 4.1.3-1
Explanatory variables considered for the propensity models

Data source	List of variables
2009 Sample file	Province of residence, language of previous interview, split household indicator, time slice group
2009 BTH* - basic	Number of call attempts, date of the last attempt, number of tracing attempts, refusal indicator, tracing indicator, number of contacts
2009 BTH* - detailed	Number of appointments, number of all attempts (regular and tracing combined), contact during a time slice, indicator variables: weekend attempt, weekday attempt, sent to senior interviewer, outcome indicators: wrong number, absent for duration of survey, language issues, interview interrupted, special circumstances, call after first refusal, call after first tracing attempt, call after tracing lead, call after first sent to senior interviewer
2008 BTH**	Number of call attempts, date of the last attempt, number of tracing attempts, refusal indicator, tracing indicator, number of contacts, response status, case outcome

* Variables derived as of LFS cut-off date
** Variables derived at the end of 2008 collection, only available for Panel 5

To simulate the conditions wherein the RD phase-in 1 would be initiated, constraints on the 2009 collection paradata were imposed. First, to simulate the intermediate call attempt cap, only BTH records with 30 or fewer call attempts were kept. Next, since the model would be run during the February LFS week, only BTH records for calls made before the LFS cut-off date, February 22, 2009, were kept.

The response rate and resolved rate, as of the LFS cut-off, were panel dependent, so logistic models were created separately for each panel. Each model was built using either the response status or the resolved status as the response variable. For the explanatory variables, all the models used the sample file variables and the basic 2009 BTH variables, while only some models used the detailed 2009 BTH variables. Some of the logistic models for Panel 5 also included the 2008 BTH variables in addition to the sample file and 2009 basic BTH variables.

Each model was built in SAS using a stepwise forward selection process, with a maximum of 20 explanatory variables entering the model. The predicted probabilities were calculated for all units in the sample, both in-progress and finalized cases. To assess the model's performance, only the in-progress cases were kept and grouped into deciles based on their predicted probabilities.

To ascertain how well the propensity model was able to identify cases most likely to be completed, two criteria were considered: the response rates for cases having predicted probabilities in the top decile group; and the spread of response rates between the top and bottom decile groups. Comparing the characteristics of the resolved status versus the response status models for the highest decile groups, the response rates were consistently higher than the resolved rate. And the range in the mean response rates between the top decile and the bottom decile groups was larger than the corresponding range in mean resolved rates. This showed that using the response status was better able to identify cases that would be completed by the end of collection. Therefore, only models using the response status were considered further.

For Panel 5, the simplest model, that included only the sample file and 2009 basic BTH variables, performed the best. This was measured by comparing the response rates at the end of collection for the top decile of in-progress cases between the models. For Panel 6, the more complex model more accurately predicted which in-progress cases were most likely to be respondents by the end of collection. Table 4.1.3-2 summarizes the variables that entered into each model. Notably, the SLID propensity models use different variables than the HES propensity model.

Table 4.1.3-2
List of explanatory variables in the final models

	List of explanatory variables in propensity model
Panel 5	Date of the last attempt, split household indicator, number of call attempts, refusal indicator, number of tracing attempts, province of residence, number of contacts, time slice group and language of previous interview
Panel 6	Date of the last attempt; call after first refusal; contact during evening time slice, number of tracing attempts, number of call attempts, special circumstances, contact during late afternoon time slice, language issues, call after first sent to senior interviewer, call after first refusal, contact during the morning time slice, number of appointments, contact during the early afternoon time slice, absent for duration of survey, call after first tracing attempt, number of all attempts, sent to senior interviewer, number of contacts, interview interrupted, call after tracing lead

When considering the cases that were in-progress as of the cut-off, the response rate at the end of collection for Panel 5 and Panel 6 were 25.7% and 26.7%, respectively. By contrast, looking at the response rate for the in-progress cases in the top decile, the response rates increase to 54.5% for Panel 5 and 72.9% for Panel 6.

4.2 Collection Phases

During the initial collection phase, collection will be conducted in the same way as in previous cycles. However cases that are part of the RD group, and that reach the intermediate call attempt cap of 30, will be set aside. Work will be suspended for these cases until the start of the RD phase-in 1.

Once the RD phase-in 1 begins, for the RD group, cases that were moved to the intermediate cap group during the initial collection phase will now be accessible. After running the propensity model, the regular in-progress cases will be divided into three groups: households that have yet to be contacted; households that have been contacted and, based on the propensity model, are most likely to complete the interview (in the top decile); and all other household that have been contacted. Furthermore, there are many households requiring tracing during collection. The tracing cases whose predicted probabilities of response are in the top 10% of all predicted probabilities among the in-progress tracing cases, will be identified and placed in another group. Over the course of the RD phase-in 1, any non-tracing, in-progress, RD cases that reach 30 call attempts will be moved to the intermediate cap group. It is hoped that by placing the in-progress sample into easily identified groups, data collection managers will be able to more efficiently staff their interviewers.

Finally, when it is time to move to RD phase-in 2, cases will be prioritized using the Priority Groups that have been used in previous cycles of collection. Laflamme and Karaganis (2010) detail the criteria that will used to determine the appropriate time to move from one phase of collection to the next. There will be no single criterion that will signal the activation or end of a RD phase, but a collection of indicators that include, among others: response rate, productivity levels and percentage of budget spent.

5. Conclusions and future work

Faced with decreasing response rates and rising survey costs, the goal for introducing a RD strategy for the 2010 SLID data collection is to increase the sample representativity, while keeping collection costs constant. Using a propensity model to identify cases that have an increased likelihood of being interviewed will hopefully help to minimize the downward trend in response rates that SLID has seen over the years. Moreover, it is hoped that more effectively staffing interviewers, focusing effort and reducing survey costs can achieve this. Using an R-indicator to assess the sample representativity will help to maintain and even improve data quality.

The next collection for SLID begins in January 2010. At this time, SLID can benefit from the RD experiences of HES, whose collection ends November 2009. Any lessons learned from HES can help to improve upon the RD strategy proposed in this paper.

After the end of the 2010 SLID collection, it will be necessary to assess the effectiveness of the RD strategy by analyzing the paradata, comparing the results between the control and RD groups. Additionally, the RD strategy will need to be considered when performing the nonresponse adjustment to account for any biases that may have been introduced by using this type of adaptive approach to data collection.

References

Dufour, S. (2009). A Review of Response Rates for Statistics Canada Surveys, unpublished report, Ottawa, Canada: Statistics Canada.

Groves, R. M. and Heeringa, S. G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, *Journal of the American Statistical Association*, 169, pp. 439-457.

Laflamme, F. and Karaganis, M. (2010). Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada, to be presented at the European Quality Conference, Helsinki, Finland.

Laflamme, F. and Mohl, D. (2007). Research and Responsive Design Options for Survey Data Collection, paper presented at the Joint Statistical Meeting, Salt Lake City, USA.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response, *Survey Methodology*, 35, pp. 101-114.

WEIGHTING AND ESTIMATION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Propensity Score Weight Adjustment for Dual Sampling Frame

C. Boudreau, M.E. Thompson and M. Iraniparast [1]

Abstract

The International Tobacco Control Policy (ITC) Netherlands Survey is an ongoing longitudinal survey of more than 2,200 smokers, which began in 2008. It aims to study smoking behaviours and the impacts of national-level tobacco control policies. It is the first survey of the ITC Project (which now has surveys in over 18 countries) to utilize a dual web+RDD sampling frame. The web sampling frame is a database that consists of more than 200,000 Dutch respondents who have agreed to participate, on a regular basis, in research studies conducted by the international survey firm TNS NIPO. Although great care was taken by TNS NIPO to ensure that their database is an accurate representation of the Dutch population, some groups remain underrepresented and others overrepresented. The sampling weights were constructed to adjust as much as possible for this, but some selection bias remained.

In this talk we present a method that utilizes the smaller RDD sample of the ITC Netherlands Survey in conjunction with a propensity score (PS) to adjust the weights of the larger web sample. The method consists of fitting a PS model where response corresponds to survey mode, and then performing a post-stratification adjustment of the web sample weights using estimated PS's. The method is easily modified to handle missing values. We illustrate with data from Wave 1 of the ITC Netherlands Survey, but the method can also be used with other dual frame surveys where one frame can be considered unbiased.

[1] C. Boudreau (cboudreau@math.uwaterloo.ca), M.E. Thompson and M. Iraniparast, University of Waterloo, Canada

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Longitudinal Estimation in the European Survey of Income and Living Conditions

Ralf Münnich and Stefan Zins[1]

Abstract

Measuring poverty and social cohesion within the European Commission is performed via the Laeken indicators. Most of these indicators are cross-sectional measures. To measure the Laeken indicators, the European Survey of Income and Living Conditions (EU-SILC) was launched as a rotational panel survey which enables measuring the change of these indicators over time.

The aim of the paper is to estimate the change of selected poverty and inequality measures and to assess their estimates' accuracy via linearization techniques. The covered measures include the GINI, the quintile share ratio, and the at-risk-of-poverty rate. Additionally, special emphasis is laid on the survey design as a rotational sampling scheme. The research includes the results of a Monte-Carlo simulation study.

Key Words: Poverty Indicators, Inequality Measures, Variance Estimation For Change, Rotational Panel.

1. Introduction

In March 2000, the Lisbon European Council has asked the Member States of the European Union to make steps towards improving social cohesion and combating poverty by 2010. In order to adequately measure poverty and social cohesion, the Laeken indicators (LI) were agreed on as a set of indicators to be measured and published each year by all member states. A thorough description of the Laeken indicators can be drawn from Eurostat (2008).
Yet, the mere publication of the LI by each EU member is necessary but not sufficient to gauge the progress towards agreed EU objectives. This would require measuring the change of LI over time, e.g. comparing the estimates in two subsequent years. This would provide a valuable tool to evaluate the effectiveness of applied policies within the EU Member States towards the eradication of poverty and improving social cohesion (cf. Eurostat, 2005, Section 2).
The European Survey of Income and Living Conditions (EU-SILC) was launched as a rotational panel survey in order to adequately allow measuring the LI as cross-sectional estimates as well as over time. Due to the European Statistics Code of Practice, the LI estimates have to be accompanied by adequate accuracy measures which generally requires applying variance estimation techniques.

The aim of the paper is to present methods for estimating the variance of change for selected LIs over time. Both, differences and ratios as measures for change are considered. The main focus of this research is put on rotational sampling scheme used within EU-SILC which consists of four rotational quarters from which one will be substituted each year. Hence, to measure change from one year to the next, a two-dimensional sample exists where three quarters are overlapping each year and one is independent between the two points of time.

The next Section gives an overview of the selected Laeken Indicators and their estimates. Subsequently, linearization techniques for variance estimation are presented, first as cross-sectional measures in Section 3 and thereafter as variance estimates for change in Section 4. Section 5, finally gives an overview of the results from a Monte-Carlo simulation study based on the EU-SILC data in order to evaluate the proposed variance estimators in a close to reality framework.

[1]Ralf Münnich and Stefan Zins, University of Trier, Faculty IV, Economic and Social Statistics Department. D-54296 Trier, Germany (Muennich@uni-trier.de; Zins@uni-trier.de).

2. The Laeken Indicators and their Estimates

The LI of interest in this study are the GINI coefficient, the at-risk-of-poverty rate, and the income quintile share ratio. Although the scope of the LI stretches far broader then the selected statistics, they cover the methodology of interest needed for most of the other income-poverty and income inequality measures. For the full set of the LI indicators we refer to Eurostat (2008) or Osier (2009).

The *at-risk-of-poverty rate* (ARPR) is the share of persons with an equivalised disposable income below the so-called poverty threshold, given as 60% of the national median equivalised disposable income. Thus, the ARPR can be written as

$$\text{ARPR} = \frac{1}{N}\sum_{j\in U}(y_j < 0.6\cdot Q_{0.5}), \qquad [1]$$

where N is total number of persons of the finite population $U=\{1, ..., N\}$ and y_j is the equivalised disposable income of the jth person in U. $Q_{0.5}$ is the median of the cumulative distribution function $F(y)$, with $Q_{0.5} = \inf_{y\in\square}(F(y) > 0.5)$ and $F(y) = \frac{1}{N}\cdot\sum_{j\in U}1(y_j \le y)$. $1(.)$ denotes the indicator function which is 1 if the condition is fulfilled and 0 otherwise. ARPR is estimated by

$$\hat{\text{ARPR}} = \hat{F}(0.6\cdot\hat{Q}_{0.5}), \qquad [2]$$

where $\hat{F}$ is an estimate for the distribution function using survey weights and $\hat{Q}_{0.5}$ is an estimate for the median. For details we refer to Osier (2009). The *S80/S20 Income quintile share ratio* (R8020) is the ratio of total income received by the 20% of the country's population with the highest income (highest quintile) to that received by the 20% of the country's population with the lowest income (lowest quintile).

$$\text{R8020} = \frac{\sum_{j\in U} y_j\cdot 1(y_j > Q_{0.8})}{\sum_{j\in U} y_i\cdot 1(y_i \le Q_{0.2})}$$

where $Q_{0.2}$ and $Q_{0.8}$ denote the lowest quintile and the highest quintile respectively. The R8020 indicator can be estimated via

$$\hat{\text{R8020}} = \frac{\sum_{j\in S} w_j\cdot(y_j - y_j\cdot 1(y_j \le \hat{Q}_{0.8}))/\sum_{j\in S} w_j\cdot(1-0.8)}{\sum_{j\in S} w_j\cdot y_j\cdot 1(y_j \le \hat{Q}_{0.2})/\sum_{j\in S} w_j\cdot 0.2},$$

where w_j is the survey weight corresponding to the jth person in some sample S. The *Gini coefficient* (GINI) which is given for a finite population as

$$\text{GINI} = \frac{1}{N}\cdot\sum_{j\in U}(2\cdot F(y_j) - 1)\cdot\frac{y_j}{\mu},$$

which is estimated by

$$\hat{\text{GINI}} = \frac{1}{\hat{N}}\cdot\hat{\mu}\cdot\sum_{i\in S} w_i\cdot(2\cdot\underbrace{\frac{1}{\hat{N}}\cdot\sum_{j\in S} w_i\cdot 1(y_j \le y_i)}_{\hat{F}(y_i)} - 1)\cdot y_i.$$

Here, the total population is estimated by $\hat{N} = \sum_{j\in S} w_j$ using the survey weights w_j.

3. Variance Estimation using Linearization Techniques

There are different approaches to estimate the variance of non-linear statistics; Resampling methods and linearization techniques. An overview of resampling methods for inequality measures is given in Kovačević and Yung (1997). Since the focus of this paper is laid on linearization methods, we refer for an overview of resampling methods to Davison and Sardy (2007) or Münnich (2007) and the references therein.

Linearization methods cover Taylor and Woodruff linearization (cf. Woodruff, 1971, Andersson and Nordberg, 1994) or more general methods such as the application of estimating equations (cf. Kovačević and Binder, 1997) and influence functions (see Deville, 1999). Demnati and Rao (2004) provide an alternative method for deriving Taylor linearization variance estimates.

A general framework based on the concept of influence functions has been presented by Deville (1999), which was first introduced into the field of robust statistics by Hampel, Ronchetti, Rousseeuw and Stahel, (1986). Deville (1999) employs a general class of non-linear statistics for a finite population U based on the concept of a measurement functional. A population parameter of interest θ can then be described as a functional $T(M)$ with respect to a finite discrete measures M which allocates a weight of 1 to all $k \in U$. If θ is to be estimated from a sample S, let $\hat{M}$ be the estimator of M, denote the measure allocating a weight w_k to any observations $k \in S$ and to all $k \notin S$ a weight 0. Where w_k can be a survey weight of any kind associated to the kth unit in the sample. The estimator for $T(M)$ is obtained by substituting M by $\hat{M}$ and therefore we may write $\hat{\theta} = T(\hat{M})$. The variance of this estimator can be approximated with that of a linear statistic

$$V(T(\hat{M})) = V\left(\sum_{k \in S} w_k \cdot u_k\right), \qquad [3]$$

where u_k is the linearized value at k given by $u_k = IT(M; k)$ and IT stands for the *influence function* of T in M. The influence function of a functional T is given by

$$IT(M;k) = \lim_{t \to \infty} \frac{1}{t} \cdot \left[T\left(M + t\delta_k\right) - T\left(M\right)\right], \qquad [4]$$

where δ_k is the Dirac measure of observation k denoting the unit mass assumed at observation k.

If expression [4] has to be estimated from an observed sample S the sample measure $\hat{M}$ is to be substituted for M in [4]. The rules of construction for influence function are essentially those used to derivate functions in differential calculus. For a set of rules see Deville (1999) and the assumptions which need to hold to apply this approximation see Deville (1999) or directly for the Laeken indicators see Osier (2009).

How to compute the linearised values of the ARPR is shown below. First we need the influence function for $F(y)$, denoted by $IF(y; k)$. The influence function for $F(y)$ at observation k can be derived as

$$IF(y;k) = \frac{1}{N} \cdot \left[1(y_k \le y) - F(y)\right] \qquad [5]$$

(cf. Osier, 2009, as a result from Deville, 1999). From Equation [2] the influence function of the ARPR can be written as (Deville, 1999):

$$IARPR(k) = \frac{1}{N} IF(0.6 \cdot Q_{0.5}; k) + f(0.6 \cdot Q_{0.5}) \cdot 0.6 \cdot IQ_{0.5}, \qquad [6]$$

where $f(y)$ is the derivative of a smoothed function of $F(y)$ and $IQ_{0.5}$ is the influence function of the median, which is given by

$$IQ_{0.5} = -\frac{1}{f(Q_{0.5}) \cdot N} \cdot \left[1(y_k \le Q_{0.5}) - 0.5\right]. \qquad [7]$$

Finally, inserting [5] and [7] into [6] we can write influence function of the ARPR at *k* as

$$IARPR(k) = \frac{1}{N}\left[1\left(y_k \le 0.6 \cdot Q_{0.5}\right) - F(0.6 \cdot Q_{0.5})\right] - \frac{1}{N}\left(\frac{f(0.6 \cdot Q_{0.5})}{f(Q_{0.5})}\right) \cdot \left[1(y_k \le Q_{0.5}) - 0.5\right].$$

Since F is a non smooth function its derivative is always 0 or not defined in the traditional way. In order to smooth F we can apply kernel density estimators (cf. Berger and Skinner, 2003). Special attention may be put on the smoothing function since the underlying income distribution can be very skewed. Spline function may be applied alternatively.

The linearized values u_k for the GINI and the R8020 can also be derived using the concept of influence functions. To see how exactly they can be derive for the GINI see Deville (1999) or Osier (2009) and for applying the approach based on estimating equations see Kovačević and Binder (1997). The influence values for the income quintile share ratio R8020 can be drawn from Hulliger and Münnich (2007) as well as Langel and Tillé (2009) or Osier (2009). An empirical comparison between linearization and resampling methods for cross-sectional measures can be found in Münnich et al. (2010).

4. Variance estimation for the difference and ratio of LI estimators

The difference $\theta_{\Delta:0,1} = \theta_1 - \theta_0$ and the ratio $\theta_{R:0,1} = \theta_1 / \theta_0$ of two indicator values at times t=0, 1 can be estimated straight forward. The variance estimates are computed according to the rules given in Section 3 using the influence values of the cross-sectional measures. This yields

$$V\left(\hat{\theta}_{R:0,1}\right) = V\left(\frac{\hat{\theta}_1}{\hat{\theta}_0}\right) \cong V\left(\frac{1}{\hat{\theta}_{t_0}}\left[u_1 + \hat{\theta}_{R:0,1} \cdot u_0\right]\right) \cong \frac{1}{\hat{\theta}_0^2}\left[V\left(\hat{\theta}_1\right) + \hat{\theta}^2_{R:0,1} \cdot V\left(\hat{\theta}_0\right) - 2 \cdot \hat{\theta}^2_{R:0,1} \cdot Cov\left(u_0, u_1\right)\right], \quad [8]$$

where u_0 and u_1 are the linearised values corresponding the indicator values θ_0 and θ_1 respectively.

The variance of the difference is similarly derived with $V\left(\hat{\theta}_{\Delta:0,1}\right) = V\left(\hat{\theta}_1\right) + V\left(\hat{\theta}_0\right) - 2 \cdot Cov\left(u_0, u_1\right)$. The variances in Equation [8] are simply the cross-sectional estimates according to [3] and do not need any further investigation. The crucial part is the estimation of the time dependent covariance of the influence values. This covariance has to be estimated considering the rotational sampling scheme. The theoretical justification is given by using the bivariate observations at two points of time where one quarter has independent observations. A more general approach with two samples is derived in Goga, Deville, and Ruiz-Gazen (2009) who first applied the theory to the estimation for change of the GINI index.

5. The Monte-Carlo Study

The main goal of the Monte-Carlo Study is to evaluate empirically the accuracy of the proposed variance estimators for change within the context of EU-SILC. The cross-sectional part of the survey consists of four quarters, each of them sampled in four successive years which adds the longitudinal dimension to the dataset. After one quarter has stayed for four years in the survey it drops out and is substituted be a newly drawn quarter. This procedure results in a cross-sectional dataset composed of four quarters per year, though the longitudinal data in two subsequent years spans only three quarters.

For the simulation study the EU-SILC 2006 longitudinal dataset was chosen as a finite population. The annual equivalized disposable household income in the years 2005 and 2006 was selected to compute the LIs. The data includes observations from all EU member states but Bulgaria, Romania, and Malta plus Iceland and Norway.
Two income distributions were considered in the study, EU-INC the actual EU-SILC income distribution excluding negative incomes and EU-SILC-ST as a truncated distribution of EU-INC where incomes above 150,000€ were removed. The population size N of EU-INC is 132,616 and for EU-INC-ST 132,443. The several income distributions where used to account for the non-robustness of the R8020 to high inequalities towards the upper tail of a distribution. Negative income values had to be excluded from the dataset since the GINI coefficient is not defined for negative incomes.

To implement a rotational sampling design the populations have been randomly split into rotational quarters. The sampling scheme can now be described in the following way:

1. A first sample of size n is independently drawn from all quarters.
2. A second sample of size $n/4$ is drawn from only one of the quarters.

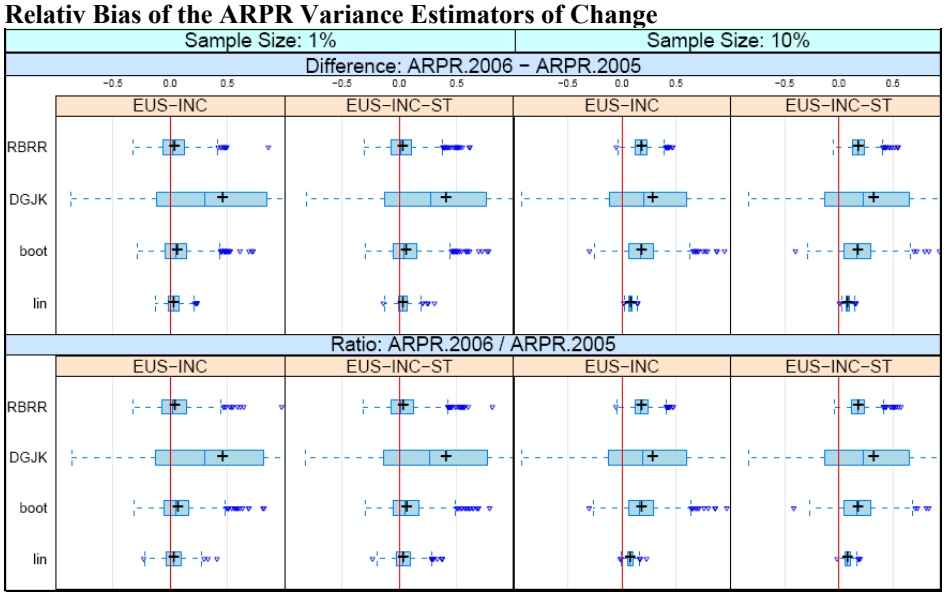
The samples in steps 1 and 2 were drawn in a stratified sampling scheme with proportional allocation where the countries were used as a stratification variable.

The variances in Equation [8] were computed using a standard variance estimator for stratified random sampling (cf. Andersson and Nordberg, 1994, p. 403). The covariance is also estimated using its standard variance estimator for stratified random sampling and applying it to the influence values at both times while omitting the linearised values originating from the independent quarter.

For some specific cases of different kinds of overlapping samples Qualité and Tillé (2008) provide an overview of variance estimators for change based on Hovitz-Thomson style estimators. See also Tam (1984) for variance estimation from overlapping samples.

Figure 5-1 shows boxplots for the relative bias of several ARPR variance estimators of change. The graph is split in four segments, each dealing with a different objective. The upper half refers to the difference estimator and the lower to the ratio estimator. The left part of the figure shows the results for a sample fraction of 1% and the right part for 10%. Each segment contains the boxplots for the two income distributions EUS-INC (left) and EUS-INC-ST (right). Four different variance estimators have been considered, three resampling methods and *lin* the linearized variance estimator in [8]. The three resampling methods used are *RBRR*, a balanced repeated replication estimator (Kovačević and Yung, 1997), *DGJK* a Delete-a-Group Jackknife (Kott, 2001) and *boot*, a non-parametric Monte-Carlo bootstrap. The linearized variance estimator performs best for all objectives. Within the resampling methods both *RBRR* and *boot* are less biased than *DGJK*, which suffers from a strong positive bias. The variance estimates from the ARPR seem to suffer from a small bias. However, the variability is diminishing in larger samples. The GINI and R8020 show similar results while having a much smaller bias and seem to be more stable. However, in these cases the combination of small samples and outliers in the income distribution yields heavily skewed variance estimation distributions, as expected. The confidence interval coverage rates show good results for EUS-INC-ST but may cause some problems for EUS-INC in the case of GINI and R8020.

Figure 5-1
Relativ Bias of the ARPR Variance Estimators of Change



6. Summary and outlook

Within the scope of this paper, we have shown that linearization methods yield accurate variance estimates for the highly non-linear statistics of the Laeken indicators. Some problems may arise when income distributions contain outlying incomes. The ARPR is less sensitive towards skewed distributions but tends to be more biased with regards to variance estimation. For the GINI and R8020 estimates seem to be relatively non-robust against outliers but very efficient in smoother cases.

Future research may incorporate calibration methods in order to stabilize the results and to gain more accurate estimates for all Laeken indicators. With regards to the bias observed in case of the ARPR the smoothing of the cumulative distribution function may be improved by using spline regression instead of kernel density estimates.

Acknowledgement

The research was done within the AMELI research project (Advanced Methodology for European Laeken Indicators; http://ameli.surveystatistics.net) which is financially supported Seventh Framework Programme of the European Commission.

References

Andersson, C. and Nordberg, L. (1994). A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys - Theory and Software Implementation, *Journal of Official Statistics*, 10, pp. 395-405.

Berger, Y. and Skinner, C. (2003). Variance estimation for a low income proportion, *Journal of Applied Statistics*, 52, pp. 457-468.

Davison, A. C. and Sardy, S. (2007). Resampling Variance Estimation in Surveys with Missing Data, *Journal of Official Statistics*, 23, pp. 371-386.

Demnati, A. and Rao, J. N.K. (2004). Linearization Variance Estimators for Survey Data, *Survey Methodology*, 30, pp. 17-26.

Deville, J. C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques, *Survey Methodology*, 25, pp. 193-203.

Eurostat (2005). Measuring progress towards a more sustainable Europe, European Commission, Eurostat, DOC KS-68-05-551-EN.

Eurostat (2008). Algorithms to compute Overarching Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC), Working group Statistics on Living Conditions, European Commission, Eurostat, DOC LC-ILC/11/08/EN – REV.

Goga, C. and Deville, J. C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data, *Biometrika*, 96, pp. 691-709.

Hulliger and Münnich (2007). Variance Estimation for Complex Surveys in the Presence of Outliers, *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, pp. 3153-3161.

Kott, P. S. (2001). The Delete-a-Group Jackknife, *Journal of Official Statistics*, 17, pp. 521-526.

Kovačević, M. C. and Binder, D. A. (1997). Variance Estimatin for Measures of Income Inequality and Polarization – The Estimating Equations Approach, *Journal of Official Statistics*, 13, pp. 41–58.

Kovačević, M. C. and Yung, W. (1997). Variance Estimation of Measures of Income Inequality and Polarization - An Empirical Study, *Survey Methodology*, 23, pp. 41-52.

Langel, M. and Tillé, Y. (2009). Statistical Inference for the Quintile Share Ratio, unpublished report, Neuchâtel, Swiss, University of Neuchâtel.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen, *Austrian Journal of Statistics*, 37, pp. 319-334.

Münnich, R., Zins, S. and Kolb, J.-P. (2010). Accuracy of Poverty and Inequality Indicators for the German EU-SILC survey, *In submission*.

Osier, G. (2009). Variance estimation for Complex Indicators of poverty and inequality using linearization Techniques, *Survey Research Methods*, 3, pp. 167-195.

Qualité, L. and Tillé, Y. (2008). Variance Estimation of Changes in Repeated Surveys and its Application to the Swiss Survey of Value Added, *Survey Methodology*, 34, pp. 173-181.

Tam, S. M. (1984). On Covariances from Overlapping Samples, *The American Statistician*, 38, pp. 288-289.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, 66, pp. 411-414.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Weighting and Variance Estimation for the German Dual Frame Household Panel Survey "PASS"

Hans Kiesl [1]

Abstract

The German Institute for Employment Research has just completed the second wave of a new (annual) panel survey focusing on low income households, which is designed as a dual frame survey. The first frame is a register of households that are currently receiving some kind of unemployment benefits; the second frame consists of an address register of the whole population. In the beginning, 6,000 households were selected from each frame (with PPS sampling of zip codes in the first stage), resulting in a large variance of design weights between the two samples. For both subsamples and for the combined sample, weights for households as well as individuals are provided, resulting in six different sets of cross sectional weights for each wave.

In this paper, we describe the challenges of weighting and variance estimation for the first two waves of our survey, including trimming of extreme weights, coordination of household and individual weights, nonresponse adjustment strategies for households and individuals, a comparison of different weight share methods to tackle changes in household composition over time, and the use of convex weighting to integrate a second wave sample of births from the smaller sampling frame (i.e. new households in need of benefits).

[1] Hans Kiesl, Institute for Employment Research (IAB), Germany (Hans.Kiesl@iab.de)

ACCOMMODATING MISSING DATA IN LONGITUDINAL SURVEY DATA ANALYSIS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Modelling and Analysis of Durations Based on Longitudinal Survey Data

J.F. Lawless and D. Mariaca Hajducek[1]

Abstract

The durations of spells that persons spend in certain life states are of interest in relation to life history processes for which longitudinal surveys provide data. In spite of advances in the analysis of life history data over the past 20 years, numerous challenges remain in dealing with data from longitudinal surveys. This article discusses the modelling and analysis of spell durations based on such data. Challenges arising from missing data, loss to follow-up of panel members, the complexity of the survey population and complex survey design are discussed. Some methods of analysis are outlined and illustrated briefly on data on jobless spells from Statistics Canada's Survey of Labour and Income Dynamics.

Key Words: Inverse Probability Of Censoring Weights, Missing Data, Pseudo-Likelihood, Survey Weights, Survival Analysis.

1. Introduction

Data on education, employment, health and other areas of an individual's life history are collected in longitudinal surveys. Such data raise the possibility of life history analysis, in which processes involving events and other variables associated with individuals over their lifetimes are studied and modelled. One framework for doing this is multistate modelling, in which individuals move among certain states during some period of time. Such models are, for example, used in studying employment, conjugal partnerships, and child-bearing. These models are specified and analyzed by considering the durations of spells spent in specific states along with transitions from one state to another. The purpose of this paper is to consider ways of analyzing the durations of such spells based on longitudinal survey data. Duration analysis and life history analysis are well developed areas with considerable software support (e.g. Andersen et al. 1993, Kalbfleisch and Prentice 2002, Cook and Lawless 2007). However, data collected through longitudinal surveys bring with them problems that are not adequately addressed by standard methods. Our objective is to discuss such problems and consider ways to deal with them.

Populations studied in longitudinal surveys tend to be heterogeneous, and individuals for a panel are typically chosen according to a complex survey design. Data on individuals in the panel are collected only intermittently (e.g., every one or two years), over a specified period of time. In the case of Statistics Canada's Survey of Labour and Income Dynamics (SLID), which we discuss later, individuals are interviewed annually over a period of six years. Missing data on specific variables is common (item non-response). In addition, individuals may be lost to follow-up before the final interview, or they may be missing at an interview but reappear for a later interview. As we discuss later, loss to follow-up may be related to an individual's previous outcomes or covariates.

This paper deals with the modelling and analysis of spell durations. The objectives of such analysis can broadly be grouped into two categories. The first is descriptive, in which the objective is to estimate features of some distribution across the population from which the panel is drawn. For example, we may wish to estimate the marginal distribution of the length of jobless spells occurring for Ontario residents in 2006, or we may wish to relate jobless spells to covariates associated with the individual or with economic conditions. The second category encompasses explanatory analysis of individuals' life history processes. In this case we are interested in association between recurrent spells of unemployment and employment, and their relationship to covariates. Longitudinal data allows spell duration distributions to be estimated more readily than can be done from retrospective data obtained from a single interview. Nevertheless, problems with missing data or measurement error can make longitudinal surveys much less useful than a longitudinal cohort study with close tracking of individuals.

2. Duration analysis for longitudinal survey data

We consider settings where individuals occupy sequences of states over time. For example, a very simple employment model has states Employed, Unemployed and Out of the Labour Force. Our interest here is in the durations of the spells spent in

[1] J.F Lawless, Dept. of Statistics and Actuarial Science, University of Waterloo, Waterloo ON Canada N2L 3G1 (jlawless@uwaterloo.ca); D. Mariaca Hajducek, Dept. of Statistics and Actuarial Science, University of Waterloo, Waterloo ON Canada N2L 3G1 (hajducek@gmail.com)

specific states. A spell in a state ends by a transition to another state, and typical data for an individual observed over a period of time consists of a sequence of states E_1, E_2, E_3,... and the durations Y_1, Y_2, Y_3,… of the spells in each of those states. In some cases we may wish to model separately the transitions from a state to each of the other states an individual may move to, but for simplicity we will consider durations without regard to which state an individual moves to.

Assume that individuals in a panel are to be observed at times *0,1,2,…, M* over a time period *[0, M]*; for simplicity we assume that time is in years and that individuals are seen at the start of each year *t*. For *t = 1,…,6* information about the previous one year period *(t-1,t]* is collected; at *t = 0* baseline information is obtained. Let the states occupied by individual *i* over *[0, M]* be (in order) $E_{i_1}, E_{i_2}, ...E_{im_i}$. Let u_{ij} and v_{ij} be the entry and exit times for state E_{ij}; note that in general, $u_{i1} < 0$ and $v_{im_i} > M$. The duration of the spell in E_{ij} is $Y_{ij} = v_{ij} - u_{ij}$. If an individual is lost to follow-up prior to time M then some of their spells may be unobserved or their duration censored. The duration of the last spell is typically censored even if the individual is followed up to time *M*; that is, we know only that $Y_{im_i} > M - u_{im_i}$.

Analysis of a single well-defined duration from survey data is well developed. Suppose in particular that a duration Y_i and covariate vector x_i are considered for individuals $i = 1,...,n$ who are selected by a complex sampling design. Estimation under Cox proportional hazards regression models was addressed by Binder (1992), who considered design-weighted estimating functions for estimation of regression coefficients and provided design-based variance estimates for regression coefficient estimators. Lin (2000) and Boudreau and Lawless (2006) considered estimation of finite population and super-population parameters, providing variance estimates for regression parameter and cumulative baseline hazard function estimators. These methods are implemented in software packages such as SAS, SUDAAN and R/SPLUS. Lawless and Boudreau (2003) reviewed similar methods for parametric regression models.

The methods of the preceding paragraph can also deal with multiple durations Y_{ij} with covariate vectors x_{ij} for individuals $i = 1,...,n$. Kovacevic and Roberts (2007) deal with multiple spells by using the common "working independence" approach based on the Cox model and robust variance estimates (e.g. Lin, 1994). However, this approach was designed for multiple durations that are observed in parallel, and is subject to bias in the case of a correlated sequence of durations (e.g. see Cook and Lawless, 2007, Section 4.4). In this case we would typically need to include information about previous durations and states visited in a model for a specific duration, in order to allow for dependencies among the durations for a given individual. In spite of this solid basis for inference about durations from survey data, however, there are several gaps in existing methodology. We discuss these next.

3. Some gaps in methodology

For simplicity we consider spell durations *Y* for a specific state *E = e*, and denote the conditional distribution function for *Y* given covariates *x* as *F(y|x)*. Care is needed in specifying the population for Y, so that Ys are comparable across time and across individuals and model parameters associated with Y are interpretable. We discuss four areas in which there are methodological gaps and challenges. These are a result of the incompleteness of the data obtained for individuals in a longitudinal survey, and the nature of the design by which individuals are selected.

3.1 Dependent loss to follow-up

Loss to follow-up (LTF) in longitudinal surveys occurs when an individual is not "seen" (that is, their information is not collected) after some interview $C < M$; we refer to C as the LTF time. We assume here that all individuals are seen at the time of selection (interview 0) and at the first subsequent interview (interview 1) but that they may become lost to follow-up at any of interviews *2,…,M-1*. Individuals who do not become LTF are seen for the last time at interview *M*. It is also possible that an individual's data are not available for some interview, but become available again at the next; we consider this in section 3.2.

A duration Y_i has a notional censoring time C_i and standard survival analysis methods, including those mentioned in Section 2, require that Y_i and C_i are independent, given covariates x_i used in the analysis. This is often violated because either (a) covariates affecting both Y_i and C_i are excluded from the model, or (b) Y_i and C_i are both related to life history prior to the time u_i at which the spell started, which is not included in the model. From this point on we will assume for the sake of exposition that individuals are seen once a year, in January. Suppose a person has a spell starting during 2006 after the interview and that the spell ends during 2007. There are then three possible scenarios: (i) if the individual was LTF as of January 2007 (that is, they were seen in January 2006 but not in January 2007), their LTF time is effectively January 2006 and we would not even know about their spell; (ii) if the individual was seen in January 2007 but not seen in January 2008 then their effective LTF time is January 2007 and Y_i is censored, with censoring time equal to their last interview date minus u_i; (iii) if the individual is seen in January 2007 and January 2008 then their duration Y_i is observed. It follows that if Y_i and the probability the individual

is LTF at the 2006, 2007 or 2008 interviews both depend on covariates not in the regression model, then censoring is dependent, and may cause substantial bias in estimates of the duration distribution.

The occurrence of a spell, and its start time u_i, may also depend on covariates. If such a covariate is also related to LTF and duration then bias may once again arise if the covariate is not in the duration model.

3.2 Other types of missing data

When a spell that starts at u_i and ends at v_i occurs, either the start time, end time or covariates may be unobserved because of LTF or (temporary) non-response. For example, for Ontario residents in the 1999 SLID panel, 29% of jobless spells had unknown start dates and of spells with known start dates, a further 18% had one or more missing covariate values. In addition, about 30% of individuals became LTF before the year 6 interview (January 2005). Reasons for missing values include unit non-response, in which an individual misses an annual interview but reappears in a later year, item non-response due to refusal or inability (e.g. in the case of a proxy respondent) to answer and LTF, discussed in the preceding section.

3.3 Variance estimation

Methods of dealing with the survey design via weighting have been developed for design-based inferences in papers mentioned in Section 2. Such methods also address cluster effects in the population. To deal with the incorporation of calibration or adjustments for LTF, we need to allow for weights with components that are estimated. Bootstrap methods are often used, although their validity has not been rigorously established in many cases. Existing methods do not satisfactorily address weighting methods for duration analysis, nor association among duration times. We discuss this in the following section, where we consider ways to deal with the issues raised above.

4. Methods for Handling Dependent LTF and Missing Data

4.1 Dependent loss to follow-up

Suppose that individuals are to be seen at times $t= 0,1,...,M$, as specified in Section 2. At time t $(t = 1,...,M)$, information concerning the one year period $(t-1, t]$ is collected. Baseline information collected at $t = 0$ typically does not include detailed information about sojourns in states occupied before then. Complete data for an individual thus consists of their baseline information plus data covering the subsequent M years. C_i denotes the last year in which information is collected for individual i; we assume everyone has data for year 1.

Two points are important in the following development: a spell can occupy (parts of) more than one year, and the probability a person is LTF at year t can depend on their previous life history. The methods we describe require that data be missing at random in the terminology of Little and Rubin (2002) and in this case we therefore assume that LTF at year t can depend only on information available up to year $t-1$. In some settings it is likely that such LTF may depend additionally on occurrences during year t but there is no way to handle this on the basis of the observed data without making untestable assumptions. This is discussed further in Section 6.

Suppose that the durations of spells in some specific state are of interest and let $D_i(t)$ represent the information on such durations for individual i in year t. Assume that a family of models relating $D_i(t)$ to a set of covariates is of interest. We denote this as

$$P_\theta(D_i(t) \mid Z_i^0(t)) = \Pr(D_i(t) \mid Z_i^0(t); \theta), \tag{1}$$

where $Z_i^0(t)$ includes information about $\{D_i(1), D_i(2),..., D_i(t-1)\}$ and external covariates $x_i(t)$ that are of interest. The parameter θ (which may be multidimensional or even contain functional components) specifies the family (1) and it is our objective to estimate θ and assess the adequacy of the fitted models.

For superpopulation inference it is assumed that for some θ_0 the expectation of the log likelihood score

$$U_{it}(\theta) = \frac{\partial \log P_\theta(D_i(t) \mid Z_i^0(t))}{\partial \theta}, \tag{2}$$

conditional on $Z_i^0(t)$, is zero for a random member of the population. However, when a set of individuals $i=1,2,...,n$ from a longitudinal survey panel experience the spells in question, this is no longer true in general. Furthermore, a panel individual may become LTF before year t. We now specify a framework to deal with this.

Let $R_{it} = I(C_i \geq t)$ denote that individual *i (i=1,...,n)* in the panel is seen at year *t*; we assume there is no missing data in $D_i(t)$ or $Z_i^0(t)$ in this case. Let $Z_i(t)$ be a vector of covariates containing survey design variables Z_i^D and other possible time-varying factors $Z_i^C(t)$ such that (a) selection for the survey panel is independent of $D_i(t)$ and $Z_i^0(t)$ given Z_i^D and (b) R_{it} is independent of $D_i(t)$ and $Z_i^0(t)$ given $Z_i^C(t)$. It is assumed that $Z_i^C(t)$ depends only on observed information up to year *t-1*, as does $Z_i^0(t)$. We denote π_i=Pr(individual *i* selected for the panel | Z_i^D) and $p_{it}(\alpha)$=Pr(R_{it} =1| $Z_i^C(t)$;α); note that whereas the inclusion probability π_i is a known function of Z_i^D, the LTF probability $p_{it}(\alpha)$ is based on a family of models indexed by parameter α. We now base estimation of θ on the estimating function

$$U^W(\theta) = \sum_{i=1}^{n} \sum_{t=1}^{M} \frac{R_{it}}{\pi_i p_{it}(\hat{\alpha})} U_{it}(\theta), \tag{3}$$

where $\hat{\alpha}$ is an estimate of α.

The estimating function $U^W(\theta) = 0$ produces a consistent estimator $\hat{\theta}$ of θ provided the LTF model $p_{it}(\alpha)$ is correct and $\hat{\alpha}$ is a consistent estimator of α (Robins et. al, 1995). The "inverse probability of censoring" (IPC) weights $p_{it}(\hat{\alpha})^{-1}$ have been utilized frequently in longitudinal survey settings (e.g. Miller et. al, 2001) but not for duration analysis. The design weights π_i^{-1} have been used on their own for duration data from surveys (e.g. Binder 1992, Lin 2000.)

To use (3) we require a model for dropout. A widely applicable approach is to model the distribution of R_{it} given $R_{i,t-1}=1$ via logistic regression as

$$\text{logit } \lambda_{it}(\alpha) = \text{logit Pr}(R_{it} = 1 \mid R_{i,t-1} = 1, Z_i^C(t); \alpha) = \alpha' Z_i^C(t). \tag{4}$$

Then, $p_{it} = \lambda_{i1}\lambda_{i2}\cdots\lambda_{it}$. The models (4) can be fitted for *t=1,...,M* from observed data, assuming that once a person is missing at year *t* (i.e. $R_{it}=0$), they are not seen again. This can be achieved by designating a person as LTF at the first year where no information on them is obtained.

The use of (3) requires that the data on spell durations be partitioned into pieces $D_i(t)$. For example, a spell that starts at time u_i in year 1 and ends at time v_i in year 2 is partitioned into pieces $D_i(1)$=*{spell started at t=u_i and is still in progress at t=1}* and $D_i(2)$=*{spell had current duration 1-u_i at t=1, and ended at t=v_i}*. The two pieces get different LTF weights p_{i1}^{-1} and p_{i2}^{-1} in (3). Standard software for the Cox model, such as **coxph** in R/SPLUS and **phreg** in SAS, allows this but many parametric survival analysis programs do not. Lawless and Mariaca Hajducek (2010) provide details concerning the application of estimating functions (3); an illustration is given in section 5.

4.2 Other types of missing data

Missing start times for spells or missing covariate values are in general problematic. Missing data patterns in longitudinal data can vary considerably across individuals. Even in cases where relatively straightforward methods are applicable, little is available in standard software. We briefly discuss some of these problems here.

Missing values can be dealt with through imputation, weighted pseudo-likelihood or likelihood methods. Replacement of a missing value by a single imputed value is unappealing when there is a good deal of missing data. Multiple imputation (Rubin, 1996) is better but requires assumptions about the processes leading to missing values. A third option, especially with categorical covariates, is to add a category called "missing value" for the covariate. This is not especially satisfactory when the objective is to assess the relationship of a covariate to a duration time variable but can be useful when prediction is of interest. The use of inverse probability weighted estimating functions (Robins et al. 1995) or maximum likelihood (Little and Rubin 2002, Cook and Lawless 2007) based on models for missing values can in principle deal with cases where the probability a value is missing is related to observed covariates or responses, but because of the need for such models, the methods are mainly useful when the same variables are missing across individuals. It is possible to extend methods in Robins et al. (1995), Lawless et al. (1999) and related references to deal with survey data but to date this has not been done for settings involving duration analysis. A serious issue is what to do when missing data patterns are not simple enough to use these methods. Some form of random imputation seems most feasible but there are considerable challenges as to how to perform the imputation and how to obtain variance estimates; see for example, Kenward and Carpenter (2007).

4.3 Variance estimation

We will discuss briefly how variance estimation can be handled in the case of losses to follow-up which are dealt with by the methods of Section 4.1. Similar methods can be developed for the inverse probability weighting or maximum likelihood methods mentioned in the preceding section.

Variance estimation based on estimating functions like (3) need to account for the estimation of parameters α in the weights and also for association among individuals and for recurrent spells within individuals. For fully parametric models this may be approached through asymptotic theory for estimating functions. For example, we consider for (3) and (4) two sets of estimating functions

$$U^W(\theta,\alpha) = \sum_{i=1}^{n}\sum_{t=1}^{M} \frac{R_{it}}{\pi_i p_{it}(\alpha)} U_{it}(\theta) \tag{5}$$

$$U^R(\alpha) = \sum_{i=1}^{n}\sum_{t=1}^{M} R_{i,t-1}\left\{\frac{R_{it}-\lambda_{it}(\alpha)}{\lambda_{it}(\alpha)(1-\lambda_{it}(\alpha))}\right\}\frac{\partial\lambda_{it}(\alpha)}{\partial\alpha}, \tag{6}$$

where (6) is a likelihood score equation score equation based on (4). An estimated covariance matrix for $\hat{\theta}$ can be obtained from the asymptotic covariance matrix for $\psi=(\theta,\alpha)$. This has the standard form $AB(A^{-1})'$ with $A=E(-\partial U/\partial\psi')$ and $B=Var(U)$, where $U=U(\psi)=(U^W(\psi)',U^R(\psi)')'$ with vectors $U,\psi,\theta,\alpha,$ etc. written in column form. The covariance matrix B should recognize cluster effects in the sample and population; see Lawless and Mariaca Hajducek (2010). Semiparametric models for which θ has functional components are more complicated. Lawless and Mariaca Hajducek (2010) consider Cox regression models.

5. Illustration

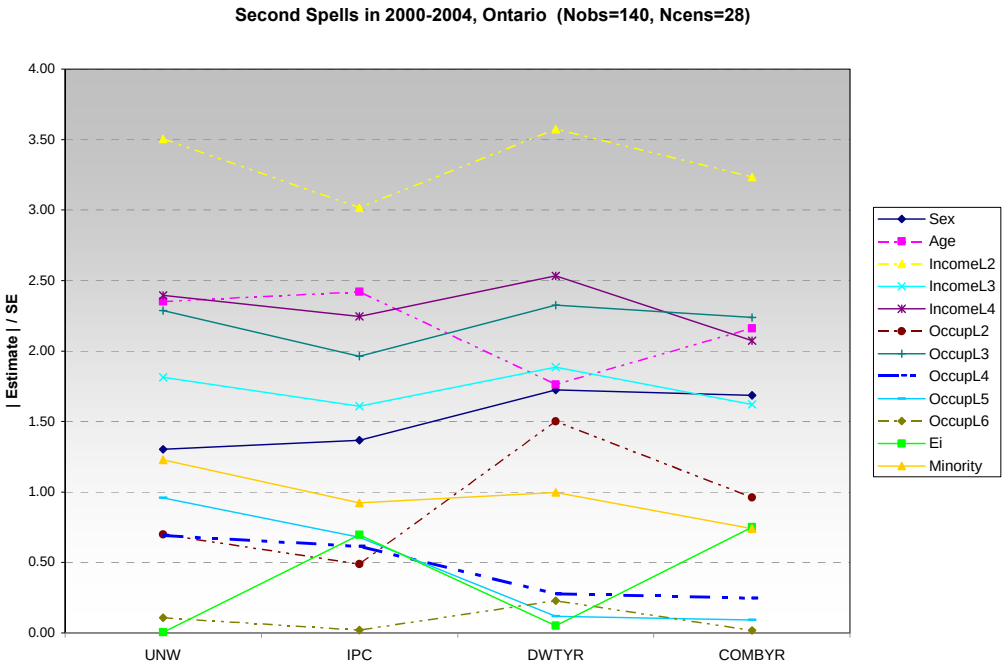
We give a short illustration of points discussed above by considering the duration of a jobless spell for the 1999 SLID panel. Labour information from the previous year is collected annually; that is, data for years 1999 to 2004 inclusive are collected at the start of each following year. A jobless spell is defined as the time period in which a person is out of work due to a permanent layoff and is looking for work.

Our example involves durations of jobless spells that started during the period 2000 to 2004, for persons 16 to 63 years old in 1999 who were resident in Ontario in that year. Spells with missing start times have been dropped. An important caveat of the illustration is that start times are assumed to be missing completely at random. In future work we plan to explore methods mentioned in Section 4.2, which would require only that start times be missing at random. Durations were modelled using Cox PH models and the estimating function $U^W(\theta)=0$ in (3). We consider up to three spells per individual. Of the 931 such spells, 236 had some form of missing value (missing start dates or covariates). Among the remaining 695 spells that were used for modelling, 441 were first, 168 were second and 86 were third spells. About 17% were censored. Categorical covariates considered here are: sex (female/male), income (low, low-med, med, high), occupation (services, manufacture, primary sector, construction, administrative, professional), employment insurance (no/yes), minority group (yes/no). The first level in each case was used as the reference category. Age was considered as a continuous variable. The start year for a spell and, for second or third spells, the mean length of previous jobless spells, were also included.

Modelling of LTF for each year involved all individuals. Missing values for categorical covariates were represented as separate categories. Twelve SLID variables and some of their interactions were assessed for inclusion in dropout models (4) for each year. Also, an indicator of whether the person had been jobless at the interview of the previous year was included. Variables that were significantly related to dropout across all years were age, education level, marital status, and employment status. Effects were in the direction one might expect; for example, persons married or in a common-law relationship were less likely to become LTF, and a person's education level was inversely related to the likelihood of LTF.

We show results here for four weighting options for estimation: (a) un-weighted ("UNW"), (b) weights estimated from modelling dropout ("IPC"), (c) longitudinal survey weights as of the start year of the spell ("DWTYR"), and (d) weights given by the inverse of the combined IPC and design probabilities ("COMBYR"). The figure below shows values of |estimate|/se(estimate) for each weighting option for the case of second jobless spells. The covariate giving the duration of the first jobless spell is not included; it is correlated with other covariates and was not significant. Standard errors were calculated recognizing clustering but not the fact that the IPC weights are being estimated. It is expected that variances estimated from a method that takes into account the IPCW randomness will be slightly smaller (Robins, et.al, 1995), which would tend to make p-values for covariates somewhat smaller.

The figure shows standardized regression coefficient estimates, and the significance of variables, to be fairly similar across the four weighting options. A reason for this may be that most jobless spells are under a year or two in length, and so the impact of LTF is not very strong. In addition, even the unweighted estimates use correct variance estimation which takes account of clustering. For the most significant variables, effects are in plausible directions. For example, higher income level in the preceding year is associated with shorter jobless spells, and higher age is associated with longer spells.



6. Concluding Remarks

The role of longitudinal survey data for detailed explanatory analysis of processes such as employment seems limited, unless problems involving missing data and accurate measurement can be resolved. We have not discussed problems of measurement error here, but we note that for certain variables this can be significant. The times attributed to past events at annual interviews are especially susceptible to error in cases where a concrete record of the event is not available. So-called seam effects (Cotton and Gilles 1998, Callegaro 2008) in which the times recorded for events are biased towards the times of data collection are well discussed, although methods of dealing satisfactorily with this appear limited. In any case, there is ample motivation for the study and development of methodology for addressing these issues.

On a final note, violations of the missing at random (MAR) requirement for the applicability of methods discussed here are almost certain to occur in any survey in which panel members are seen or contacted at widely spaced times. In that case the role of auxiliary data from administrative sources, or perhaps from tracing some of the individuals who are lost to follow-up, is very important. Methods for the incorporation of such data into duration analysis deserve attention. In the absence of such data, one can perform sensitivity analysis to assess the effect of non-MAR loss to follow-up or other types of missing data (e.g. Scharfstein and Robins 2002).

Acknowledgments

We thank Georgia Roberts, Christian Boudreau and Mary Thompson for comments and suggestions, and Georgia Roberts for her helpfulness during the second author's internship at Statistics Canada in the fall of 2008. We are grateful to Georgia Roberts and Pat Newcombe–Welch of the Southwestern Ontario Research Data Centre at the University of Waterloo for help

in dealing with SLID data. The support of MITACS and the National Program for Complex Data Structures (NSERC) is gratefully acknowledged.

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data, *Biometrika, 79*, pp. 139-147.

Boudreau, C. and Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data, *The Canadian Journal of Statistics, 34*, pp. 203-216.

Callegaro, M. (2008). Seam effects in longitudinal surveys, *Journal of Official Statistics, 24*, pp. 387-409.

Cook, R.J. and Lawless, J.F. (2007). *The Statistical Analysis of Recurrent Events*, Springer, New York.

Cotton, C. and Giles, P. (1998). The seam effect in the Survey of Labour and Income Dynamics, *SLID working paper No.75F0002,* Statistics Canada, Ottawa, ON.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition, Wiley, New York.

Kenward, M.C. and Carpenter, J. (2007). Multiple imputation: current perspective, *Statistical Methods in Medical Research, 16*, pp.199-218.

Kovacevic, M. S. and Roberts, G.R. (2007). Modelling durations of multiple spells from longitudinal survey data, *Survey Methodology, 33,* pp. 13-22.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edition, Hoboken, NJ.

Lawless, J.F. and Boudreau, C. (2003). Modelling and Analysis of Duration Data from Longitudinal Surveys*, Proceedings of Statistics Canada Symposium 2002: Modelling Survey Data for Social and Economic Research*.

Lawless, J.F. and Mariaca Hajducek, D. (2010). Analysis of durations in longitudinal studies with intermittent ascertainment of information, Unpublished manuscript.

Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression, *Journal of the Royal Statistical Society B, 61*, pp. 413-438.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine, 13,* pp. 2233-2247.

Lin, D.Y. (2000), On fitting Cox's proportional hazards models to survey data, *Biometrika, 87*, pp. 37-47.

Little L.J.A. and Rubin D.B. (2002). *Statistical Analysis of Missing Data*, 2nd edition, Wiley, New York.

Miller, M.E., Ten Have, T.R., Reboussin, B.A., Lohman, K.K. and Rejeski, W.J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple cause nonresponse, *Journal of the American Statistical Association*, *96,* pp.844-857.

Robins, J.M, Rotnitzky, A. and Zhao, L.P (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association, 90*, pp. 106-121.

Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association, 91*, pp. 473-489.

Scharfstein, D.O. and Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring, *Biometrika, 89,* pp. 617-634.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Analysis of Longitudinal Surveys with Missing Responses

Changbao Wu and Ivan Carrillo Garcia[1]

Abstract

Longitudinal surveys have emerged in recent years as an important data collection tool for population studies where the primary interest is to examine population changes over time at the individual level. The generalized estimating equation (GEE) approach is the most popular statistical inference tool for longitudinal studies. The vast majority of existing literature on the GEE method, however, uses the method for non-survey settings, and issues related to complex sampling designs are ignored.

We propose methods for the analysis of longitudinal surveys when the response variable contains missing values. Our methods are built within the GEE framework, with a major focus on using the GEE method when missing responses are handled through imputation.

We first argue why and further show how the survey weights can be incorporated into the so-called Pseudo GEE method under a joint randomization framework, and the missing responses are handled either by a re-weighting method or by imputation. Consistency of the resulting GEE estimators of the regression coefficients are established under certain regularity conditions. Linearization variance estimators are developed under the assumption that the finite population sampling fraction is small or negligible, a scenario often held for large scale population surveys. Finite sample performances of the proposed estimators are investigated through a simulation study. The results show that the proposed GEE estimators and the linearization variance estimators perform well under several sampling designs for both continuous and binary responses.

[1] Changbao Wu, University of Waterloo, Canada (cbwu@math.uwaterloo.ca); Ivan Carrillo Garcia, Statistics Canada

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

An Analysis of Intention to Quit Smoking in the ITC4 Survey Accounting for Missingness in Response and Covariate

Baojiang Chen and Mary Thompson[1]

Abstract

Data from longitudinal studies often feature both incomplete response and incomplete covariate data. The impact of missing data often depends on the frequency with which data are missing and the strength of the association between the missing data indicators and the response variables. When both response and covariate data may be incomplete it is important to take the association between the missing data indicators for these two processes into account through joint models. Inverse probability weighted generalized estimating equations are developed to deal with data that are missing at random. Empirical studies demonstrate that the consistent estimators arising from the proposed methods have very small empirical biases in moderate samples, and are more efficient than alternative methods which ignore the association between the missing data processes. An application to the International Tobacco Control (ITC) Four Country Survey demonstrates the usefulness of the proposed method.

Key Words: Association, Missing Response, Missing Covariates.

1. Introduction

Research on smoking has demonstrated that many smokers simply do not want to smoke. About 80% report that they want to quit. Between 40% and 50% try to quit in any given year, although only 3%--5% quit successfully for at least 12 months (Centers for Disease Control and Prevention 2002; Health Canada 2002; Hyland et al. 2004). It is desirable to study the factors that influence intention to quit, which is often a precursor of quit attempts.

The present study used data from the first six waves of the International Tobacco Control (ITC) Four Country Survey, a cohort survey of representative samples of adult smokers in Canada, the United States, the UK, and Australia. In this survey, respondents in each country are surveyed annually using parallel survey protocols and measures. A respondent's data can be incomplete due to missing responses and missing covariate data, or study subjects failing to answer questions. Problems of inference arise if the mechanism leading to the missing data is related to the values of response or covariates. For example, respondent fatigue or degree of addiction could be associated with non-response on specific items in a lengthy and probing questionnaire. In such cases, analyses based on individuals with complete data only can lead to invalid inferences. Under a missing completely at random (MCAR) mechanism (Little and Rubin 1987), analyses based on generalized estimating equations (GEE) (Liang and Zeger 1986) yield consistent estimates for the regression parameters. When the data are missing at random (MAR) or missing not at random (MNAR) (Little and Rubin 1987), analyses based on GEE give inconsistent parameter estimates. Robins, Rotnitzky, and Zhao (1995) developed a class of estimators based on an inverse probability weighted generalized estimating equations (IPWGEE) approach in a regression setting when the data are MAR.

The purpose of this illustrative analysis is to evaluate the dependence on some covariates of the intention to quit smoking in four countries -- Canada, United States, UK, and Australia -- when there are missing values for the responses and covariates. The approach used here is based on inverse probability weighted estimating equations in which the association between the missingness of the response and covariate is addressed. For details, please refer to Chen, Yi and Cook (2009).

The remainder of this paper is organized as follows. In Section 2, we review the method for dealing with the missing response and missing covariates introduced by Chen, Yi and Cook (2009). Data arising from the International Tobacco Control Four Country Survey are analyzed in the application in Section 3. Conclusions are made in Section 4.

[1]Baojiang Chen, Department of Biostatistics, University of Washington, Seattle, Washington, US 98195 (bjchen@u.washington.edu); Mary Thompson, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1 (methomps@uwaterloo.ca)

2. The proposed method

In this section we give a brief introduction of the method of Chen, Yi and Cook (2009) to address the missing response and missing covariates problem in longitudinal data analysis.

2.1 Model formulation

Let Y_{ij} denote the response for subject i at wave j for $j = 1, \ldots, J$, and $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$. Let $\mu_{ij} = E(Y_{ij} | X_i, Z_i)$, where $X_i = (X_{i1}, \ldots, X_{iJ})'$ is the covariate that is subject to missingness, and $Z_i = (Z_{i1}, \ldots, Z_{iJ})'$ is the covariate vector that is always observed. The response model is often given by

$$g(\mu_{ij}) = X_{ij}'\beta_x + Z_{ij}'\beta_z$$

for $j = 1, \ldots, J, i = 1, \ldots, n$, where g is a monotone differentiable link function, and $\beta = (\beta_x', \beta_z')'$ is a $p \times 1$ vector of regression coefficients of interest. Let $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})'$.

For the missing data processes, we let R_{ij}^Y denote the missing indicator of the response Y_{ij}, which is equal to 1 if Y_{ij} is observed and 0 otherwise, and similarly define R_{ij}^X as the missing indicator of the covariate X_{ij}. Let $\lambda_{ij}^Y = P(R_{ij}^Y = 1 | H_{ij}^Y, H_{ij}^X, X_i, Y_i)$ and $\lambda_{ij}^X = P(R_{ij}^X = 1 | H_{ij}^Y, H_{ij}^X, X_i, Y_i)$, where $H_{ij}^Y = (R_{i1}^Y, \ldots, R_{i,j-1}^Y)$ and $H_{ij}^X = (R_{i1}^X, \ldots, R_{i,j-1}^X)$ denote the history of the missing indicators of the response and covariate for subject i, at wave j. Typically, we employ marginal logistic regression models for λ_{ij}^Y and λ_{ij}^X at each assessment time. In particular, we specify

$$\text{logit}(\lambda_{ij}^Y) = u_{ij}'\alpha_y \text{ and } \text{logit}(\lambda_{ij}^X) = v_{ij}'\alpha_x, \quad (1)$$

where u_{ij} and v_{ij} contain functions of $(H_{ij}^Y, H_{ij}^X, Y_i, X_i, Z_i)$, $j = 2, 3, \ldots, J$, and α_y and α_x are regression parameters; let $\alpha_{xy} = (\alpha_y', \alpha_x')'$.

At each time point j, the observation status of the response and covariate may be associated within subjects because of common factors affecting the two observation processes. To model this association we treat the response and covariate observation processes symmetrically and define the conditional odds ratio

$$\psi_{ij} = \frac{P(R_{ij}^Y=1, R_{ij}^X=1 | H_{ij}^Y, H_{ij}^X, X_i, Y_i, Z_i) P(R_{ij}^Y=0, R_{ij}^X=0 | H_{ij}^Y, H_{ij}^X, X_i, Y_i, Z_i)}{P(R_{ij}^Y=1, R_{ij}^X=0 | H_{ij}^Y, H_{ij}^X, X_i, Y_i, Z_i) P(R_{ij}^Y=0, R_{ij}^X=1 | H_{ij}^Y, H_{ij}^X, X_i, Y_i, Z_i)},$$

which is typically modeled as

$$\log(\psi_{ij}) = \phi_j, \quad j = 2, \ldots, J.$$

In this paper, we consider a missing at random mechanism and assume that

$$P(R_{ij}^Y = \eta_j^Y, R_{ij}^X = \eta_j^X | H_{ij}^Y, H_{ij}^X, X_i, Y_i, Z_i) = P(R_{ij}^Y = \eta_j^Y, R_{ij}^X = \eta_j^X | H_{ij}^Y, H_{ij}^X, X_i^{(a)}, Y_i^{(a)}, Z_i) \quad (2)$$

for each time point j, where $X_i^{(a)}$ denotes the observed parts X_i, and $Y_i^{(a)}$ denotes the observed parts of Y_i.

2.2 Estimation and inference

For the estimation of the parameter β, we may employ the augmented inverse probability weighted generalized estimating equations (AIPWGEE). Define $D_i = \partial\mu_i/\partial\beta^t$, $A_i(\alpha) = [w_{ijk}]_{J\times J}$, where $w_{ijk} = I(R_{ij}^Y = 1, R_{ik}^Y = 1, R_{ik}^X = 1)/\pi_{ijk}^{YYX}$, and $\pi_{ijk}^{YYX} = P(R_{ij}^Y = 1, R_{ik}^Y = 1, R_{ik}^X = 1 | X_i, Y_i, Z_i)$. Let $M_i = \kappa V_i^{-1/2}[C_i^{-1} * A_i(\alpha)]V_i^{-1/2}$, where $A * B = [a_{ij}b_{ij}]$ denotes the Hadamard product of $J \times J$ matrices $A = [a_{ij}]$ and $B = [b_{ij}]$, κ is a dispersion parameter, $V_i = diag\{var(Y_{ij} | X_i, Z_i), j = 1, \ldots, J\}$, and C_i is a working correlation matrix. The AIPWGEE is given by

$$\sum_{i=1}^n U_i(\beta, \alpha) = 0, \quad (3)$$

where $U_i(\beta, \alpha) = D_i M_i (Y_i - \mu_i) - \eta A_i(\alpha)$, $A_i(\alpha)$ is an observed data vector that is free of β, η is the regression coefficient of $A_i(\alpha)$ in the population regression of $D_i M_i (Y_i - \mu_i)$ on $(A_i', S_i')'$, and S_i is the score function of α for subject i.

In practice the parameters α of the missing data model are unknown, and one must replace α in (2) with a consistent estimate. The estimation procedure is relatively standard and we do not give details here. Denote the resultant estimator $\hat{\beta}$, and under some regularity conditions, $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean 0 and covariance matrix $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, where $\Gamma = E[\partial U_i/\partial\beta^t]$, and Σ is the covariance matrix of the residuals from the regression of $D_i M_i (Y_i - \mu_i)$ on $(A_i', S_i')'$.

3. ITC four country survey

The ITC Four Country Survey began in 2002, with approximately annual waves. The sample size was maintained at approximately 2000 in each country at each wave, through recruitment of new respondents to replace dropouts. Respondents in the ITC Four Country Survey are smokers aged ≥ 18 years at recruitment; those who stop smoking are retained in the sample.

The response of interest is the intention to quit smoking for current smokers, having smoking status 1, 2 and 3. (People who have smoking status 4, 5, 6, or 7 are current quitters and do not have the intention to quit variable.) Intention to quit is expressed here as a binary variable (1--Yes, 0 -- No). The covariates of interest include gender, age (18-24, 25-39, 40-54, ≥ 55), income (Low - <£ 15000/$30000, Moderate - between £ 15001/$30001 and £ 30000/$59999, High - >£ 30001/$60000), cigarettes per day, time to first cigarette of the day in minutes, and education level (Low, Moderate, High). However, there are missing data for the variables intention to quit and income level. The missing proportions for these variables are listed in Table 3-1 at each wave for each country.

Table 3-1
Missing proportion for the response and covariates for smoking status 1, 2 and 3

Country	Canada		United States		UK		Australia	
	Intend%	Income%	Intend%	Income%	Intend%	Income%	Intend%	Income%
Wave 1	13.78	19.47	16.15	20.64	12.76	20.17	11.28	15.76
Wave 2	5.76	11.94	9.86	13.95	3.87	10.39	3.98	8.25
Wave 3	1.32	6.94	1.60	5.93	1.31	8.92	1.46	6.10
Wave 4	1.52	7.26	2.60	6.63	1.84	9.18	1.05	5.65
Wave 5	1.96	7.25	1.90	5.54	1.89	8.82	1.28	6.56
Wave 6	1.58	7.27	1.84	5.69	1.47	10.05	1.28	6.09
Total	4.92	10.64	6.30	10.40	4.44	11.81	3.74	8.40

The proportions missing the income variable diminish with wave. This can be attributed in part to the fact that at each wave, those who had responded the previous time were asked if there had been any change in income category, while those who had not responded were asked about their income again. The proportions missing the intention to quit variable also decrease with wave, and the reason for this is not clear.

3.1 Statistical models

Here we build some models for the ITC Four Country data for the proposed method. Let Y_{ij} denote the response for subject i at wave $j, j = 1, \ldots, 6$, and $\mu_{ij} = P(Y_{ij} = 1 | X_{ij})$. The response model is given by

$$\text{logit}(\mu_{ij}) = X'_{ij}\beta,$$

where X_{ij} is the covariate vector, including gender, age, square root of cigarettes per day (sqrtcpd), logarithm of the time to first cigarette in minutes (logftime), education level, and income level.

For the missing data models, we specify $\text{logit}(\lambda^y_{ij}) = u'_{ij}\alpha_y$ and $\text{logit}(\lambda^x_{ij}) = v'_{ij}\alpha_x$, where u_{ij} and v_{ij} are the covariates that influence the missingness of response and covariate respectively, which may include a function of previous missing indicators, gender, age, square root of cigarettes per day (sqrtcpd), logarithm of the time to first cigarette in minutes (logftime), education level, and income level. For the associations of the missing indicators, we build models $\log(\psi_{ij}) = \phi_j$, $j = 2, \ldots, 6$.

3.2 Results

Tables 3.2-1 to 3.2-5 list the results for the response and missing data models. Here we give a detailed commentary on the analysis results in Canada. The other three countries are very similar. Tables 3.2-1 and 3.2-2 report the response models and missing data models in Canada.

For each country data set, we compare three methods: the proposed method which accommodates the missingness, the complete case analysis that accommodates the missingness only by modeling the indicator that both the response and covariate are observed, and the complete case analysis that ignores the missingness. We point out that the complete case analysis give biased estimates when data are missing

at random, while the proposed method gives consistent and efficient estimates when the missing data model is correctly specified (Chen, Yi and Cook, 2009).

From Table 3.2-2, it is seen that the missing at random assumption may be more appropriate than a missing completely at random assumption, for both the response and the covariate, since logftime, age, education and income are significant in the missing response model, and gender, logftime, sqrtcpd, age and education are significant in the missing covariate model. In particular, the longer the time until the first cigarette of the day (or the less "addicted" the smoker), the greater the probability of responding. There are serial dependences in the missing indicators for the response and covariates, since r_{j-1}^{y} is moderately significant in the missing response model, and r_{j-1}^{y} and r_{j-1}^{x} are significant in the missing covariate model. Also, there are some associations between the missingness of the response and covariate, since ϕ_1, ϕ_2, ϕ_3, ϕ_4 and ϕ_5 are significant. It seems there is no significant association at wave 6 since ϕ_6 is not significant. Note that the empirical estimate for ϕ_6 is 0.0402, which is very small.

Table 3.2-1 reports the estimates for the response models for unstructured working correlation assumptions in the GEE model using the proposed method, as well as the complete case analysis with and without conventional inverse probability weights for Canada. The two complete case analyses give close results, and they are different from those obtained from the proposed method. Here we comment that as with many bias reduction techniques the standard errors are higher for the proposed method, since the variation for the missing data processes is also involved in the estimation and inferences. For all the methods, it is seen that gender is not significant, indicating that there is no significant difference in intention to quit between male and female. Logftime is not significant in the proposed method, indicating that the time to first cigarette has no significant effect on the intention to quit smoking, while it is significant in the complete case analysis. This difference may be caused by bias inherent in the complete case analysis. Perhaps the dependence of intention to quit on the time to first cigarette is more marked for people who are observed for both the response and income variable. The significance of sqrtcpd in the proposed method indicates that the more cigarettes per day, the less likely an intention to quit, while sqrtcpd is not significant in the complete case analysis. This may also be caused by the bias inherent in the complete case analysis result. Perhaps the dependence of intention to quit on cigarettes per day is more marked for people who are missing either the response or income variable. Age (25-39) and age (40-54) are not significant in the proposed method, indicating that there is no significant difference in intention to quit for people in age group 25 to 54 compared to people in age group 18 to 24, while age (40-54) is significant in the complete case analysis, indicating that the dependence of intention to quit on age group between 40 and 54 is more marked for people with complete data. This also reflects the higher standard errors in the proposed analysis. Age ($\geq$55) is significant in all the methods, indicating that people in age group greater than 55 are less likely to intend to quit smoking compared to people in age group 18--24. Significance of educationM in all the methods indicates that people with moderate education level are more likely to intend to quit compared to people with low level education; however, educationH is not significant in any of the methods, suggesting that there is no significant difference in intention to quit between people with high level and low level education. Significance of incomeH in all the methods indicates that people with high income level are

Table 3.2-1
Response models for ITC4 data

	Canada			United States			UK			Australia		
Parameter	Est	S.e	p	Est	S.e	p	Est	S.e	p	Est	S.e	p
Proposed method:												
Intercept	2.091	0.242	<.001	1.520	0.198	<.001	1.219	0.256	<.001	1.579	0.329	<.001
GenderM	-0.074	0.081	0.357	-0.108	0.072	0.130	-0.130	0.105	0.211	-0.214	0.100	0.032
Logftime	0.026	0.027	0.322	0.118	0.024	<.001	0.031	0.029	0.262	0.053	0.030	0.074
sqrtcpd	-0.183	0.033	<.001	-0.119	0.029	<.001	-0.123	0.047	0.008	-0.057	0.032	0.072
Age(25-39)	0.217	0.184	0.238	-0.260	0.140	0.063	-0.064	0.153	0.671	-0.012	0.272	0.964
Age(40-54)	-0.200	0.180	0.264	-0.461	0.137	0.001	-0.484	0.145	0.001	-0.669	0.275	0.015
Age(≥55)	-0.555	0.185	0.002	-1.107	0.142	<.001	-1.079	0.191	<.001	-0.990	0.283	0.001
EducationM	0.196	0.089	0.027	0.178	0.080	0.025	0.468	0.147	0.001	0.143	0.127	0.257
EducationH	0.088	0.122	0.469	0.256	0.099	0.009	0.570	0.125	<.001	0.406	0.142	0.004
IncomeM	0.127	0.085	0.135	0.159	0.077	0.038	0.206	0.113	0.067	0.187	0.104	0.071
IncomeH	0.237	0.100	0.017	0.189	0.086	0.028	0.145	0.152	0.340	0.053	0.126	0.669
Complete case weighted GEE:												
Intercept	1.251	0.132	<.001	1.397	0.124	<.001	1.057	0.144	<.001	1.558	0.129	<.001
GenderM	-0.092	0.069	0.179	-0.132	0.058	0.022	-0.212	0.060	<.001	-0.143	0.066	0.029
Logftime	0.067	0.015	<.001	0.058	0.013	<.001	0.021	0.015	0.143	0.042	0.014	0.002
sqrtcpd	-0.009	0.100	0.332	-0.051	0.009	<.001	-0.042	0.009	<.001	-0.019	0.008	0.024
Age(25-39)	0.162	0.126	0.198	-0.248	0.120	0.038	-0.035	0.132	0.788	-0.200	0.121	0.097
Age(40-54)	-0.265	0.121	0.027	-0.488	0.116	<.001	-0.609	0.130	<.001	-0.821	0.118	<.001
Age(≥55)	-0.672	0.128	<.001	-1.011	0.119	<.001	-1.073	0.134	<.001	-1.194	0.129	<.001
EducationM	0.250	0.074	0.001	0.220	0.064	0.001	0.292	0.071	<.001	0.196	0.078	0.011
EducationH	0.088	0.102	0.389	0.225	0.088	0.010	0.311	0.094	0.001	0.329	0.100	0.001
IncomeM	0.074	0.074	0.313	0.113	0.064	0.077	0.155	0.067	0.019	0.200	0.074	0.006
IncomeH	0.231	0.087	0.007	0.115	0.077	0.132	0.296	0.076	<.001	0.153	0.080	0.056
Complete case GEE:												
Intercept	1.243	0.130	<.001	1.391	0.124	<.001	1.037	0.145	<.001	1.545	0.130	<.001
GenderM	-0.091	0.068	0.181	-0.134	0.058	0.021	-0.193	0.060	0.001	-0.143	0.066	0.029
Logftime	0.065	0.015	<.001	0.060	0.013	<.001	0.023	0.015	0.107	0.042	0.014	0.002
sqrtcpd	-0.007	0.100	0.426	-0.052	0.009	<.001	-0.042	0.009	<.001	-0.020	0.008	0.017
Age(25-39)	0.159	0.126	0.207	-0.247	0.120	0.039	-0.026	0.133	0.840	-0.195	0.122	0.110
Age(40-54)	-0.268	0.121	0.026	-0.485	0.116	<.001	-0.612	0.130	<.001	-0.814	0.120	<.001
Age(≥55)	-0.688	0.128	<.001	-1.016	0.119	<.001	-1.077	0.135	<.001	-1.188	0.130	<.001
EducationM	0.259	0.074	0.000	0.223	0.064	0.001	0.290	0.071	<.001	0.191	0.078	0.013
EducationH	0.097	0.102	0.338	0.226	0.087	0.009	0.318	0.094	0.001	0.331	0.099	0.001
IncomeM	0.082	0.073	0.262	0.113	0.064	0.075	0.149	0.067	0.025	0.208	0.074	0.004
IncomeH	0.228	0.086	0.007	0.109	0.077	0.153	0.286	0.076	<.001	0.163	0.080	0.041

more likely to intend to quit than people with low income level, but incomeM is not significant, indicating that there is no significant difference for people with moderate level income and low level income in the intention to quit.

Table 3.2-2
Missing data models for ITC4 data: Canada

Parameter	Estimate	S.e	p	Estimate	S.e	p
Intercept	2.3144	0.2140	<.0001	0.7122	0.1413	<.0001
GenderM	0.0657	0.1217	0.5893	0.1897	0.0890	0.0330
Logftime	0.4291	0.0125	<.0001	0.3415	0.0114	<.0001
sqrtcpd	0.0063	0.0176	0.7194	0.0357	0.0127	0.0049
Age(25-39)	0.0086	0.2233	0.9691	0.3949	0.1498	0.0084
Age(40-54)	-0.0279	0.2194	0.8989	0.5171	0.1478	0.0005
Age(≥55)	-0.4601	0.2301	0.0456	-0.1443	0.1521	0.3426
EducationM	0.2516	0.1343	0.0610	0.1581	0.0958	0.0989
EducationH	0.2992	0.1939	0.1228	0.5916	0.1523	0.0001
r_{j-1}^{x}	0.3850	0.3283	0.2410	7.9643	0.2607	<.0001
r_{j-1}^{x}:IncomeM	0.5499	0.1614	0.0007			
r_{j-1}^{x}:IncomeH	-0.3822	0.1591	0.0163			
r_{j-1}^{y}	0.5883	0.3188	0.0650	-4.6311	0.1757	<.0001
Association						
ϕ_1	4.9701	1.2835	0.0001			
ϕ_2	3.5034	1.0584	0.0009			
ϕ_3	1.7849	0.8936	0.0458			
ϕ_4	2.5133	0.9230	0.0065			
ϕ_5	1.3455	0.6810	0.0482			
ϕ_6	0.8465	0.5096	0.0967			

Table 3.2-3
Missing data models for ITC4 data: United States

Parameter	Estimate	S.e	p	Estimate	S.e	p
Intercept	2.1944	0.1939	<.0001	0.9384	0.1422	<.0001
GenderM	-0.1029	0.1133	0.3638	-0.0597	0.0885	0.4997
Logftime	0.4367	0.0112	<.0001	0.3705	0.0106	<.0001
sqrtcpd	0.0122	0.0166	0.4609	0.0483	0.0124	0.0001
Age(25-39)	0.0275	0.1972	0.8893	0.4901	0.1517	0.0012
Age(40-54)	0.3348	0.1955	0.0867	0.4787	0.1467	0.0011
Age(≥55)	-0.0567	0.2050	0.7822	0.1914	0.1527	0.2102
EducationM	0.0755	0.1213	0.5339	0.0889	0.0950	0.3491
EducationH	0.1923	0.1787	0.2820	0.1821	0.1362	0.1813
r_{j-1}^{x}	-1.3784	0.3351	<.0001	7.9744	0.2903	<.0001
r_{j-1}^{x}:IncomeM	-0.0751	0.1867	0.6874			
r_{j-1}^{x}:IncomeH	0.0902	0.2021	0.6556			
r_{j-1}^{y}	2.4721	0.3347	<.0001	-4.7668	0.2047	<.0001
Association						
ϕ_1	5.0828	1.4512	0.0005			
ϕ_2	4.2512	1.0868	0.0001			
ϕ_3	3.4018	1.6228	0.0361			
ϕ_4	2.9335	1.0354	0.0046			
ϕ_5	0.7749	0.5244	0.1395			
ϕ_6	-0.3517	0.5328	0.5092			

Table 3.2-4
Missing data models for ITC4 data: UK

Parameter	Estimate	S.e	p	Estimate	S.e	p
Intercept	2.2744	0.2658	<.0001	0.4124	0.1688	0.0146
GenderM	0.0795	0.1298	0.5403	0.2819	0.0871	0.0012
Logftime	0.4240	0.0125	<.0001	0.3146	0.0116	<.0001
sqrtcpd	-0.0125	0.0243	0.6084	0.0175	0.0163	0.2833
Age(25-39)	0.2152	0.2591	0.4063	0.7454	0.1685	<.0001
Age(40-54)	0.2305	0.2599	0.3753	0.6626	0.1668	0.0001
Age(≥55)	0.0493	0.2632	0.8515	0.0619	0.1659	0.7092
EducationM	-0.1635	0.1541	0.2887	0.2356	0.1061	0.0264
EducationH	0.1443	0.2218	0.5151	0.3588	0.1515	0.0179
r_{j-1}^{x}	0.2980	0.3720	0.4232	9.0376	0.3205	<.0001
r_{j-1}^{x}:IncomeM	0.4139	0.1817	0.0227			
r_{j-1}^{x}:IncomeH	-0.1813	0.1805	0.3150			
r_{j-1}^{y}	0.8882	0.3605	0.0138	-5.5074	0.2544	<.0001
Association						
ϕ_1	5.0221	1.4043	0.0003			
ϕ_2	3.3933	1.3434	0.0115			
ϕ_3	2.5297	1.2118	0.0368			
ϕ_4	2.7053	1.1533	0.0190			
ϕ_5	0.4146	0.7030	0.5553			
ϕ_6	1.0570	0.8856	0.2327			

Table 3.2-5
Missing data models for ITC4 data: Australia

Parameter	Estimate	S.e	p	Estimate	S.e	p
Intercept	2.3486	0.1975	<.0001	0.7787	0.1357	<.0001
GenderM	-0.0340	0.1296	0.7930	0.1463	0.0964	0.1291
Logftime	0.3953	0.0123	<.0001	0.3106	0.0113	<.0001
sqrtcpd	-0.0282	0.0208	0.1748	0.0316	0.0143	0.0270
Age(25-39)	0.4498	0.1993	0.0240	0.7526	0.1431	<.0001
Age(40-54)	0.4592	0.2040	0.0244	0.7949	0.1466	<.0001
Age(≥55)	0.3425	0.2298	0.1361	0.2426	0.1570	0.1223
EducationM	-0.1353	0.1596	0.3964	0.0415	0.1199	0.7296
EducationH	-0.2087	0.1932	0.2802	0.2882	0.1564	0.0653
r_{j-1}^{x}	0.4696	0.3765	0.2124	8.0150	0.2688	<.0001
r_{j-1}^{x}:IncomeM	-0.0010	0.1792	0.9954			
r_{j-1}^{x}:IncomeH	0.0286	0.1808	0.8745			
r_{j-1}^{y}	0.6997	0.3645	0.0549	-4.7754	0.1876	<.0001
Association						
ϕ_1	5.1733	1.5114	0.0006			
ϕ_2	3.2662	1.4836	0.0277			
ϕ_3	1.7442	1.1269	0.1217			
ϕ_4	1.4635	0.7352	0.0465			
ϕ_5	1.2803	0.4740	0.0069			
ϕ_6	1.3352	0.8523	0.1172			

4. Conclusions

In this paper, we apply the augmented inverse probability weighted generalized estimating equations to the ITC Four Country Survey data by assuming the data are missing at random. This approach involves modeling the missing data process and weighting the estimating equations by the inverse of a probability that is calculated based on the models for the missing data processes. The results in the missing data analyses suggest that the missing at random may be reasonable, so this method can correct for bias that the
complete case analysis does not. For example, in this illustration, the important common factor in missingness of both might be degree of addiction as measured by (the inverse of) logftime; for example, the more addicted smokers may be more likely not to respond to both because they are becoming less patient toward the end of the interview. We can reduce bias due to dropout by appropriate weighting. Only for the
US does the proposed method sharpen the conclusion about dependence of intention to quit on income. However, in all countries the estimation of dependence of intention to quit on sqrtcpd shows a bigger effect size and may be better because we are in effect bringing back in more highly addicted dropouts.

Acknowledgements

This work is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The ITC project is supported by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897).

References

Centers for Disease Control and Prevention. (2002). Cigarette smoking among adults-- United States, 2002, Morbidity and Mortality Weekly Report, 51, pp. 642--645.

Chen, B., Yi, G. Y. and Cook, R. J. (2009). Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data which are Missing at Random, *Journal the American Statistical Association*, in press.

Health Canada, Tobacco Control Programme. (2002). *Canadian Tobacco Use Monitoring Survey*. February-December 2001. Ottawa, Canada.

Hyland, A., Li, Q., Bauer, J. Giovino, G., Steger, C. and Cummings, K. M. (2004). Predictors of cessation in a cohort of current and former smokers followed over 13 years, *Nicotine & Tobacco Research*, 6, pp. S363--S369.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, 73, pp. 13--22.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc., 1st ed.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, 89, 846--866.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, 90, 106--121.

FACTORS AND IMPACTS OF NON-RESPONSE

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Factors Associated with Different Patterns of Non-Response in English Longitudinal Study of Ageing (ELSA)

Hayley Cheshire and David Hussey [1]

Abstract

Understanding the factors associated with attrition is of key relevance for those managing longitudinal studies. If analysts can identify those groups most likely to stay in or drop out of the study, the survey design process can be tailored (e.g. through fieldwork practices) to maximise likely response.

A review of the current literature relating to people aged 55 and over (Bhamra et al, 2008) has identified some factors related to attrition – for example being older, cognitively impaired, having lower socio-economic status, and being less well educated.

We propose to study the factors associated with different patterns of participation from wave 1 to wave 3 of the English Longitudinal Study of Ageing (ELSA). ELSA is a study of people aged 50 and over and their younger partners. A total of 12,100 individuals were included at baseline and subsequently followed up as part of the study.

The following comparison groups will be used:
- Completed interviews at all waves
- Dropped out at wave 2, but came back at wave 3
- Dropped out at wave 2
- Dropped out at wave 3

Our analyses will expand the current focus on demographic factors to include survey variables which can be used to provide some indication of level of commitment to the study at wave 1. Of key interest are factors relating to response behaviour, for example – the use of extremes or mid-points on answer scales, item non-response, willingness to consult documents during interview, and consent to linkage of government administrative data.

[1] Hayley Cheshire (Hayley.Cheshire@natcen.ac.uk) and David Hussey (David.Hussey@natcen.ac.uk), National Centre for Social Research, U.K.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Empirical Investigation of Nonresponse Bias Due to Attrition in National Survey of College Graduates (NSCG)

Donsig Jang, John Finamore, David Hall, Steve Cohen, Flora Lan and Fan Zhang[1]

Abstract

The NSCG is a biennial survey whose main goal is to produce estimates representing the target population of U.S. scientists and engineers at a fixed reference date. The Decennial Census long form is used as its sampling frame. Because the complete sampling frame is available only once a decade, NSCG uses the data for several rounds of its survey with periodic supplemental samples of new graduates in sciences and engineering fields.

In a longitudinal survey like NSCG, follow-up surveys should include nonrespondents in its sample to minimize bias. However, most initial nonrespondents become persistent refusals and, therefore, difficult to gain cooperation in the next round. For this reason, the NSCG sample had been followed in three subsequent rounds only if individuals continued to respond to the survey. However, it is anticipated that the inclusion of only respondents in the next round of NSCG would cause a substantial survey bias even if customary nonresponse weighting adjustments were made. To understand the nonresponse bias due to the longitudinal nonrespondents, we compared estimates from the sample based on 1990s Decennial long form respondents collected in 2003 to those from a new sample drawn from Census 2000 initially surveyed in 2003. In this paper, we will present results from this investigation that will show empirical insights about attrition effects on survey bias.

[1] Donsig Jang, Mathematica Policy Research, U.S.A. (DJang@Mathematica-Mpr.com); John Finamore (john.m.finamore@census.gov) and David Hall, U.S. Census Bureau, U.S.A.; Steve Cohen, Flora Lan and Fan Zhang, National Science Foundation, U.S.A.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Factors associated with participation in the GAZEL cohort

Marie Zins, Jean François Chastang, Mireille Coeuret-Pellicer, Annette Leclerc, Sébastien Bonenfant, Alice Guéguen, Anna Ozguler and Marcel Goldberg[1]

Abstract

Background: The GAZEL cohort was selected in 1989 from Électricité de France-Gaz de France employees aged 35 to 50. A postal questionnaire was used for selection, and 20,625 subjects (15,011 men and 5,614 women) agreed to take part. Follow-up consisted of an annual postal questionnaire and an invitation to go to a health examination centre for a medical check-up. The initial participation rate was 44.5%. Each year about 75% of the subjects mailed back the questionnaire, and 44.7% visited a health examination centre.

Objectives: Study the socio-demographic, behavioural, occupational and health-related factors associated with the effects that being selected had on participation in the annual follow-up and the health examination, and quantify their role.

Methods: On selection, volunteers were compared with non-participants using variables collected systematically from the company's medical administrative databases (absenteeism, mortality, occupational exposure) in logistic regression models. Mixed models were used to study the probability of responding to the annual questionnaires during follow-up, while logistic regression models were used to study the probability of participation in the health examination.

Results: The various steps – selection, follow-up and visit to a health centre – are not always affected by the same factors. The magnitude of the selection effects varies from step to step. This study describes the potential biases that this can generate.

[1] Marie Zins (marie.zins@inserm.fr), Jean François Chastang, Mireille Coeuret-Pellicer, Annette Leclerc, Sébastien Bonenfant, Alice Guéguen (alice.guéguen@inserm.fr), Anna Ozguler and Marcel Goldberg, INSERM, France

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Strategies for studying non-response bias in the Coset (Cohorte santé et travail) and Constances (Cohorte des consultants des centres d'examens de santé) cohort

Gaëlle Santin[1], Laetitia Bénézet[1], Alice Guéguen[2], Rémi Sitta[2,3], Stéphanie Gauvin[1], Hélène Sarter[1], Nicolas Razafindratsima[4], Marie Zins[2,3], Béatrice Geoffroy-Perez[1], Marcel Goldberg[1,2]

Abstract

The Coset program and the Constances cohort are longitudinal surveys currently in the pilot phase. One of their objectives is to describe, and track changes in, the health of workers by occupation. The expected participation rate on inclusion is unlikely to exceed 30%. To adjust for non-response biases, the strategy to be adopted is based on the use of auxiliary information from existing files on participants' and non-participants' health care claim payments, hospitalizations and employment history. For purposes of discussing this strategy, a complementary survey will be conducted on a sample of non-participants. These strategies will be described using the pilot phase of the farm workers cohort of the Coset program.

Key words: Longitudinal survey, cohort; non-response, non-response bias.

1. Introduction

Non-response can lead to recurring problems in surveys (biased estimates, inflated variance), and such problems will arise in the Coset program (cohorts for workplace epidemiological monitoring) and the Constances cohort (health examination centre consultants). One of the purposes of these cohorts, currently in their pilot phase, is to describe, and track changes in, the health status of the population currently or formerly in the labour force, based on occupational activity among other factors. In studying the epidemiology of occupational risks, it is essential to use cohorts, since they make it possible to take account of lifetime occupational exposures. This is important because for some pathologies such as cancer, there is a long latency period (sometimes extending over several decades) between exposure to risk and the onset of the disease (Goldberg, 2008). The expected participation rate in the Coset program and the Constances cohort is unlikely to exceed 20% to 30%.

This article will describe the Coset program, with particular emphasis on the pilot phase of the Coset farm workers cohort, which will begin in the first half of 2010 (the Constances cohort has been described elsewhere in this symposium (Gueguen, 2009)). During the pilot phase, the following two methods will be implemented to minimize non-response biases in the estimates on inclusion:

- First, the use of administrative files containing auxiliary information, available for participants and non-participants alike. This information, relating to health care claim payments, hospitalizations and employment history, is not subject to the limitations associated with information provided on a voluntary basis, and it will be used to adjust the weights resulting from the sampling plan;
- Second, the use of double sampling, by means of conducting a complementary survey of non-respondents.

The estimates obtained after adjustment for non-response by reweighting and those obtained from double sampling will be compared. This will provide a basis for discussing from several perspectives the biases related to non-response at the time of inclusion, with a view to extending the Coset farm workers cohort and the Constances cohort nationwide. Since the pilot phase is currently under way, no results will be presented.

[1]Gaëlle Santin, Laetitia Bénézet, Stéphanie Gauvin, Hélène Sarter, Béatrice Geoffroy-Perez, Marcel Goldberg, Institut de veille sanitaire, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex, France
[2]Alice Guéguen, Rémi Sitta, Marie Zins, Marcel Goldberg, Unité mixte Inserm-Cnamts 687/ IFR69, 16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France
[3]Rémi Sitta, Marie Zins, Équipe RPP-C du Cetaf, 16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France
[4]Nicolas Razafindratsima, Institut national d'études démographiques, 133 boulevard Davout 75980 Paris Cedex 20, France

2. General description of the Coset program

2.1 Objectives

The main objectives of the Coset program are to describe at time "t" and to track over time the morbidity/mortality of persons in the labour force according to occupational activity (socio-occupational category and industry) and, looking at the links between morbidity/mortality of persons in the labour force and occupational exposures, to describe how these links evolve over time.
The Coset program should therefore serve to describe the distribution of morbidity or mortality and frequencies of exposure to various health hazards at the scale of the labour force in France, that is, to produce both unbiased quantitative estimates (incidence rates, prevalences, etc.) for the labour force in France at the time of inclusion (longitudinal representativeness) and unbiased quantitative estimates for the same population in a given year (cross-sectional representativeness).

2.2 General method: Study of data from three cohorts

Data in the Coset program will be collected from persons in the labour force who are covered by the three main social protection systems, which together extend to 95% of the working population in France. They will therefore come from three cohorts:

- for the wage-earning labour force, covered by the *Régime général de sécurité sociale* (RGSS), representing roughly 80% of labour force participants in France, the data will come from the Constances cohort (Gueguen, 2009). This is a general cohort consisting of persons covered by the RGSS, and it is currently being created by the *Institut national de la santé et de la recherche médicale* (Inserm). This cohort serves a number of objectives and contributes to many fields of research; among other things, it is used to collect information needed to monitor occupational risks. That information will be used in the Coset program; ultimately, the Constances cohort will include approximately 200,000 persons;
- a cohort, currently in its pilot phase, comprised of the farm worker labour force covered by *Mutualité sociale agricole* (MSA), and a cohort, currently at the discussion stage, comprised of the work force of self-employed workers covered by the *Régime social des indépendants* (RSI). These two cohorts will be created by the *Institut de veille sanitaire* in partnership with the social protection systems concerned. Ultimately, each of them should include 35,000 persons.

Insofar as possible, the methods used for the three cohorts will be similar.

2.3 Pilot phase of the Coset farm workers cohort

For the pilot phase of the Coset farm workers cohort, the persons approached to participate will be drawn randomly from among wage-earning and self-employed farm workers served by one of the five pilot departmental offices of the MSA, aged 18 to 65 and having been insured for at least 90 days in 2008, whatever their type of activity. The sampling frame used will be the retirement insurance frame of MSA members. Ten thousand persons will be selected randomly, numbering 2,000 per departmental office. For each office, the random sample, constituted by simple random sampling, will be stratified by sex, age and employment status (wage earner or self-employed) and will be proportional to the size of the strata.
Data will be collected using a self-administered mail questionnaire. The information collected will concern health status (muscle and joint problems, mental health problems, cardio-vascular and respiratory problems, cancer, etc.), current and past occupational activity (status, type of contract, work time) and exposures to various past or present workplace hazards (organizational and psychosocial constraints, pain, noise, hazards of chemical, physical or biological origin). A reminder will be mailed out one month after the initial mailing. Respondents will be followed up on an annual basis: they will be mailed a short questionnaire focusing on their health status and any new occupational events that occurred during the year.

3. Treatment of total non-response in the pilot phase

The expected response rate for this pilot phase is between 20% and 30%. Therefore, some estimates will probably be affected by biases related to non-response. This part describes the strategies for compensating for non-response in the pilot phase of the Coset farm workers cohort.

3.1. Adjustment for total non-response by reweighting

To obtain unbiased estimates, there are plans to adjust the sampling weights for non-response, since non-participants will probably differ from the participants and these differences may be related to the variables of interest (health status, employment history). The adjustment can be made by forming homogeneous response groups, using the score method (Eltinge, 1997), or by generalized calibration (Sautory, 2003).

- On which variables should an adjustment be made for non-response?

The study of selection biases is an integral part of epidemiology, since this field is continually faced with non-response in its surveys (Goldberg, 2001b). It has been shown that participation in an epidemiology survey is related to the person's age, social category and health status (Oleske, 2007) and to behaviours that pose a health risk, such as alcohol and tobacco consumption (Goldberg, 2001a). Since all these variables are also related to the Coset variables of interest, it is essential to take them into account when treating non-response.

- What are the available data sources?

During the inclusion phase of the pilot study, there are plans to access many items of auxiliary information, available in the databases of the MSA and the file of the *Système national d'information inter-régime de l'assurance maladie* (SNIIR-AM). This information will relate firstly to the health of the individuals randomly selected (payment of claims for medications and consultations, hospitalizations, work accidents and occupational illnesses) and secondly to their occupational activity (employment status, work time). Sociodemographic information such as age and social category will also be available from these same files.
These numerous items of auxiliary information will be available for both participants and non-participants. Moreover, at the time of inclusion, this information will be available going back at least two years. During follow-up of the cohort, the information will be collected annually. Considering its nature and diversity, this information should make it possible to achieve a good adjustment for non-response to minimize biases in the estimates and, insofar as possible, to reduce non-response to an ignorable phenomenon.

However, the use of auxiliary information for purposes of obtaining unbiased estimates is open to discussion, especially since the expected response rate is low. So as to be able to discuss this method, a complementary survey of a sample of non-participants will be conducted during the pilot selection phase.

3.2. Complementary survey of a sample of non-participants

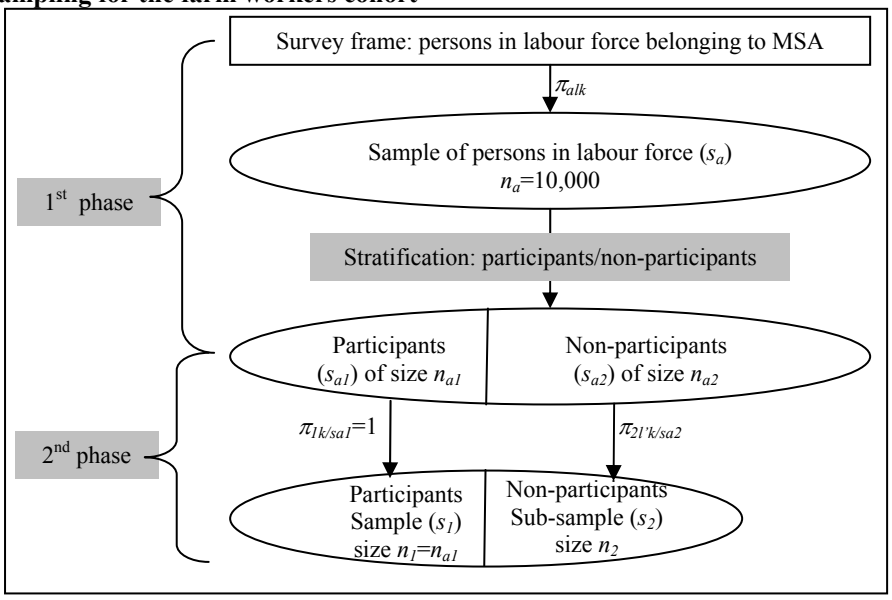
3.2.1 General principle

Hansen and Hurwitz (Hansen, 1946) were the first to conduct a complementary survey of non-participants in order to obtain unbiased estimates of means and totals, by adopting the framework of double sampling (or two-stage sampling).
Applied to the Coset farm workers cohort, the principle is the following (Särndal, 1992) (*cf.* Figure 3.2.1-1):
A sample (s_a) of 10,000 members of the labour force was randomly selected from the MSA retirement insurance frame, with, for individual k, a probability of inclusion π_{alk} (with l corresponding to the stratum: departmental MSA office x sex x age class x status). On completion of data collection by means of a mail questionnaire, the sample (s_a) consists of a sample of participants (s_{a1}) and a sample of non-participants (s_{a2}); it may therefore be partitioned into a "participants" stratum and a "non-participants" stratum. A second-stage sample is selected with the following probabilities of inclusion:

- If individual k belongs to (s_{a1}), he will have a probability of inclusion $\pi_{1k/sa1}$ equal to 1. The sample obtained (s_1) corresponds exactly to (s_{a1}) and its size is equal to the number of participants in the initial survey.
- If individual k belongs to (s_{a2}), he will have a probability of inclusion $\pi_{2l'k/sa2}$ (with l' corresponding to the stratum to which the individual belongs, consisting of the department). This yields a sample (s_2) of size n_2. The complementary survey will then be conducted using sample (s_2).

The sufficient condition for obtaining unbiased estimates is to have a response rate close to 100% for sample (s_2). Accordingly, the operational protocol for the complementary survey of non-participants calls for collection methods that maximize the response rate (short questionnaire; telephone or face-to-face survey).

Figure 3.2.1-1
Principle of double sampling for the farm workers cohort



3.2.2 Sampling for the complementary survey

The target population for the random sampling in the complementary survey consists of all individuals who did not participate in the initial mail survey of the Coset farm workers cohort. Subjects will be selected by stratified random sampling, based on the subject's departmental MSA office. The size of the sample was determined according to a principle of parsimony between collection costs and participant numbers, which must be large enough to obtain reliable estimates.
A sample of 100 non-participants per departmental office will be selected randomly, and the persons sampled will be randomly divided into two groups that will be subject to different collection methods: 70 will be interviewed primarily by telephone, while 30 will be interviewed primarily face to face. In all, 500 persons will thus be sampled.

3.2.3 Data collection

3.2.3.1 Data collected

The questionnaire for the complementary survey is based on the initial questionnaire; however, to maximize the response rate, it has been considerably reduced and the length of the interview will not exceed ten minutes. Only the main variables of interest have been retained (overall health status and employment status) and the variables considered in advance to be associated with non-response, such as sex, age, marital status, behaviours posing a risk for health (alcohol, tobacco), overall health status, education level, employment status, number of occupational episodes in work history, occupation and main occupational exposures.
Because data related to the collection process (paradata) are useful for understanding non-response in the complementary survey (Beaumont, 2005), they will also be collected. These include number of contacts, dates and times of contact and duration of interview.

3.2.3.2 Data collection methods

Data collection methods were chosen based on their capacity to obtain a maximum response rate. For this reason, at the outset a telephone survey was chosen rather than a mail survey. Since telephone numbers are not available in the MSA frames, it is likely that a certain percentage of telephone numbers will not be found, because at present there is no comprehensive telephone directory in France. It was therefore decided that persons not reached by telephone would be interviewed face to face. However, to be able to study the potential biases related to the different data collection methods (by telephone and face-to-face), it was ultimately decided that one group would be earmarked to have a face-to-face interview rather than a telephone interview.
Interviewers will capture responses directly, using the CATI (computer assisted telephone interview) and CAPI (computer assisted personal interview) methods.
To maximize the chances of contacting people and because the survey population consists of persons in the labour force, calls and visits will be mainly made during time blocks when people are most likely to be at home (in the evening or on Saturdays).

Interviewers will make up to 15 call attempts for telephone interviews and up to 3 visits for face-to-face interviews, on different days and at different times.

3.3 Data analyses

In the end, for a parameter of interest Y , several estimates will be calculated and compared:

- $\hat{Y}^{s_{a1}}$ obtained from sample (s_{a1}) ;
- $\hat{Y}^{s_{a1}Aj}$ obtained from sample (s_{a1}) after adjustment for non-response by reweighting;
- $\hat{Y}^{s_2}$ obtained from sample (s_2) ;
- $\hat{Y}^{ED}$ obtained from double sampling ($s_1 \cup s_2$).

To get an idea of the magnitude of the bias associated with non-participation and to grasp the effect of adjustment, estimates $\hat{Y}^{s_{a1}}$ and $\hat{Y}^{s_2}$, $\hat{Y}^{s_{a1}}$ and $\hat{Y}^{ED}$, $\hat{Y}^{s_{a1}}$ and $\hat{Y}^{s_{a1}Aj}$, $\hat{Y}^{s_{a1}Aj}$ and $\hat{Y}^{ED}$ can be compared. Initially, it may be considered that two estimates are not too different if their confidence intervals overlap; other, more elaborate criteria for comparison may also be chosen subsequently.

If the response rate for the complementary survey is close to 100% and there is no other source of bias than non-participation (e.g., measurement error bias), the estimate $\hat{Y}^{ED}$ will be approximately unbiased. By comparing these estimates, it will be possible to gain a better understanding of biases related to non-participation, and to discuss the quality of the auxiliary information contained in informational databases for purposes of correcting the non-participation bias.

However, it is unrealistic to expect to obtain a response rate of close to 100% for the complementary survey. Nevertheless, considering the data collection procedures employed in this survey, it seems reasonable to expect a response rate of around 70%. The estimate $\hat{Y}^{ED}$ can then be corrected for non-participation, based on the auxiliary information contained in the informational databases, to obtain an estimate $\hat{Y}^{EDAj}$ and compare it to $\hat{Y}^{s_{a1}Aj}$. If these two estimates are close, it can be considered that the use of auxiliary information yields estimates equivalent to those that would have been obtained if one had sought to maximize the participation rate in the initial survey. If they differ, several hypotheses will have to be discussed, such as the insufficiency of the auxiliary information to correct for non-participation or the existence of measurement biases in the complementary survey (different collection procedures, less specific responses of persons interviewed in the complementary survey).

4. Discussion and prospects

In the pilot phase of the Coset farm workers cohort, a comparison of the estimates obtained after adjustment for non-response by reweighting and those obtained from the double sampling will be conducted for the variables common to the initial survey and the complementary survey.

However, this approach has several limitations. First, this comparison cannot be made for all variables of interest, since the complementary survey is limited to certain questions. Also, because different collection methods are used for the initial survey and the complementary survey, other forms of bias could result. Moreover, if the auxiliary information used in adjusting for non-response by reweighting is not appropriate, this might cause an increase in the variances for the estimates $\hat{Y}^{s_{a1}Aj}$, especially since the expected participation rate for the initial survey is low. Also, the variances of estimates $\hat{Y}^{ED}$ and $\hat{Y}^{s_2}$ are likely to be high, owing to the small number of participants in the complementary survey. This may pose a problem, since the comparison of the estimates is to be based on the confidence intervals.

Nevertheless, this approach has several advantages. The auxiliary information contains numerous items related to the variables of interest (health status, employment history) and participation, and this information is comprehensive, is not volunteer-based and is available going back at least two years. For its part, the complementary survey can be conducted in such a way as to maximize the response rate. This is so because it will be based on a limited sample, the address frame is of very good quality and the partnership with MSA, which has a high profile among its members, promises good co-operation on the part of the persons approached.

Lastly, by comparing these strategies, it will be possible to have a comprehensive approach for both understanding and compensating for non-participation in the Coset farm workers cohort.

The results are expected during 2012, and they will be used to adjust the implementation procedures for the Coset farm workers cohort when it is extended nationwide, planned for 2012.

Acknowledgments

The authors wish to thank the head office of Mutualité sociale agricole for its involvement and its active participation in establishing the Coset farm workers cohort and in collecting information from its members.

References

Beaumont, J-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31, 227-231.

Eltinge, J. and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells, with an application to income nonresponse in the US Consumer Expenditure Survey, *Survey Methodology*, 23, pp. 33-40.

Goldberg, M., Chastang, J.F., Leclerc, A., Zins, M., Bonenfant, S., Bugel, I., Kaniewski, N., Schmaus, A., Niedhammer, I., Piciotti, M., Chevalier, A., Godard, C. and Imbernon, E. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population, *Am J Epidemiol,* 154, p. 373-84.

Goldberg, M. and Luce D. (2001). Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et consequences *Rev Epidemiol Santé Publ*, 49, p. 477-492.

Goldberg, M., and Imbernon, E. (2008). Quels dispositifs épidémiologiques d'observation de la santé en relation avec le travail?, *RFAS*, 2-3, p.21-44.

Gueguen, A., Sitta, R., Bénézet, L., Santin, G., Lanoe, JL., Goldberg, M. and Zins M. (2009). Contribution of administrative and medical administrative databases to the Constances cohort, *Proceedings :Symposium 2009, Longitudinal Surveys: from Design to Analysis,* Statistics Canada.

Hansen, M. and Hurwitz, W. (1946). The problem of nonresponse in sample surveys, *JASA,* 41, p.517-29.

Oleske, D.M., Kwasny, M.M., Lavender, S.A. and Andersson, G.B. (2007). Participation in occupational health longitudinal studies : predictors of missed visits and dropouts, *Ann Epidemiol*, 17, p. 9-18.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Sautory, O. (2003). Calmar 2: A new version of the Calmar calibration adjustment program, *Proceedings: Symposium 2003, Challenges in Survey Taking for the Next Decadee,* Statistics Canada.

GENERAL METHODOLOGICAL ISSUES

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Sample Allocation for the 2010 Decade of the National Survey of College Graduates

John Finamore, David Hall, Donsig Jang, Stephen Cohen, Flora Lan and Fan Zhang[1]

Abstract

The National Survey of College Graduates (NSCG) is a biennial longitudinal survey that derived its current sample from the 2000 decennial census long form. With the American Community Survey (ACS) replacing the long form, we are planning to use the ACS as a sampling frame for the 2010 decade of the NSCG. In the planning phase, we examined NSCG design options for the 2010 decade and decided on a rotating panel design. To transition into this rotating panel design, part of the 2010 NSCG sample will be selected from the ACS frame and part will be carried forward from the 2000 decade NSCG sample. In subsequent survey cycles, the ACS-based sample will be carried forward, but the 2000 decade cases will rotate out and will be replaced by sample from more recent ACS years.

The 2010 decade of the NSCG will be designed to use the ACS sampling frame to provide statistically reliable estimates for the key NSCG analytical domains. Based on current funding, the 2010 and 2012 NSCG will select approximately 130,000 sample cases from the ACS-based frame over the course of two NSCG survey cycles. This paper will discuss our research to identify the NSCG key analytical domains, establish the reliability thresholds for these NSCG domains, and develop an algorithm to determine the sample allocation under the reliability thresholds. The sample allocation algorithm will initially focus on the 2012 NSCG design (the first complete ACS-based design), but will also investigate allocating the NSCG sample in 2010, 2014, and beyond.

[1] John Finamore (john.m.finamore@census.gov) and David Hall, U.S. Census Bureau, U.S.A.; Donsig Jang, Mathematica Policy Research, U.S.A. (DJang@Mathematica-Mpr.com) ; Stephen Cohen, Flora Lan and Fan Zhang, National Science Foundation, U.S.A.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Life Pathways Project: Design and Methodological Issues

Trivina Kang, Melvin Chan, Tan Teck Kiang, and David Hogan[1]

Abstract

This paper describes the Life Pathways Project, a longitudinal study of about 30,000 students from multiple cohorts in Singapore. Although such large scale studies are common internationally, such efforts are relatively new in Singapore, especially in the educational arena. This study introduces the wide array of outcomes measured in this study, its research design and gives details of how data was collected. It also surfaces a major challenge that the study faced with regard to attrition. This issue was most pronounced among the post-secondary cohort of students and a high proportion of students dropped out despite follow-up efforts and incentives. This raises questions of what other strategies need to be employed to engage youth to remain active participants in such longitudinal surveys.

Key Words: Youth, Schooling, Longitudinal Studies.

1. Introduction

In 2002, the Centre for Research in Pedagogy and Practice (CRPP) was established at the National Institute of Education, Singapore's sole teacher education institution. Supported by the Ministry of Education, this centre has become the largest educational research centre in the Asia Pacific. The primary focus of CRPP is on research in teaching and learning, specifically understanding what is happening in Singapore classrooms and developing new and innovative ways to enhance this experience.

Over the years, researchers within CRPP and NIE have conducted numerous projects. This paper describes one of the projects, the Life Pathways Project (LPP), which was designed to collect data on a broad array of student outcomes. This project distinguishes itself from other projects, not only because it is longitudinal in nature and also because it seeks to measure academic as well as non-academic outcomes of a large representative sample of students in Singapore. These outcomes include student academic performance in English and Mathematics assessments, as well as self reported data on student economic, social, civic, psychological skills, dispositions and understandings.

2. Purpose

Although such longitudinal studies have been conducted in other countries, this was the first large-scale effort of its kind in Singapore. Singapore students have received top places in international competitions like the Third International Mathematics and Science Competitions (TIMSS) and are known for their academic achievements. But as the suite of recent Ministry of Education initiatives like Teach Less Learn More, Social and Emotional Learning, Innovation and Enterprise, signalled, the Singapore MOE is increasingly committed to developing students who not only strive toward academic excellence but are also well-positioned and equipped for meaningful social participation in various social and institutional domains. As such, the LPP is well positioned to contribute to this broadened understanding of outcomes of schooling by addressing the development of a range of capacities that young people putatively need in order to participate successfully and rewardingly in the institutional and social life of Singapore.

The array of skills and capacities measured in the Life Pathways Study include:

- Economic capacities and participation ("Human capital")
    - Conventional measures of human capital: Academic achievement; school attainment (years of schooling)

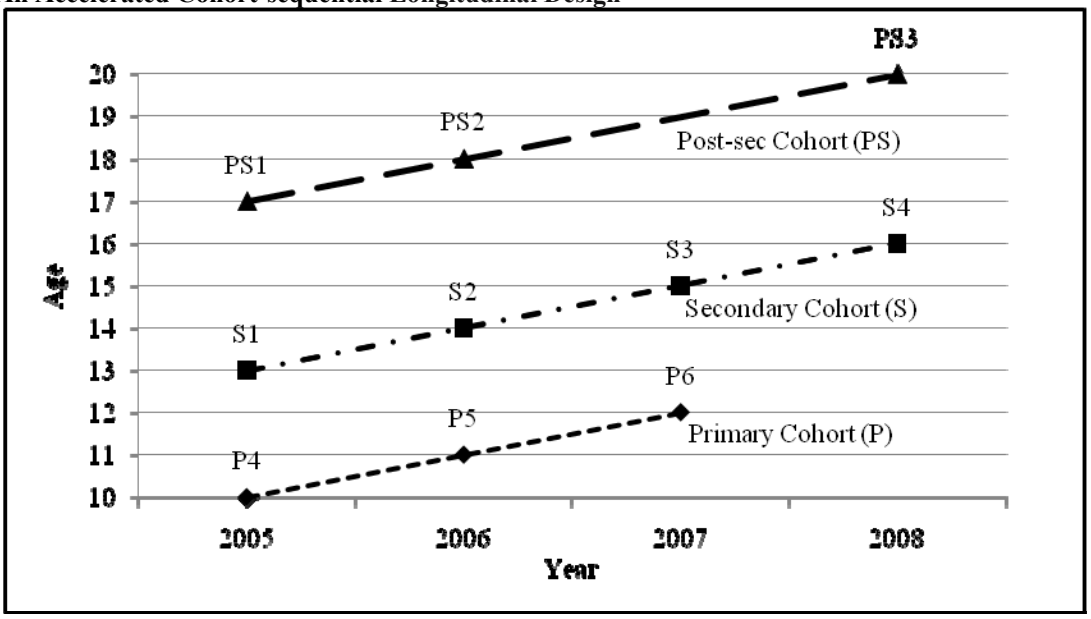
[1]Trivina Kang, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616 (trivina.kang@nie.edu.sg); Melvin Chan, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616; Tan Teck Kiang, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616; David Hogan, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616

    - "Knowledge economy" skills and capacities (e.g., orientation to risk taking; orientation to live long learning; orientation to individual and team work; adaptability; sense of agency/self efficacy; interpersonal problem solving skills; time management; multi-literacies; coping skills; effort regulation; information management; interdisciplinary work skills, etc)
    - Labor market participation and attainments (income, promotion, etc)
- Social skills and participation ("Social capital")
    - Social Networks; Friendships;
    - Intimacy (family, friends, boy/girl friends);
    - Trust & Attachment / Membership;
    - Leadership skills;
    - Community group memberships
- Civic capacities and participation ("civic capital")
    - Civic Beliefs (good citizenship, justice, national identity, etc);
    - Civic Agency (participation in political groups, voting intentions, reading/ discussing the news)
    - Civic Dispositions (interest/engagement, trust, respect for others);
    - Civic identity/national identities (sense of attachment, membership)
- Subjectivity (the organization of the self)
    - Self beliefs (self efficacy/subjective agency, self concept, self confidence, locus of control);
    - Values;
    - Goals and aspirations;
    - Identity formations (social, personal);
    - Reflexivity/self understanding (incl. meta-cognition)
- Subjective Wellbeing
    - Existential aspirations and satisfaction;
    - Mental health morbidities (anxiety, depression, exam anxiety);
    - Positive and negative affect;
    - Domain-specific life satisfaction (school, family, friendship, etc);
    - Life satisfaction & happiness
- Life Goals, Choices and Pathways
    - Life goals and aspirations
    - School engagement and attainment;
    - Work, career and labor market participation;
    - Social participation (networks, friendship, intimacy; civil society);
    - Family formation;
    - Participation in civic / political society
    - Self formation (biography)

3. Research Design

We have identified three cohorts of students – Primary, Secondary and Post-secondary and our design allows us to have data across this spectrum of about 10 years of schooling as the Primary 4 (P4, Age 10) cohort overlaps with the Secondary 1 (S1, Age 13) cohort and this then overlaps with the Post Secondary (PS1,Age 17) cohort.

Figure 3-1
An Accelerated Cohort-sequential Longitudinal Design



The P4 and S1 cohorts include all students in all classes in some 77 schools. This allows cross-sectional multilevel analysis of school level variables (Level 3) on student (Level 2) and on time (Level 1).

In total, 4 survey instruments were developed for the two younger cohort, P4 and S1 cohort and 3 instruments for the PS1. The survey instruments are differentiated in terms of their key outcome measures, but all share common predictor modules. The outcomes measures the instruments focused on include Form 1: Economic and social capital; Form 2: Subjective well-being; Form 3: Life planning and identities, and Form 4: Citizenship and civic capital. As the Post-secondary students are older and hence able to handle more survey questions, items in Form 3 in the Secondary survey were distributed to Forms 2 and 4 in the Post Secondary survey. Hence, only 3 forms were administered to this cohort.

At the first administration (at Age 10,13 and 17), students were randomly allocated one of the forms. In subsequent years, they would be routed to the same form upon logging into survey website with their individual identification numbers. Respondents were given a unique Internet address (URL) and survey password where they can access and complete the annual survey.

In addition, students in the Secondary cohort took a pen and paper assessment. In 2005, Educational Assessment Australia (EEA) at the University of New South Wales (UNSW) was engaged to design and produce annual Mathematics and English assessment instruments for this project. The assessment materials were developed by EAA as part of the International Competitions for Schools (English and Mathematics). These instruments are henceforth termed the "UNSW" assessments. The English instrument contained 64 questions that comprised the following domains: Reading (literacy); Reading (factual); Language (textual); Language (Syntactical); and Spelling. The Mathematics instrument had 40 questions which comprised the following: Number; Measurement; Chance and Data; Space and Geometry; and Algebra. These UNSW assessments allowed the project to obtain objective academic performance data across the entire Secondary cohort that was calibrated over time.

In Secondary 1, approximately half of the secondary cohort students in each selected school were randomly assigned to the English UNSW assessment while the other half were assigned the Mathematics assessment. Students continued with their assigned subject assessment across the 4 years.

4. The Issue of Attrition

As mentioned above, the samples for the study were large and once identified, entire cohort of students in the level (P4, S1 or PS1) took part in the study. As students had to be routed to the same survey instrument annually, we used their individual ID numbers to identify them for the relevant forms. For the P4 and S1 cohort, we were able to obtain cooperation and support from the schools (37 Primary schools and 38 Secondary schools) to arrange for time slots where students would log into our website and complete our surveys. Although there was attrition, primarily because students left the school, as can be seen from Table 4-1 and 4-2, the response rates were manageable for the Primary and Secondary cohort. Response rate for the online surveys were follows: 94.4% for Primary, 71.5% for Secondary. However, it is challenging to prevent steep attrition for the Post-secondary cohort. This is largely due to the fact that the post-secondary institutions were less amenable to give us access to their students. Despite repeated follow-up calls and incentives such as movie vouchers and lucky draws offering prizes like Apple notebooks, it was difficult to get the students to respond.

Table 4-1
Number of Respondents in the Life Pathways Study

	Primary	Secondary	Post-secondary
Wave 1	9620	11627	8836
Wave 2	10849	11904	5819
Wave 3	10978	11743	2318
Wave 4	--	11198	--
Total	9084	8303	2077

Table 4-2
Number of Respondents for the UNSW English and Mathematics Assessments

	UNSW Mathematics	UNSW English
Wave 1	6052	6063
Wave 2	6099	6109
Wave 3	5501	5536
Wave 4	5815	5977
Total	4265	4246

Given this high attrition rate, there has been a limit to our use of the post-secondary cohort longitudinal data for modelling purposes. We have also realised that despite the use of technology and online surveys that has the advantage of ensuring that all fields in the survey are completed, the age-old issue of ensuring that respondents participate, and in the case of longitudinal studies, continue to participate is one that needs to be addressed. In a context like Singapore, where there probably less involvement and interest in survey participation, ensuring that attrition rates remains low is an area that requires serious consideration in research design. Future research should also address the possibility that maybe online surveys may not be appropriate for all population groups, even though it is cost efficient and allows for easily compilation of data. We would also need further explorations of the type of incentives that would be attractive for respondents, especially those in Generation Y.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Experiences with the design and analysis of longitudinal data at Statistics New Zealand

Deborah Brunning[1]

Abstract

Prior to the turn of the century, Statistics New Zealand had little experience with the design and analysis of longitudinal data. However, over the last decade, the increased interest of policy makers in New Zealand, like their counterparts in many other countries, has identified a need for more information to enable the study of patterns and dynamics beyond what is achievable with repeated cross-sectional snapshots. To respond to this need, over the past 10 years Statistics New Zealand has: designed and run 7 waves of an 8 wave survey to measure income and employment dynamics (known as SoFIE); conducted, in partnership with the Department of Labour, 2 waves of a 3 wave longitudinal survey to measure migrant's settlement experiences in New Zealand (LisNZ); developed a longitudinal business database, by combining data from a number of sources; and developed a dataset which enables the examination of the longitudinal patterns and dynamics of both employers and employees using administrative data from the taxation system (LEED).

In this paper we will discuss our experiences with these developments. There have been both achievements and significant challenges encountered in this work, in areas such as the design and implementation of collection methodologies, use of computer assisted methods, and confidentiality and data access. We will examine how these experiences are influencing our thinking about future longitudinal data collections at Statistics New Zealand.

Key Words: Longitudinal, Business, Administrative Data, Social, Computer Assisted Interviewing, Data Access.

1. Introduction

Statistics New Zealand has been involved with four longitudinal collections over the last decade. This paper presents a brief overview of these four longitudinal collections, followed by a discussion about some of the experiences with this work, and how these influence the thinking for future longitudinal data collections at Statistics New Zealand.

1.1 An overview of Statistics New Zealand's Social longitudinal collections

Survey of Family, Income and Employment (SoFIE)
The Survey of Family, Income and Employment (SoFIE) is New Zealand's national survey designed to study changes in individuals, family and household income, and the factors that influence these changes in individuals, such as labour force involvement and family composition. SoFIE is a fixed panel design and attempts to re-interview the same group of individuals over eight years (or waves). At wave 1, about 15,000 households were randomly selected to take part in SoFIE, from which we collected data from 22,000 eligible individuals in 11,500 households.

Both a household and personal questionnaire is used to collect information from SoFIE respondents at each wave. The household questionnaire is answered by one adult in each household and collects information on the household characteristics (household type, family type) and standard of living (type of housing, appliances owned). The personal questionnaire is answered by every adult SoFIE respondent and collects information on demographics, labour market history and current activity, family, education and income.

Longitudinal Immigration Survey: New Zealand (LisNZ)
The Longitudinal Immigration Survey: New Zealand (LisNZ) is a longitudinal survey of immigrants, developed by the Department of Labour in partnership with Statistics New Zealand (who carried out the survey). The main objective of the LisNZ is to inform immigration policies, and assist with developing settlement services.

The LisNZ interviewed the same group of migrants at three waves; 6 months (wave 1), 18 months (wave 2), and 36 months (wave 3) after taking up permanent residence in New Zealand. The survey sample was selected from migrants aged 16 years and over who were approved for permanent residence in New Zealand from 1 November 2004 to 31 October 2005. Migrants

[1]Deborah Brunning, Statistics New Zealand, P.O. Box 2922, Wellington, New Zealand, deborah.brunning@statsgovt.nz, www.stats.govt.nz

were sampled at the time they were granted residence. For migrants who gained residence approval offshore, the wave 1 interview could have been as much as 18 months after their approval date, as migrants with residence approval offshore have 12 months to arrive in New Zealand and take up residence. Wave 1 interviews began in May 2005 and wave 3 interviews were completed in November 2009.

The LisNZ was designed to produce estimates for several subpopulations defined by the immigration approval category and region of origin. It was estimated that 5,000 completed interviews in wave 3 would be required to produce estimates of the required accuracy. Therefore, a large initial sample of over 12,200 was selected via a stratified systematic random design.

For practical and operational reasons the LisNZ survey population was restricted to migrants who were living at the first wave interview in the North Island, South Island or Waiheke Island (thus excluding most smaller offshore New Zealand islands), and could also understand at least one of the seven designated survey languages (English, Mandarin, Cantonese, Samoan, Korean, Hindi and Punjabi).

1.2 An overview of Statistics New Zealand's Business longitudinal collections

Linked Employer – Employee Data (LEED)
The Linked Employer-Employee Data (LEED) integrates existing employer and employee information together, providing insight into the operation of the New Zealand labour market and its relationship to business performance. LEED uses existing administrative data from two sources: tax data from the Inland Revenue and business data from Statistics New Zealand's Business Frame (BF). The LEED dataset is created by linking a longitudinal employer series from the BF to a longitudinal series of Employer Monthly Schedule (EMS) payroll data. The EMS payroll data provides data for all individuals who receive income from which tax is deducted at source. Within this tax data, individual employees are associated with an employer at a given time, which provides a unique employer-employee pairing. The LEED dataset provides an important link between business data and data about individuals.

The first release of LEED statistics was in February 2006, which provided a back series of data from the June 1999 quarter to the December 2004 quarter. Quarterly LEED statistics have continued to be produced for each quarter since the first release.

Longitudinal business database (LBD)
The Longitudinal Business Database (LBD) is a prototype longitudinal database of business information originally created as the end result of the two year IBULDD (Improved Business Understanding via Longitudinal Database Development) feasibility project.

The LBD integrates administrative and survey data that already exists from Statistics New Zealand and other government agencies. The core of the LBD is the Longitudinal Business Frame (LBF). The LBF is a longitudinal register of businesses and contains the longitudinal information (such as industry, ownership type and sector) on a wide population of firms. Another key component is the tax data from the Inland Revenue which provides business information on sales and purchases; financial performance and position variables; salaries, wages and employee counts. The LBD also includes a number of Statistics New Zealand sample surveys that measure business practices and performance.

The LBD is a comprehensive register of businesses that have been active at any time since 2000. The dynamics of business births, deaths, expansion, and contractions are captured in the longitudinal data.

2. Experiences

2.1 Social experiences

The SoFIE is Statistics New Zealand's first ever experience with a longitudinal social survey, therefore there was no previous longitudinal experience or expertise already within Statistics New Zealand to draw upon. The initial approach taken was heavily influenced by previous cross-sectional experiences. However, this wasn't enough, as we have since learnt that longitudinal collections require a different way of approaching the survey process (from design to analysis) than cross-sectional collections. The development and ongoing operations of the LisNZ, in most respects, has been less turbulent than SoFIE, as expertise gained and lessons learnt from early SoFIE experiences were applied to the LisNZ, where applicable.

The SoFIE processing system covers all required transformation (such as deriving variables, imputation, non-response adjustments, and calibration) of the data, from post editing (after collection) to having a data set suitable for analytical use and producing microdata for release. Different issues with aspects of the processing system have arisen since it was originally designed and built, but both the frequency and magnitude of issues continued to increase as the waves increased. A substantial

amount of resource over the last five years has been used to investigate, document, and provide fixes to some of the issues encountered. However, despite these efforts the system as a whole became so unstable that it was unable to be run leading up to the scheduled release of the wave -5 data in late 2008. The scheduled SoFIE releases had to be deferred, to allow time for the processing system to be stabilised. The SoFIE system stabilisation project is currently underway and is due for completion mid 2010. With a suitably operating processing system, Statistics New Zealand are aiming to get back on schedule with SoFIE releases, with the planned release of wave 1-7 data in late 2010. The issues encountered with the processing of SoFIE data have not affected the actual collection of data, which has continued in the field as planned.

Computer assisted person interviewing (CAPI) is used for both SoFIE and LisNZ. The SoFIE was the first of Statistics New Zealand's surveys to begin developing the questionnaire and field collection using any form of computer assisted interviewing (CAI). Now all Statistics New Zealand's household social surveys (both cross-sectional and longitudinal) use either CAPI or computer assisted telephone interviewing (CATI). The planned use of CAI from the start of both social longitudinal collections was a major advantage in addressing the needs of collecting the required information. The type and detail of information collected from respondents and the required complex routing between questions simply would not have been possible via a paper-aided person interview. Other key advantages of CAI include:

- allowing the use of dependent data, where information collected from earlier waves is incorporated into the current wave interview;
- unit record and consistency checks can be done during the interview, to solve queries and inconsistencies while the interviewer has the respondent with them;
- enabling easier collection and retrieval of respondent information for respondent management.

Questionnaire development, testing and programming is a resource intensive process. SoFIE was developed to incorporate the use of questionnaire modules (or sections) that would only be asked in particular waves of the personal questionnaire. The 'health' module contains a section of health related questions (self-rated health status, incidence of chronic disease, major health events, risk factors and health service use) which are only asked at wave interviews three, five and seven. While questions on the type and value of asset and liabilities are asked in a different module in wave interviews two, four, six and eight.
Even though SoFIE had incorporated modules, making any other changes to the questionnaire between waves (after the initial development) was restrictive.

2.2 Business experiences

The use of data from the New Zealand taxation system is a core component of both LEED and LBD. Statistics New Zealand's use of tax data is allowed for in the Tax Administration Act, providing certain conditions are satisfied. For example, the use of the data must not affect the integrity of the tax system. Initial challenges for both collections involved demonstrating that all the legislation, security, and confidentiality requirements for using tax data and the integration of different sources of data could be met.

Tax data is collected by the Inland Revenue for the purpose of administering New Zealand's taxation system. Statistics New Zealand's use of this data for LEED and LBD is for the production of statistics, which is a different purpose than it was collected for. Therefore cleaning, transformation and integration processes are required to produce the required microdata. Examples of some of the challenges faced in applying these processes were:

- difference between collection units and statistical units;
- changes in administration numbers as a result of changes in business structures, such as mergers or splits;
- missing or miscoded tax numbers for employees or businesses;
- methodology used to link different data sources together.

The LBD also integrates other (non tax) administration data and survey data. A key part of the development work of the LBD involved establishing common linkages, units and timeframes to all the source data. For example, one of the business surveys included is the Annual Enterprise Survey (AES), which has collection units based on the Kind of Activity Units (KAU). The common unit used in LBD is the enterprise level. The AES collection units have to be aggregated to the enterprise level before they can be linked into the LBD.

Providing access to the LEED microdata was the very first experience for Statistics New Zealand in providing external users with access to any business microdata (either longitudinal or cross-sectional). Microdata access for longitudinal business data has been a greater challenge than that of longitudinal social data due to the actual make-up of the business data. For example, New Zealand has very few large businesses and in some industries such as telecommunications or transport, these large businesses have a high risk of being identified within the microdata. Extra security and confidentiality measures are used for

both LEED and LBD to address this risk. For example, restrictions are placed on the physical access site of the microdata, who has access to the microdata and the way in which any produced output from the microdata is released outside of the access site.

2.3 Data access and confidentiality

A key data access challenge faced by official statistics agencies is ensuring that both the safety (protection of respondents' confidentiality) and the utility of the microdata are maintained when provided to external users. Statistics New Zealand provides several different ways for external users to access microdata based on the user needs and data requirements. The confidentiality techniques applied to microdata for protection is dependent on the level of restriction placed on access to the microdata. In general, the more public the access, the greater the amount of disclosure control applied.

It is possible to produce a microdata file from a cross-sectional survey with adequate levels of safety and utility for public release. The risk of being able to identify respondents on longitudinal microdata is much greater than that related to cross-sectional, simply due to its longitudinal nature. For example, information showing transitions over time can be highly identifying to respondents. Achieving an adequate level of safety in longitudinal microdata for public release would likely make the utility level far too low for any users needs.

Statistics New Zealand's approach to providing access to longitudinal microdata is to make it available via our Data Lab environment only. Data Labs are contained within each of our office buildings in three different cities and are controlled environments for pre-agreed users and research projects. All output is confidentiality checked prior to being released outside of the Data Lab. The amount of time involved for project approval, initial set up in the Data Lab, as well as getting output checked and released is sometimes viewed by users as the greatest barrier to access. Reflection and review of our confidentiality checking processes, purpose and techniques are underway to identify aspects where timeliness can be improved while still ensuring the safety of the microdata.

The Data Lab environment and access to microdata in general, was also a new experience for Statistics New Zealand over the last decade. Much of the drive behind providing access to microdata (and the procedures and guidelines for this access) has been enacted as a result of the input from many of the external users of Statistics New Zealand's longitudinal data. One example of this is the requested release of unweighted counts by SoFIE researchers in the Data Lab. The resulting discussions and consideration of this request lead to a change in Statistics New Zealand's confidentiality practice. In March 2008, a new policy was endorsed, which allowed unweighted counts from social surveys to be released and published under certain conditions for the first time.

2.4 Resources

Resources are limited and it is always an ongoing challenge for any organisation to decide where and how these resources are allocated or spent.

Reflecting back on the early stages of SoFIE and considering the actual amount of resources used to this point, the amount of initial and ongoing resources required had been under estimated. Longitudinal collections require a large long-term resource investment. However, it is not just the size of available resource that is important, but the way in which the resources are assigned and used within different areas or stages of the collection. The current rebuilding of the SoFIE processing system has highlighted the importance of getting the initial resource investment correct. Fix-ups or rebuilding systems at a later stage will result in the use of more resources in total.

People are one of the most important types of resource. The development and retention of longitudinal expertise and knowledge of staff is an ongoing challenge, especially given the long-term nature and complexity of longitudinal collections in comparison to cross-sectional collections. Over the last decade Statistics New Zealand has invested heavily in developing staff capabilities in the design, field collection and processing aspects of longitudinal collections. In the past there has not been as much focus on analysis or the actual use of our own longitudinal data. An ongoing focus for the future is to make better use of our own data, as well as better assisting others to use our data.

3. Conclusion

For Statistics New Zealand the longitudinal aspects of SoFIE, LisNZ, LEED and LBD have provided our data users with the ability to investigate and explore changes to New Zealand businesses or individuals over time. This simply would not have been possible with cross-sectional collections. The experience gained from these collections has provided Statistics New Zealand with the opportunity to develop longitudinal knowledge and expertise to draw upon in the future.

REDESIGN OF LARGE-SCALE LONGITUDINAL SURVEYS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Continuity and Innovation in the Design of *Understanding Society:* the UK Household Longitudinal Study

Dr Heather Laurie[1]

Abstract

This paper discusses the design decisions made in the incorporation of the British Household Panel Survey (BHPS), a long-running household panel survey in the UK, into a new and larger study, *Understanding Society*: the UK Household Longitudinal Study. The paper sets out the background to the new study, the key features of the new study and its potential for analysis, the survey design of *Understanding Society* and the decisions made in incorporating the BHPS sample into the new design, and the decisions around continuity and innovation in the content of the questionnaires.

1. Introduction

For survey designers longitudinal studies present specific problems in adapting to change, especially when the design has been more or less constant over the life of a long-running study. This paper discusses such a situation where The British Household Panel Study (BHPS) faced a transition to a new and larger study, *Understanding Society* the UK Household Longitudinal Study. The BHPS is a household panel survey of around 8,000 households in the UK which has completed 18 annual waves of data collection. The BHPS has been the major source of household panel data for the UK since 1991 and is widely used by academic and policy researchers. Following extensive consultation with the user community, a contract to establish a new and larger household panel of 40,000 households, called *Understanding Society*: the UK Household Longitudinal Study, was awarded to the research team responsible for the BHPS who are based at the Institute for Social and Economic Research (ISER), University of Essex. Wave 1 of *Understanding Society* began in January 2009 with wave 2 commencing in January 2010. The design of *Understanding Society* includes the incorporation of existing BHPS sample members from wave 2 of the new study. This paper outlines the design of *Understanding Society* and how experience from the BHPS informed design and implementation decisions in setting up the new study. The rationales for decisions made in incorporating the BHPS into the new study are considered. These include reconciling the competing demands for continuity and innovation, decisions on the timing of interviews throughout the year, questionnaire content, and maintaining panel loyalty while making the transition to a different fieldwork organisation.

1.1 Background to Understanding Society

Understanding Society is a new household panel survey funded by the UK Economic and Social Research Council with co-funding from UK government departments and motivated by the success of longitudinal studies in the UK. The UK has a diverse and rich portfolio of longitudinal studies including not only the BHPS but the British Birth Cohort studies (National Child Development Survey, British Cohort Study 1970, Millennium Cohort Study), studies of ageing such as the English Longitudinal Study of Ageing (ELSA), youth cohort studies such as the Longitudinal Study of Young People in England (LSYPE), and Census longitudinal studies. The BHPS is a widely used dataset in the UK but it was recognised that as a mature panel, there was a need to look to the future of longitudinal data resources for the coming decades in the UK. In taking the BHPS forward into a new era, continuity with the existing design and data was important. But it was also clear that there were new demands which would inevitably lead to changes to the BHPS model and which needed careful consideration. One of the key elements for the new study was sample size as it was recognised that a larger sample size than was available on the BHPS would enable a broader range of analysis and finer grained analysis of sub-groups within the population. The total achieved sample for *Understanding Society* was set at 40,000 households including the existing sample of BHPS households.

[1] Dr Heather Laurie, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. laurh@essex.ac.uk

1.2 Key features and analysis potential

There are eight key features of the design of *Understanding Society* that reflect its scientific rationale and can be exploited to generate major innovations in scientific research. These offer significant analysis potential beyond those available within the BHPS and include the following.

(i) The **large sample size** of 40,000 households provides an opportunity to explore issues where other longitudinal surveys are too small. Analyses of small subgroups, such as teenage parents, disabled people, or the young unemployed, which are of key policy concern but which, when sub-groups are disaggregated by other characteristics may become too small for robust analysis. Analysis at regional and sub-regional levels, allowing examination of the effects of geographical variation in policy, also becomes possible. And the large sample size allows high-resolution analysis of events in time, for example focussing on single-year age cohorts. Events which are relatively rare in the population, such as divorce for example, will occur in greater numbers any given wave something which enables robust analysis without having to wait for a number of waves to build up sufficient numbers of events for analysis.

(ii) The **household focus of the design** means data will be collected on all members of sampled households and their interactions within the household. This has major advantages for important research areas such as consumption and income, where within-household sharing of resources is important, or demographic change, where the household itself is often the object of study. Compared to individual-based birth cohorts, it will give better and more continuous information on the family and household environment within which early development takes place. The ability to observe multiple generations and all siblings will allow examination of long-term transmission processes and isolate the effects of commonly shared family background characteristics. Compared with existing studies, *Understanding Society* will provide much greater opportunities to explore linkages outside the household.

(iii) A **full age range sample** complements age-focused cohort studies in the UK such as those sampling elderly people or young people and provides a unique look at behaviours and transitions in mid-life. For example, even for issues of pensions and long-term care, usually associated with old age, the policy setting is heavily influenced by behaviour occurring earlier in people's lives when decisions about savings, pensions and investments are taken. Moreover the large sample size means that all cohorts can be analysed at a common point in time.

(iv) The inclusion of an **Innovation Panel for methodological research and testing** provides a vehicle for a range of methodological work to inform not only the survey development but to contribute to survey methodological developments in general.

(v) The **multi-topic design** aims to meet a wide range of disciplinary and inter-disciplinary research needs. *Understanding Society* is a multi-topic survey for the study of a range of life course domains. While meeting the needs of 'traditional' quantitative social science disciplines such as economics, sociology and social policy, it also serves other disciplines (e.g. the biomedical sciences) and make possible a wider set of methodological approaches. It will facilitate genuine interdisciplinary research: within the social sciences (e.g. geography and economics); within the biomedical sciences (e.g. psychology and genetics); and between the two.

(vi) The incorporation of an **ethnicity research** agenda within *Understanding Society* recognises the increasing prominence of research into ethnic difference for our understanding of the make-up of British society and issues of diversity and commonality. It also emphasises the potential of such research to transform our understanding of individual and social processes. Through its coverage of large numbers of minority group members over time and of ethnicity-relevant research domains, the ethnicity strand will enable critical advances in knowledge of these topics and the processes involved.

(vii) *Understanding Society* will support collection of a wide range of biomarkers and health indicators. This opens up exciting prospects for advances at the interface between social science and biomedical research. It will provide the opportunity to assess exposure and antecedent factors of health status, understanding disease mechanisms (e.g. gene-environment interaction, gene-to-function links), household and socioeconomic effects and analysis of outcomes using direct assessments or data linkage.

(viii) Extensive data linkage to administrative records and geo-coded data is planned with permissions to link to health, education, pension and state benefit records being requested of survey participants. These data will provide significant new areas for analysis where the rich, contextual social survey data can be used in combination with administrative data.

2. Survey design

2.1 Household panel design

The design of *Understanding Society* follows that of the BHPS and other national household panels. The study design is a longitudinal sample of individuals representing the whole UK population, and interviewed within a household context. The sample consists of private residential addresses drawn from the UK Postcode Address File (PAF) and all members of private households found at those addresses at wave 1 are designated as original sample members. The sample is an equal probability sample with a total of 2640 Primary Sampling Units (PSU is a postcode sector in the UK) containing 18 issued addresses per PSU. In Northern Ireland the sample is a simple random sample. Each monthly sample consists of 110 PSUs with each month providing representative data for the UK with a key aim of the design being to enable analysts to produce quarterly estimates for the UK. The issued sample size assumes a minimum household response rate of 60% after the exclusion of ineligible addresses identified during fieldwork. At each wave all sample members aged 10 and over are eligible for interview and individuals are followed as they move and form new households. Other individuals who form households with sample members after wave 1 become eligible for interview as long as they are resident with an original sample member. As with the BHPS, the following rules mean that the sample will remain representative of the UK 2009 population as it changes, subject to weighting and except for new immigrants to the UK. The inclusion of a new immigrant sample is planned for the future. (see http://www.iser.essex.ac.uk/survey/bhps for details of the BHPS design and content and http://www.understandingsociety.org.uk for further details of *Understanding Society*).

Beyond the basic specification of the household panel design, there are some key differences between the *Understanding Society* and BHPS designs. The first of these is the inclusion of an Innovation Panel, a longitudinal sample of 1500 households which goes into the field a year in advance of the main sample and is used for methodological testing and experimentation. As with the main sample, this is an equal probability sample of Great Britain (in this case south of the Caledonian Canal) and excluding Northern Ireland with 120 PSUs and 23 issued addresses per PSU. The Innovation Panel is used for randomised experiments in fieldwork procedures such as incentives and materials provided to respondents, split ballot experiments on the use of showcards, alternative question wording, the use of dependent interviewing in measures of change, as well as for sensitivity testing of certain items (see Laurie, Burton and Uhrig (2008) for preliminary results from wave 1 of the Innovation Panel).

Secondly, the design includes an ethnic minority oversample in medium to high density areas of ethnic minorities which, from 2001 UK Census data, are estimated to cover 80 – 90% of the ethnic minority populations of interest in the UK. The oversample is designed to achieve an additional 1,000 interviews in each of five key groups: Indian, Pakistani, Bangladeshi, Caribbean and Black African with eligible households identified through a screening survey. In addition all mixed background respondents identified in the screening interview including Chinese, other Asian and Middle Eastern respondents are eligible for inclusion. The boost design also incorporates a comparison sample from the general population sample of some 500 households. The estimated sample sizes for the various components of *Understanding Society* are set out in Table 2.1-1 below.

Table 2.1-1

Estimated Achieved Sample Sizes (Households) waves 1 and 2

	New sample *	BHPS Sample**	EM Boost	Total
Wave 1	29,850	8,100	4,220	42,170
Wave 2	25,370	6,900	3,590	35,860

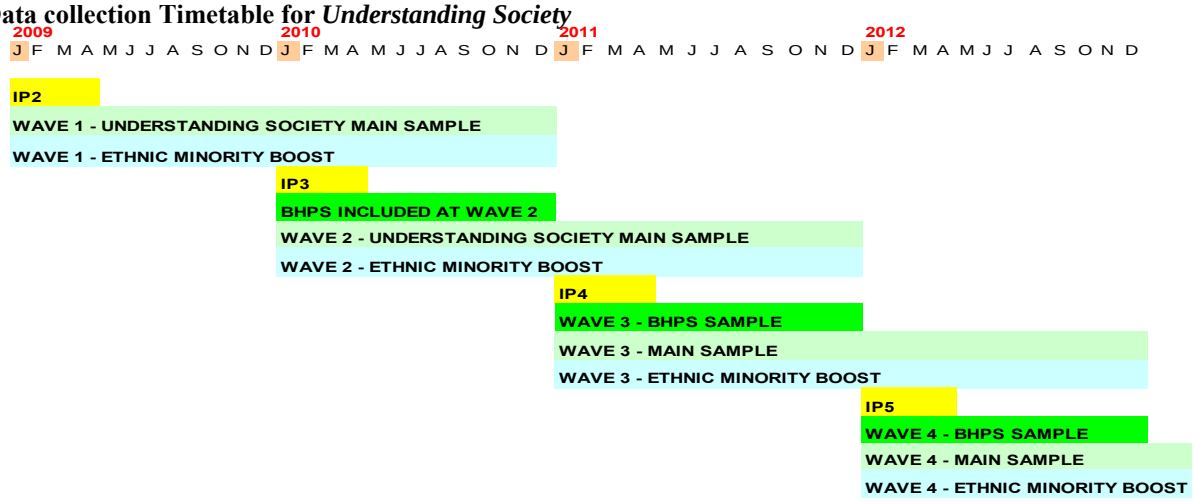
* Includes the Innovation Panel of 1500 households and the 500 household comparison sample for the boost.
** Note that BHPS Wave 18 is considered here as part of *Understanding Society* Wave 1

2.2 Timing of data collection

The BHPS is an annual interview and for 18 years, fieldwork has started on September 1st with the main fieldwork period running until December of the same calendar year. This has been followed by a relatively long tail of a further three months to deal with cases which were difficult to contact or away during the fieldwork period, tracing movers and refusal conversion. The design of *Understanding Society* also has an annual interview at 12 month intervals but given the large sample size and fieldwork capacity for getting the interviews completed, it has continuous fieldwork over a 24 month period for each wave.

Respondents will maintain their wave 1 sample month throughout the life of the study even if they move address between waves. The implication of this design is that there are overlapping waves where year 2 of wave 1 overlaps with year 1 of wave 2. Monthly samples for *Understanding Society* began in January 2009 and the BHPS sample is being incorporated from wave 2 of the study in 2010. Table 2.2-1 sets out the data collection timetable for each wave.

Table 2.2-1
Data collection Timetable for *Understanding Society*



One difficult decision with respect to the BHPS sample was whether to maintain the wave 1 to wave 18 data collection timetable or alternatively, to integrate the BHPS sample fully into the *Understanding Society* fieldwork timetable by assigning it to monthly samples. There were arguments for and against keeping the BHPS data collection in the same fieldwork period as in previous waves. Respondents expected to be interviewed between September and December each year and were generally ready for an interviewer to call on them at that time of year. For analysts, keeping the same data collection period would ensure consistency in terms of any seasonal effects and provide continuity for analyses which were built up over a number of waves of data. The arguments for incorporating the BHPS sample into the *Understanding Society* collection timetable were primarily to do with a desire to fully integrate the BHPS into the new study for the future. In addition, keeping the BHPS sample data collection in the same period as at previous waves would produce a lumpy and difficult to manage distribution for fieldwork.

On balance, the view taken was that incorporating the BHPS into the monthly sample design would enhance the possibilities of using the BHPS in conjunction with the new sample rather than being seen as something of a separate sub-study within *Understanding Society*. In terms of the longer term strategy for the study as a whole integrating the BHPS into the monthly sample design seemed the most coherent approach. Despite this decision, there were concerns about the longer than normal gap between interviews for the BHPS sample. The wave 18 interviews took place in late 2008 and into early 2009, so the BHPS sample could not be incorporated into wave 1 of the new study starting in January 2009. Wave 2 of *Understanding Society* started in January 2010 and if the BHPS sample were evenly spread across the two calendar year period for wave 2, the gap between interviews could be as great as 36 - 38 months. This was seen as unacceptable so the decision was taken to randomly allocate the BHPS sample across the first twelve months of the wave 2 fieldwork period giving a maximum gap between interviews of two years with 50% of the sample being interviewed within around 18 – 20 months of their previous wave 18 interview. This means that the BHPS sample will be included in the first year of data collection at each wave in the future.

2.3 Data collection mode

The BHPS has always been conducted with face to face interviews in respondent's homes, which since 1999 have been Computer Assisted Personal Interviews (CAPI). The intention on *Understanding Society* is to move to mixed modes of data collection at some waves, partly due to fieldwork costs but also to provide respondents with a greater choice of mode by which to answer. Waves 1 to 3 of *Understanding Society* are being conducted face to face using CAPI as establishing the panel in the early waves through maximising response rates and minimising attrition at those waves is recognised to be critical for the long-term quality and health of the study. It is likely that either wave 4 or wave 5 of the study will use a mixed mode data collection strategy and the Innovation Panel is being used to test different sequential mode strategies in order to determine which is most effective in terms of overall response rates and fieldwork efficiency. Wave 2 of the Innovation Panel (conducted April – June 2009) experimentally tested sequential mixed modes using telephone and face to face interviews and we anticipate including web interviews at a later phase of the development work. Analysis of these data, including extensive para-data on calling

patterns, is currently underway. The Innovation Panel is therefore a resource not only for informing the development of the main study but also for conducting innovative methodological research of wider interest to survey practitioners and methodologists.

2.4 Changing fieldwork organisations

For the BHPS sample, their first contact with *Understanding Society* will be a face to face interview at wave 2 (2010). For eighteen years, the BHPS has been conducted by GfKNOP, a London based survey organisation sub-contracted by the survey team at ISER to carry out the interviews and deliver the data. However, the process of commissioning the new study resulted in a different fieldwork provider, the National Centre for Social Research (NatCen), being awarded the data collection contract for waves 1 and 2 of the study. This change presented certain integration issues for the BHPS sample, many of whom had been interviewed by the same interviewer for all or many waves of the BHPS. Our concern was that this change in fieldwork provider might impact negatively on response amongst the BHPS sample. No matter how loyal they may have been to the BHPS over the years, it could be expected that this change might lead some respondents to decide to withdraw from the study, something we clearly wished to avoid as far as possible. In addition to the change of interviewers and fieldwork organisation, BHPS respondents needed to have a new name and associated new logo and 'look' introduced to them.

Special procedures to introduce BHPS respondents to the new survey and to interviewers new to them were developed. These included a series of mailings between the wave 18 BHPS interview and the first contact at wave 2, informing respondents about the changes to the study and encouraging them to see these in a positive light. Their importance as founder members of the study was stressed, the new name, design of logos and materials introduced to them, additional promotional material was sent to them, details about NatCen were provided and they were informed of their new month of interview. Prior to the interviewer calling at the address, an advance letter tailored for the BHPS sample is being sent which interviewers personalise with their name and telephone number details. As is standard practice on the BHPS, there is a dedicated phone-line, email address and participant website so that respondents can contact the survey team directly if they have any concerns. As wave 2 has just begun, we have no evidence to date of how the transition to the new study is progressing for the BHPS sample but will be monitoring progress closely and making any interventions needed during fieldwork.

3. Questionnaire design and content

3.1 Continuity and innovation in content

The questionnaire structure for *Understanding Society* has a similar format as BHPS and consists of:

- Household questionnaire (including household membership roster) of 15 minutes
- Individual interview for all adults aged 16 and over (32 minutes)
- Self-completion questionnaire for all interviewed adults (7 minutes)
- Proxy interview for adults aged 16 and over who are unable to be interviewed (10 minutes)
- Self-completion Youth questionnaire for children aged 10 – 15 (on the BHPS the Youth questionnaire is from age 11 – 15).

All questionnaires within the CAPI script and associated fieldwork materials for respondents are translated into nine languages including Welsh, and bi-lingual interviewers or translators provided where necessary.

While the structure of the interviews uses the same instruments as BHPS questionnaire, there is significant demand from the user community for new or extended content relative to the BHPS to be included on *Understanding Society*. The BHPS sample will receive the new questionnaire from wave 2 and it has been important to maintain some comparability with BHPS in order to minimise disruption to long-running time series on key measures. As a result, there is significant use of BHPS questions in *Understanding Society*. A guiding principle in making decisions about any specific question, was that where an existing BHPS question had no better replacement that would clearly produce better quality data, it would be carried forward without changes. However, where there was a clear rationale in terms of improved data quality for changing or improving a BHPS question we included new, revised or replacement questions.

Substantive areas where either new content relative to the BHPS or extended content has been required include:

- Family and social networks outside the household
- Attitudes and behaviours related to environmental issues
- More on illicit and risky behaviour especially for young people
- More on neighbourhood attachment and social engagement
- Cultural activities and use of leisure time

- More on parenting attitudes, behaviour and parenting styles
- Psychological attributes
- Cognitive ability/functioning measures
- More on health outcomes and health related behaviour
- Quality of sleep
- Well-being
- Quality of marital relationships
- Risk and trust
- Collection of data about younger children < 10 from parent
- More data on transition into young adulthood
- Discrimination and harassment
- Ethnic and national identity

While these new areas of coverage were seen as centrally important, the survey also needs to collect the key social, economic and employment data which form the backbone of the longitudinal content. One of the major problems faced was the reduced individual interview length on *Understanding Society* which is significantly shorter than has been the case on the BHPS (32.5 minutes vs. 40 minutes respectively). This length constraint has meant that a fairly high proportion of BHPS questions are either not being included at all, or more usually will be asked less frequently than annually on *Understanding Society*.

3.2 Annual and rotating modular design

The solution to the length constraints was to move away from the BHPS model where around 80% of the questionnaire was annual repeated measures and 20% rotating thematic content and with most people being eligible to be asked all questions. *Understanding Society* has adopted a model where there is more use of (i) questions asked regularly, but not every wave; (ii) triggered questions asked only after key events such as the birth of a child or moving house; (iii) questions asked only at particular ages; and (iv) sub-samples based on demographic characteristics. In contrast to BHPS, around 50% of the individual questionnaire is devoted to annual repeated measures and the remainder to rotating thematic modules carried either every two or three years depending on the subject matter and expected rates of change. The annual repeating measures are largely BHPS questions carried forward which will allow longer time series analysis to continue for that portion of the sample as well as providing continuity with the new *Understanding Society* sample as the longitudinal data come on stream over the next few waves.

A key design task in the initial stages was to agree which measures were critical for the annual panel design and which questions should be carried annually. The annual repeated measures include basic demographic characteristics; changes between waves - employment, education, training, fertility, partnering, geographic mobility, health conditions; health status (e.g. SF12), disability; labour market activity and employment status, job search; current job characteristics, basic employment conditions, hours of paid work, second jobs; childcare, other caring within and outside household; child development at key ages; income and earnings; life satisfaction; political affiliation; transport and communication access; education; retirement planning/expectations at key ages; consumption expenditure; housing characteristics; housing expenditure ; household facilities, car ownership. In addition, the wave 1 questionnaire included factual background measures such as place of birth, educational qualifications, and details of parental background along with some life history information including a cohabitation, marriage and fertility history and a migration history for the ethnic minority boost.

3.3 Bio-measure collection

The collection of direct physical measures and samples is scheduled to begin during wave 2 of Understanding Society and to continue through wave 3. While the available funding does not allow collection from all sample members we estimate that between 25,000 and 30,000 cases will have these measures, providing a significant proportion of the total sample. The aim is to collect these measures from all BHPS sample measures so that the longitudinal survey data can be used in conjunction with the bio-measure data and provide early longitudinal results.

There are two types of bio-measures being collected. The first are direct anthropometric measures and the second are samples to be stored for use in later analysis. Health researchers are increasingly incorporating summary measures derived from these biomarkers in the analysis of pathways to disease and in the exploration of the emergence of health disparities (for example Chandola, Brunner and Marmot, 2006; Seeman et al, 2004; Sabbah et al, 2008). The data collection for this element of the study will be done either by nurse interviewers or by trained survey interviewers. The measures to be taken are:

- Whole blood (nurse only)
- Lung function (spirometry) (nurse only)
- Blood pressure and pulse rate

- Height and Weight
- Waist circumference
- Bio-electrical impedance
- Grip strength
- Saliva (interviewer only)
- Dried blood spots (interviewer only)

In the UK, there are well-developed protocols and standards for collecting physical measures using nurse interviewers and this is done on several large scale studies. The collection of some of the proposed measures by trained interviewers, in particular taking blood pressure and pulse rate, saliva, and dried blood spots has never been done on large scale surveys in the UK. The protocols and interviewer training for this element are innovative for the UK. There is substantial experience in major US population surveys of successful use of these 'minimally invasive' approaches being carried out by well-trained survey interviewers (McDade et al, 2007, Lindau and McDade, 2008; Weinstein et al 2008) and we are drawing on this experience in developing our procedures.

In addition to the physical measures, wave 3 will collect cognitive ability/functioning measures for the whole sample. Overall, these data should provide an opportunity to assess exposure to and antecedent factors of people's current health status; a better understanding of disease mechanisms such as gene-environment interactions; household and socioeconomic effects on health and an analysis of outcomes using direct assessments and data linkage.

4. Conclusion

Understanding Society is a highly ambitious study in both scope and coverage and is designed to keep the UK at the leading edge of social science research. A major feature is the intention to be a bio-social survey providing new opportunities for inter-disciplinary research across the medical and social sciences. As a data resource, *Understanding Society* therefore offers data of multiple types from multiple sources for a longitudinal sample of households and individuals over time.

The BHPS has been a huge success over the past eighteen years and is highly values as a longitudinal data resource for the UK. The data are widely used by academic researchers within the UK and across the world and it is regularly used by government departments and policy makers as a source of high quality, authoritative longitudinal data for the UK. The BHPS has been designed to the highest academic standards and has constantly evolved throughout its life, with the transition to *Understanding Society* being the most recent development to ensure a bright future for the continuation of the BHPS in the coming years. *Understanding Society* is an ambitious study with a wider scope than the BHPS in terms of design, content and potential analysis uses. Getting it right is therefore of critical importance, with a successful transition into the new study for the BHPS sample being an important marker of the success of the study. The design decisions facing the survey team at ISER have not been easy and in many cases we will not know how successful we have been until analyses using the BHPS sample in conjunction with the new *Understanding Society* sample begin to appear. Judgements have had to be made based on the available evidence, advice from the user community, meeting the longitudinal analysis needs and aims of the study, and our experience of designing and managing longitudinal surveys. We hope we have found the right balance between continuity and innovation demanded by *Understanding Society* while remaining true to the principles of providing high quality longitudinal data always followed on the BHPS.

References

Chandola, T., Brunner, E. and Marmot, M. (2006). Chronic stress at work and the metabolic syndrome: prospective study. *British Medical Journal* 332: 521-525.

Laurie, H., Burton, J. and Uhrig, S.C. Noah (eds) (2008). Understanding Society: Some preliminary results from the wave 1 Innovation Panel, *Understanding Society Working paper Series, 2008 – 03, Institute for Social and Economic Research, University of Essex* http://research.understandingsociety.org.uk/files/working-papers/2008/usocwp-2008-03.pdf

McDade, T.W., Williams, S. and Snodgrass, J.J. (2007). What a drop can do: dried blood spots as a minimally-invasive method for integrating biomarkers into population-based research. *Demography* 44(4): 899-925.

Lindau, S.T. and McDade, T.W. (2008). Minimally-invasive and innovative methods for biomeasure collection in population-based research. In M. Weinstein, J.W.Vaupel and K.W. Wachter (Eds) *Biosocial Surveys*. Washington D.C., National Academies Press.

Sabbah, W., Watt, R.G., Shelham, A. and Tsakos, A. (2008). Effects of allostatic load on the social gradient in ischaemic heart disease and periodontal disease: evidence from the Third National Health and Nutrition Examination Survey. *Journal of Epidemiology and Community Health* 62: 415-420.

Seeman, T.E., Crimmins, E., Huang, M.-H. et al. (2004). Cumulative biological risk and socio-economic differences in mortality: MacArthur Studies of Successful Aging. *Social Science and Medicine* 58: 1985-1997.

Weinstein, M., Vaupel, J.W. and Wachter, K.W. (Eds, 2008). *Biosocial Surveys*. Washington D.C., National Academies Press.

For further information see:

http://www.understandingsociety.org.uk
http://www.iser.essex.ac.uk

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Survival and Revival of the Survey of Income and Program Participation

David S. Johnson [1]

Abstract

For the past two decades, the Survey of Income and Program Participation (SIPP) has been the leading source of data about the economic well-being of Americans. SIPP has been used to evaluate the effectiveness of government programs by many federal, state, and local agencies, academic institutions, and private research and policy study bodies.

Recently, the U.S. Census Bureau initiated a project to reengineer the SIPP in order to provide crucial information in a timely manner and at reduced cost through reengineered survey design, improvements in processing efficiency, and a focused content scope. The main purpose of SIPP is to provide a nationally representative sample that can be used to evaluate the annual and sub-annual dynamics of income, the movements into and out of government transfer programs, and the effect on family and social context of individuals and households. The main activities of this reengineering process include: (1) improvements in the collection instrument and processing system; (2) development of an Event History Calendar in the instrument; (3) use of administrative records data to supplement and evaluate survey data; and (4) development of survey content and use of reimbursable supplements, through interactions with stakeholders.

An important activity initiated with the development of reengineered SIPP improvements was the interaction and consultation with stakeholders on both content and design of the proposed improvements for the re-engineered SIPP. In addition, the Bureau began fielding the current SIPP collection in September, 2008. Most recently, other activities in the re-engineering process have included:

- Evaluating a test of a paper version of an Event History Calendar (EHC) questionnaire.
- Planning for a larger scale test of an automated EHC questionnaire in early 2010.
- Reconstitution of an American Statistical Association/Statistical Research Methodology SIPP advisory subcommittee.
- Open meetings to discuss information needed to obtain recommendations from a National Academy of Sciences-Committee National Statistics panel commissioned to advise on the plans for and research of use of administrative records in the re-engineered SIPP.
- Procurement of administrative record files and consultation services for some national level and selected state level data on government programs to assess data quality of both the paper and automated test data.
- Plan for in depth consultation on training for Field Representatives in the Event History Calendar methodology of interviewing.

[1] David S. Johnson, U.S. Census Bureau, U.S.A. (david.s.johnson@census.gov)

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Results from the Canadian Household Panel Survey Pilot

Andrew Heisz[1]

Abstract

In January 2006, a conference on longitudinal surveys hosted by Statistics Canada, the Social and Humanities Research Council of Canada and the Canadian Institute of Health Research concluded that Canada lacks a longitudinal survey which collects information on multiple subjects such as family, human capital, income, labour and health and follows respondents for a long period of time. Following this conference, Statistics Canada received funds from the Policy Research Data Gaps fund to support a pilot survey for a new Canadian Household Panel Survey Pilot (CHPS-Pilot). Consultations on the design and content were held with academic and policy experts in 2007 and 2008, and the pilot survey was conducted in the fall of 2008. The objectives of the pilot survey were to: (1) test a questionnaire and measure the quality of data collected; (2) evaluate several design features and; (3) test reactions to the survey from respondents and field workers. The pilot survey achieved a response rate of 76%, with a mean household interview time of 68 minutes. Several innovative design features were tested, and found to be viable. Response to the survey, whether from respondents or interviewers, was generally positive. This paper highlights these and other results from the CHPS-Pilot.

Key Words: Canadian Household Panel Survey, Longitudinal Surveys, Pilot Surveys.

1. Introduction

In the fall of 2008, Statistics Canada, in partnership with Human Resources and Social Development Canada (HRSDC) and the Canadian academic community, fielded a the Canadian Household Panel Survey Pilot (CHPS-Pilot). This paper describes the background of the project, the steps taken in the development of the pilot survey, the methodological results, and the reaction to the pilot survey from respondents and field workers. A separate study will evaluate the questionnaire and quality of the data collected in the pilot.

2. Pilot development

2.1 Background

In January 2006, Statistics Canada, the Social Science and Humanities Research Council of Canada and the Canadian Institute of Health Research hosted a conference "Longitudinal Social and Health Surveys in an International Perspective"[2]. The objective of this conference was to take stock of longitudinal social and health surveys fielded in Canada and elsewhere. At that time, Statistics Canada's longitudinal social and health survey program was nearing 15 years old, and the conference provided an opportunity to discuss the successes and shortcomings of Statistics Canada's longitudinal portfolio. All of Statistics Canada's longitudinal surveys were discussed and comparisons were drawn to longitudinal surveys in other countries.

Among other things, the conference identified an important data gap for Canada: Canada lacks a "general household panel survey". A general household panel survey is a multitopic longitudinal household survey with a sample representative of the population. Canada has longitudinal surveys that focus on specific topics, like the National Population Health Survey and the Survey of Labour and Income Dynamics (SLID). Canada also has longitudinal surveys that focus on particular sub-populations like the National Longitudinal Survey of Children and Youth, the Youth in Transition Survey, and the Longitudinal Survey of Immigrants to Canada. These surveys are valuable as they allow for in-depth research on a particular topic or sub-population. However, a general household panel survey would allow for research that stretches beyond traditional subject matter domains, enabling researchers to see how events in one domain may affect others, perhaps much later in life. It would allow researchers to investigate how lives evolve in social contexts, the most immediate social context being the family, but also larger social contexts such as friendships, neighborhoods and public services. The design of such surveys calls for interviewing all

[1]Andrew Heisz, Income Statistics Division, Statistics Canada, 5th Floor, Jean Talon Building, 170 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, andrew.heisz@statcan.gc.ca
[2] Most papers presented at this conference are available at www.ciqss.umontreal.ca/Longit/index.html.

household members, allowing analysis of family dynamics and their interactions with other domains in ways that would be impossible for other surveys.

A number of other recommendations came from the "Longitudinal Social and Health Surveys in an International Perspective" conference that relate to the design of the new household panel survey:

• It should focus on four subject matter domains: (1) Labour and Income, (2) Family, (3) Human Capital Development, and (4) Health.
• It should expand the number of topics that could be captured over the life of a panel by using rotating modules.
• It should have an indefinite panel length.
• It should not be used for the production of cross sectional income estimates.
• It should engage academic and policy experts, and keep them engaged as partners in the ongoing survey governance.
• It should be flexible, with the ability to adapt to emerging research and policy needs.
• It should be comparable in design and content to international general household panel surveys such as the German Socio Economic Panel (GSOEP), the British Household Panel Survey (BHPS) and the Household Income and Labour Dynamics in Australia survey (HILDA).
• It should be easy to use, and emphasize the importance of minimizing processing time, simplifying weighting and methodology, and reducing the learning time needed to understand the dataset[3].

Following the Montreal conference, the Policy Research Data Gaps fund provided three years of funding for the CHPS-Pilot.

2.2 Governance and content development

The CHPS-Pilot project was developed under a tripartite governance system with each of Statistics Canada, HRSDC and the academic community represented. A steering committee made up of two Director Generals from Statistics Canada, a Director General from HRSDC and two academics directed the project. The survey was managed at Statistics Canada in the Income Statistics Division.

Content development took place from February 2007 through March 2008. Content development was driven by four academic expert groups, each responsible for one of the four major subject matter domains identified above. These four expert groups, comprising about 20 Canadian academics, advised on content needs and priorities, and discussed possible research uses for the new survey. Each of the four academic expert groups prepared a report indicating what data should be collected for each domain, including some indication on which items could be used as rotating content as opposed to content that would appear in each year. These reports were developed over several months, during which time there were a number of meetings held to exchange preliminary ideas. A draft of the pilot survey was then produced.

Policy research experts from federal departments were consulted to comment on the draft survey. Researchers from HRSDC, the Canadian Mortgage and Housing Corporation, and the Bank of Canada were consulted on data needs, resulting in numerous changes to the draft survey. Qualitative testing of the questionnaire in the form of one-on-one interview testing was undertaken from January through March 2008.

Application development and data processing took place in Special Surveys Division of Statistics Canada. As with all social surveys at Statistics Canada, methodology was the responsibility of Household Surveys Methodology Division while collection was carried out by Collection Planning and Management Division.

2.3 Survey design

The CHPS-Pilot survey was conducted between October 15th and December 31st, 2008. The name of the pilot, for the purposes of field work, was the Living in Canada Survey – Pilot. The survey design followed the model established by the HILDA, GSOEP and BHPS for general household panel surveys. Briefly:

Target population: the target population was all Canadians living in households, excluding institutional and on-reserve populations. Households were sampled (the frame and sampling approach is discussed below), and all household members became permanent sample members, regardless of their age.

Target respondent: the survey was to interview all household members aged 15 and over living in sampled households.

[3] Please see Picot, Berthelot and Webber (2006) for more details.

Mode: the collection method was non-proxy computer assisted personal interviews (CAPI), although the option was given to interviewers to conduct the interview over the telephone using the CAPI application if a face to face meeting could not be scheduled.

Sample size and frame: The sample comprised 1,200 dwellings selected from the Labour Force Survey (LFS) Rotate-out frame and 1,400 dwellings selected from the LFS Area-frame. Roughly equal sized samples were drawn from each of four provinces - Ontario, Quebec, Saskatchewan and New Brunswick. To minimise collection costs, the pilot-survey had a small sample size and used a highly clustered design. As a result, the pilot survey was not expected to yield estimates of the population.

Following rules: The following rules for the survey were not firmly established as the intent was for a single wave pilot only. However, they would have been similar to those used in the HILDA, BHPS or GSOEP surveys. In those surveys, all permanent sample members are followed indefinitely, cohabitants of permanent sample members are also interviewed, and children of permanent sample members themselves become permanent sample members.

2.4 Pilot survey content

The pilot survey was to develop a wave-1 questionnaire; the intent of a wave-1 questionnaire is to establish a foundation of data upon which the later waves of the survey could build. Accordingly, the pilot survey included some questions of a retrospective nature, as well as questions designed to generate a baseline picture in human capital, labour, health and well-being.

The survey was structured as a number of components.

• The entry component made a roster of all household members and collected basic demographics about each of them. This component was responded to by a person identified to be knowledgeable about all household members (a "person most knowledgeable" or PMK)
• A household component, which was a set of questions about the household or its members, that one could reasonably expect the PMK to answer on behalf of all household members.
• A member component, which was to be asked of each household member aged 15 and over. Household members aged 0-14 did not receive an interview.

The household component asked questions on housing, childcare use, monthly expenditures on key items and total monthly household income. It also included questions on food, financial and housing security, as well as material deprivation.

The member component included retrospective questions on marriages and common law relationships, parenting history and fertility intentions, educational history, and jobless spell histories. It also contained questions on current labour market activities, characteristics of current jobs, and new questions targeted towards the self-employed, skills used at work, and questions on employment expectations. Other questions sought main activities in the event that the respondent was not working. Four modules were identified to gather information on work to retirement transitions: two were used in the pilot, with the intent that the other two would be used in the second wave of the survey. These questions were asked of respondents 45 years of age and older. Finally, the member component included questions on demographics and life satisfaction.

3. Pilot results

3.1 Pilot objectives

The objectives of the CHPS-Pilot Survey were to (1) test a questionnaire and measure the quality of data collected; (2) evaluate several design features and; (3) test reactions to the survey from respondents and field workers. As noted earlier, a paper describing the questionnaire and data quality is forthcoming. In terms of design features, the specific goals were to:

1. Establish a response rate
2. Gather information to help chose the best frame for this survey
3. Examine the difficulty completing interviews in larger households
4. Examine the use of the telephone interview
5. Get information on questionnaire and question length

In this section we evaluate these design features as well as report on reactions to the survey.

3.2 Response rate

It is important to have a high response rate for any survey, but for a longitudinal one, perhaps it is even more important, as a low response rate in the first wave may cast doubt on the representivity of the survey for the life of the panel. Moreover, responses to the wave-1 survey permit the construction of auxiliary information to adjust for attrition in subsequent waves, which could extend the life of the panel.

Canada, like elsewhere in the world, has seen a decline in willingness to respond to the first wave of longitudinal surveys. For example, in SLID, a CATI panel survey where a new first wave has been launched every three years since 1993, the wave-1 response rate has fallen from over 90% in the early panels of the survey, to mid-70% rates in the late 2000s.

Table 3.2-1 shows the response rate from the CHPS-Pilot survey. In future waves of the survey, one would target an individual level response rate, but in the first wave, where membership of the sample is unknown due the household level sample frame, the target response rate must be set at the household level. The target response rate for the CHPS-Pilot was for 80% of households to be partially interviewed (a partial interview is one where some but not all eligible members were interviewed). This target response rate was established based upon recent response rates of other CAPI surveys at Statistics Canada. Altogether, there were 2,356 addresses issued, yielding 2,122 households eligible for interview. The (compete plus partial) response rate was 76%, which was slightly lower than the target. Overall, 5,453 individuals were enumerated, with 3,205 eligible for interview. Fully 91% of those eligible for interview were interviewed.

Table 3.2-1
Response rates from the CHPS-Pilot

Wave 1 household outcomes		
Addresses issued	2,356	
Out of scope	237	
Multiples (additions to sample)	3	
Eligible households	2,122	100%
Refusal and non-contact	501	24%
Complete plus partial household coverage	1,621	76%
Complete household coverage	1,405	66%
Wave 1 individual outcomes		
Enumerated individuals	5,453	
Ineligible children (Under 15)	2,248	
Enumerated adults	3,205	100%
Refusals, non-contacts and partial interviews	298	9%
Complete member interviews	2,907	91%

One way of evaluating the success of this response rate would be to look at wave 1 response rates for other successful household panel surveys to determine if the CHPS-Pilot met the standards set by international surveys. Results published for the 1991 Wave-1 BHPS (Lynn (2006)) and the 2001 Wave-1 HILDA (Watson and Wooden (2002)) indicate that the 1991 BHPS achieved a 74% partial plus complete response rate, while the 2001 HILDA achieved a 66% partial plus complete response rate, indicating that the 76% CHPS-Pilot response rate may have been enough to match the successes of the international surveys. A possible area for improvement in the CHPS-Pilot was in the share of completed member interviews which was as high as 95% in the BHPS, but only 91% in the CHPS-Pilot.

3.3 Frames

As noted above, the CHPS-Pilot used two frames to draw sample from – the LFS Rotate-out frame and the LFS Area frame. The objective of this approach was to evaluate which frame would yield a higher response rate. The Rotate-out frame included dwellings previously used in the September 2007 LFS. In most cases, respondents at these dwellings would have previously spent up to six months in the LFS, which could have a negative effect on response. However, the benefit to using the Rotate-out frame is that it contains a large amount of auxiliary information on respondents which could be used to improve wave-1 non-response adjustment. In contrast, respondents at dwellings in the Area-frame sample would not have had any prior experience with Statistics Canada surveys (aside from the Census), but the auxiliary information available on this frame is much more limited. It should be noted that no tracing of former LFS respondents selected from the Rotate-out frame was implemented in the pilot; rather, interviews were conducted with the current occupants of the dwellings. Ideally, we would

have compared response rates from the two frames after attempting to trace Rotate-out frame members who had moved in the year since the September 2007 LFS.

Table 3.3-1
Results comparing LFS-Area Frame and LFS-Rotate-out Frame

	Area frame		Rotate outs	
In Scope Households	1,180		942	
Total Non Response	265	22%	236	25%
No Contact	18	2%	18	2%
Refused	174	15%	167	18%
Mental/Physical Limitations	9	1%	6	1%
Language Barrier	9	1%	4	0%
Other	55	5%	41	4%

Results from the two frames are shown in table 3.3-1. It can be seen that the refusal rate from the Area frame (at 15%) was lower than the Rotate-out frame (at 12%), but not by a great amount. Moreover, non-contact in the Rotate-out frame (at 2%) would, if anything, have been higher if tracing were implemented. However, the slightly higher response rate in the Area frame comes at the cost of loss of the useful auxiliary information that would come with the Rotate-out frame. As a result, it is ambiguous which of the Area frame or the Rotate-out frame would have been better to use for the survey. However, the results of this test suggest that the effect of the prior response burden of the LFS on the response rate would be minor were we to choose to use the Rotate-out frame.

3.4 Difficulty completing large households

One important concern for the success of the survey was the response burden the survey design places on large households, and the corresponding difficulty completing interviews for each respondent in large households. Table 3.4-1 shows statistics on household completion rates for households of various sizes. First, large households were quite rare. Only 16.3% of households required three or more interviews, while 4.2% required four or more. However, large households were harder to complete. Fully 99% of one-respondent households are completed, compared to 85% of 2-respondent households, 66% of 3-respondent households 76% of 4-respondent households, and 56% of 5-respondent households.

Most (52%) respondents lived in households with two target respondents. On average it took 4.6 attempts to interview both respondents in the household. Interestingly, the mean number of contacts required to complete a household with two respondents was not more than a household with one respondent, with the average number of contacts being 4.5 for a one respondent household. However, more contacts, and hence more interviewer effort were required for households with three or more respondents.

Table 3.4-1
Results comparing LFS-Area Frame and LFS-Rotate-out Frame

Household Size	1	2	3	4	5
% households	31.2	52.2	12.1	3.3	0.9
% complete	98.8	85.3	66.2	76.1	56
attempts for a complete	4.5	4.6	6.4	6.9	7.3

3.5 The telephone interview

As noted earlier, the CHPS-Pilot was a CAPI interview. Interviewers were instructed to make their first contact with the household a personal contact, and wherever possible to interview each respondent in person. However, interviewers were given the freedom to complete interviews with difficult to reach respondents over the telephone. This was done in order to reduce response burden in large households, increase the share of completed households, and reduce collection costs. Moreover, in the event that certain cases were referred to senior interviewers, these were most likely to have been conducted over the telephone.

Table 5 shows the number of attempts, number of contacts, and the number of complete cases for the two collection methods. An attempt is an effort to contact a household, a contact is an attempt that successfully reaches a household member, and a complete case is a contact that results in a case status being finalized, which means that interviewers will not attempt to contact the household again.

Table 3.5-1
Comparison of Interview Attempts by Personal and Telephone Collection

Collection method	Number of attempts	Number of contacts	Number of complete cases
In Person	9,266	4,715	1,164
Telephone	5,069	3,015	442

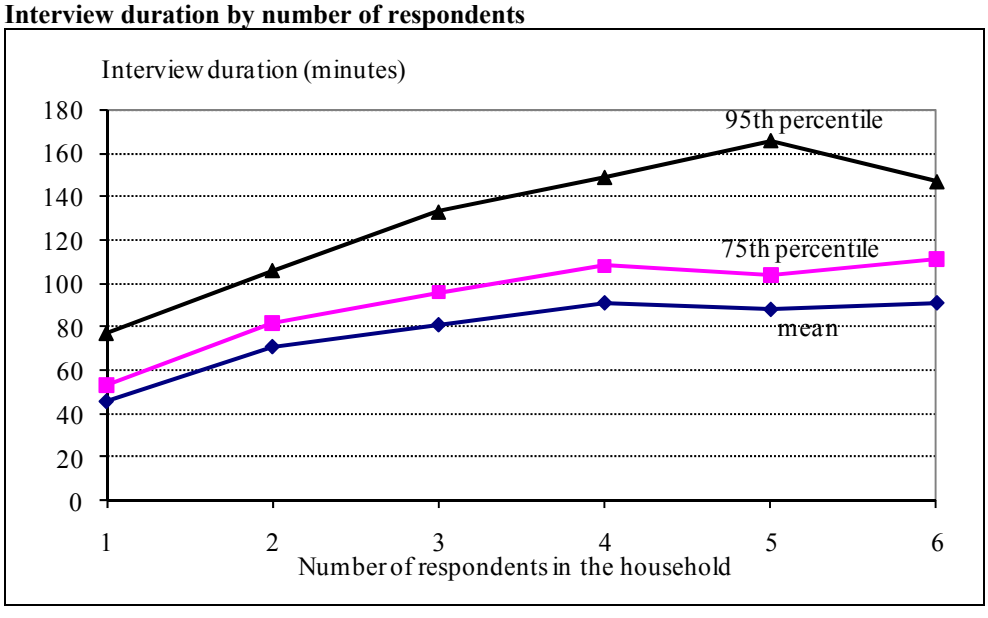
Fully 35% of attempts, 39% of contacts and 27% of complete cases were conducted over the telephone, which are relatively high rates for a survey conducted by personal interview. These rates are lower, but still remain high if we exclude from the calculations attempts made by senior interviewers (which are usually made by telephone). This reveals the importance of telephone collection in the survey for achieving the response rate described above. It may be that telephone interviews should be integrated more into the survey plan, perhaps with the development of a more formal procedure for telephone usage, or through the use of CATI instruments (in addition to CAPI) and CATI call centers.

3.6 Interview and question length

The expected interview length was possibly the most highly scrutinized metric during the design of the CHPS-Pilot as there was concern that seeking to interview all household members aged 15+ would lead to to an excessive amount of time spent in some households. While there was some discussion that we limit the number of interviews to four per household, it was decided that we would instead limit the interview length through careful pruning of the questionnaire. This was done using audit trail information on interview time per question from other surveys, as well as through discussions with subject matter and collection experts. Where audit trail information was not available, ad-hoc evaluations, such as timing of questions in mock-interview situations, were used.

In the end, average interview time was 12.6 minutes for the household component and 24.8 minutes for the member component. These interview times were quite close to what we had anticipated before collection. Considering only households where all target respondents were interviewed, the average interview time was 68 minutes (including entry and exit). Figure 3.6-1 shows interview lengths by number of target respondents (showing only completed cases) indicating that the average interview length, was 46 minutes in a single respondent household, 71 minutes in a two respondent household, 81 minutes in a three respondent household, and 91 minutes or less in 4, 5 or 6 respondent households. While 91 minutes was seen as a long interview time, in the context of CHPS-Pilot, where the interview time is shared among household members, it was seen as acceptable. However, figure 3.6-1 also reveals that in that some households, the interview time was much longer than the average. For instance, 5% of four respondent households required more than 149 minutes of interview time to complete, indicating that measures to reduce the length of unusually long interviews are warranted.

Figure 3.6-1
Interview duration by number of respondents



3.7 Reactions to the survey

In total, 4 program managers, 10 senior interviewers, and 109 regular interviewers were involved in data collection for the CHPS-Pilot, and the training of field workers yielded many opportunities for survey developers to receive informal reactions to the survey. Moreover, there were weekly conference calls with the program managers during the field period to debrief survey managers on progress in the field as well as to raise any questions or concerns. Finally, a debriefing questionnaire was given to interviewers after collection to receive direct feedback on the survey.

In general, interviewers were found to be very receptive to the survey. According to the results of the debriefing questionnaire, respondents were also receptive: 95% of the interviewers said respondents had a positive or neutral attitude towards the survey. A large number of interviewers (80%) felt that respondents clearly understood the purpose of the survey, and 75% of interviewers felt that conducting personal interviews did not cause problems. Several interviewers reported that the use of telephone interviewing allowed them to overcome scheduling difficulties

Nevertheless, interviewers reported back on a number of concerns with the survey design. Among the major concerns, almost half of the interviewers found the interview to be too long, and 61% felt that having multiple respondents or needing multiple visits caused a problem during collection. Scheduling of interviews and repeat visits for large households were two areas reported as being difficult.

4. Conclusion

Results from the CHPS-Pilot continue to be evaluated. In particular, the attention of the survey team at Statistics Canada and the academic experts has moved towards an evaluation of content appearing in the CHPS-Pilot. An evaluation of the content will appear in a separate report.

The CHPS-Pilot demonstrated the feasibility of a general household panel survey in Canada, established a probable wave 1 response rate, a target interview length, and yielded useful information on frame use. These results and other expertise gained from the CHPS-Pilot will inform future work on longitudinal household survey development at Statistics Canada.

References

Picot, G., Berthelot, J.-M. and Webber M. (2006). Possible Future Directions for Longitudinal Surveys at Statistics Canada, *Longitudinal Social and Health Surveys in an International Perspective Conference*, Montreal, Canada.

Watson, N. and Wooden M. (2002). The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology, *Hilda Project Technical Paper Series,* No.1/02, May 2002, University of Melbourne, Melbourne, Australia.

Lynn, P. (2006). Quality Profile: British Household Panel Survey, Version 2.0: Waves 1 to 13: 1991-2003, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, England.

LATENT MODELS AND BAYESIAN ESTIMATION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Latent Growth Curve Modelling of Life Satisfaction Trajectories in the British Household Panel Survey

Maria de Fátima Salgueiro, Marcel D. T. Vieira and Peter W. F. Smith[1]

Abstract

Life satisfaction trajectories are modelled using data from the British Household Panel Survey (BHPS), a national representative survey conducted on an annual basis. Respondents have been asked to rate their satisfaction levels with income, house/flat, job, social life, amount of leisure time, and its use. Employees who were interviewed in all waves from 1991 to 2005 and fully answered all life satisfaction variables in all waves are considered. Latent growth curve modelling is used to model variation in the two perceived life satisfaction latent factors considered. The benefits of taking into account the complex survey design are discussed.

Key Words: Latent Growth Curve Models, Longitudinal Data, Subjective Well-Being.

1. Introduction

Recent years have seen a growth in interest in subjective well-being (SWB) by social scientists. Several measures have been proposed, the choice of measurement instrument influencing the assessment of SWB and its determinants (Peasgood, 2007). The British Household Panel Survey (BHPS) is a national representative survey conducted, since 1991, on an annual basis. Several SWB measures are available in the BHPS. Since wave 1996, in addition to a question on overall life satisfaction, respondents have been asked to rate their satisfaction levels with eight domain dimensions (health, income, house/flat, spouse/partner, job, social life, amount of leisure time, and its use). Statistical approaches adopted in the literature to model SWB often include ordered probit models and fixed effects models. Random effects models and cross-lagged structural equation models have also been proposed to model longitudinal survey data (see, *e.g.*, Berrington *et al.,* 2008).

In the current paper trajectories of life satisfaction are investigated and modelled using BHPS data. Employees who were interviewed in all waves 1 to 15 and fully answered all life satisfaction variables in all waves are considered. Latent growth curve modelling (LGCM) is used to model both within-individual and between-individual level variation in the two perceived life satisfaction latent factors considered. Possible determinants of life satisfaction include age, gender, having children, family income, level of education and number of working hours. The benefits of taking into account the complex survey design are discussed. The structure of the paper is as follows. Section 2 describes the conceptual framework and the variables under analysis and presents some descriptive statistics to characterize the sample. Section 3 presents the conditional LGCM proposed to describe and explain *Leisure* and *Material Satisfaction* growth trajectories and summarizes the methodological options undertaken concerning the statistical modelling. The main results are described in Section 4, and Section 5 contains a brief discussion.

2. The data

1.1 Conceptual framework and variables under analysis

Data from the British Household Panel Survey (BHPS) are used. Four waves are considered: years 2002 to 2005. The sample includes 2255 respondents: employees, who have remained employed throughout the period under analysis, aged 16 years old or more, who have participated in the survey since wave one (so that longitudinal weights can be used), with full interview outcome.

Six Life Satisfaction variables are used as dimensions of SWB. Variables were measured on a 7 point Likert type scale and include Satisfaction with Social life; Amount of leisure time; Use of leisure time; Household income; House / flat; and Job.

[1]Maria de Fátima Salgueiro, ISCTE-IUL – Lisbon University Institute, Av. Forças Armadas, 1649-026 Lisboa, Portugal (fatima.salgueiro@iscte.pt); Marcel D.T. Vieira, Universidade Federal de Juíz de Fora, Brasil (marcel.vieira@ufjf.edu.br); Peter W.F. Smith, Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom (P.W.Smith@soton.ac.uk)

The first 3 variables are expected to be indicators of a *Leisure Satisfaction* factor, whereas the last 3 are expected to measure *Material satisfaction.* The 2 latent factors should be positively correlated. Nine possible determinants of life satisfaction are considered as explanatory variables, namely Gender; Age category; Number of children in the household; Qualification; Social class; Marital status; Perceived health status; Number of hours worked per week; and Household income.

As far as the survey design is concerned, the initial sample (wave one) of the BHPS was selected by a stratified multistage clustered design. Information regarding the primary sampling units (PSUs), the strata and the longitudinal weights are available in the BHPS datasets and are used in the analysis. In the BHPS longitudinal weights account for losses between each adjacent pair of waves, as well as for the initial weight at wave one (for further details see Taylor *et al.,* 2008).

1.2 Some descriptive statistics

The sample includes 2255 respondents, 49% are female and 51% are male. The distribution of the age group follows: 4.7% are between 16 and 21 years old; 13.9% between 22 and 29; 23.7% between 30 and 39; 32.1% between 40 and 49 and the remaining 25.6% are 50 years old or older. Regarding marital status, 74.2% live with company and 25.8% live alone. The number of own children in the household ranges from 0 (74.2%) to 6, and is less than or equal to 2 in 94% of the households. For 77% of the respondents the number of working hours is 30 hours per week, or more. Health status is perceived as excellent for 28% of the respondents; as very good for 50.3%; as fair for 16.9%; and as poor or very poor for 4.8%.

Descriptive statistics for the six satisfaction variables in 2002 show that, for most variables, the relative frequency of answers ranging from 5 (satisfied) to 7 (completely satisfied) is above 60%. The exception is for satisfaction with the amount of leisure time, with a value of 50.6%. This suggests the majority of the respondents have reasonably high satisfaction levels. When average satisfaction values are considered, the average score is above 4.3 for all six variables in all the four years considered in the analysis. The mean satisfaction levels have remained quite stable over the time period under investigation.

3. The proposed latent growth curve model

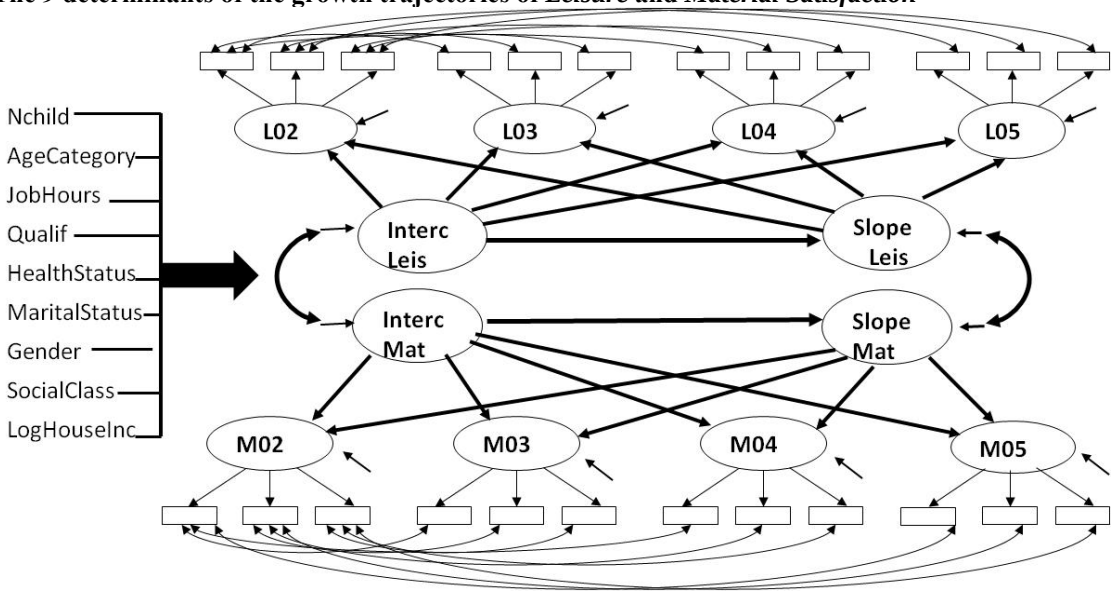
Latent growth curve models (LGCM) are part of the general family of random effects models – for an introduction see Preacher *et al.* (2008). An important advantage of this methodology is that it is carried out using structural equation modeling (SEM) methodology, sharing many of its strengths and weaknesses (see Duncan *et al.*, 2006). This allows the data analyst to take advantage of standard SEM features, such as global tests and assessment of model-data fit, multiple indicator latent variables and simultaneously modeling of multiple processes with regressions amongst the latent growth parameters. Yet, large sample sizes are still required.

LGCM permit the investigation of inter-individual differences in change over time, and allows for the investigation of the antecedents and consequences of change. The incorporation of both time-varying and time-invariant covariates is also possible. In LGCM time is incorporated in the model via constraints on the factor loadings in a latent variable model. Growth parameters are therefore specified as latent variables, and constraints are imposed on the factor loadings. The shape of the growth trajectory (linear, quadratic) depends on the number of latent variables specified in the model. The repeated observations are regarded as imperfect measures of an underlying latent trajectory. The shape of such trajectory can be estimated using the mean and the covariance structure of the observed repeated measures. For a more detailed explanation see Boolen and Curran (2006).

3.1 The conditional LGCM proposed to describe and explain *Leisure* and *Material Satisfaction* growth trajectories

Figure 3.1-1 displays the two-process conditional LGCM proposed to describe and explain the joint growth trajectories of the *Leisure* and M*aterial* S*atisfaction* latent factors.

Figure 3.1-1
The 9 determinants of the growth trajectories of *Leisure* and *Material Satisfaction*



Leisure and *Material Satisfaction* latent factors are both measured by three manifest variables at each time point. An invariant factorial structure is assumed over time and the error terms of the repeated measures are allowed to be correlated over time (for simplicity of display, only some of the corresponding arrows are displayed in Figure 3.1-1). Initial *Leisure* and *Material* levels of satisfaction are assumed to influence the corresponding growth trajectories, and for that reason slopes are regressed on intercepts.

In order to explain the joint growth trajectories, the two random intercept and the two random slope variables are regressed on the nine explanatory variables. The number of children in the household and the logarithm of the household income are modeled as continuous variables. For the remaining seven categorical explanatory variables dummy variables were created and are included in the model. All manifest variables are treated as ordinal variables. Linear growth is assumed, with the usual coding convention 0, 1, 2, 3 for the loadings, 2002 being the baseline year.

3.2 Main methodological options

The statistical modeling is conducted in three steps. First, Confirmatory Factor Analysis (CFA) is used to assess the adequacy of a model with two correlated factors (*Leisure* and *Material*) to account for the 6 life satisfaction dimensions, at each time point. In a second step the two-process second-order conditional LGCM presented in Figure 3.1-1 is considered to jointly explain the growth trajectories of the two latent life satisfaction factors, taking into account the complex survey design. In a third step the same model is estimated without incorporating the sampling design information and results are compared to those previously obtained in step 2. All models are estimated in Mplus 5 (Muthén and Muthén, 1998-2007), using a robust least squares estimator (WLSMV) available (Asparouhov, 2005). Only valid cases are considered. The option Complex is used to incorporate sampling design features in the inference procedures and standard errors are estimated using a sandwich type estimator (Muthén, 1998-2004).

4. Results

4.1 Main findings taking into account the complex survey design

Table 4.1-1 displays the estimates (and corresponding standard errors in parenthesis) for the main parameters in the fitted LGCM.

Table 4.1-1
Estimates (and standard errors) for the main parameters in the conditional LGCM

	LEISURE	MATERIAL
Mean of the Intercept	0 (---)	0 (---)
Mean of the Slope	-0.133 (0.235) n.s.	-0.178 (0.171) n.s.
Variance of the Intercept	0.599 (0.021)	0.372 (0.025)
Variance of the Slope	0.088 (0.016)	0.046 (0.009)
Regression of the Slope on the Intercept	0.051 (0.039) n.s.	-0.023 (0.035) n.s.
Covariance between random intercepts	0.373 (0.017)	
Covariance between random slopes	0.054 (0.008)	

n.s. denotes a non-significant estimate

Similar results are obtained if the complex survey design is not considered. It is also possible to conclude that there are several non-significant predictors of the life satisfaction average growth trajectories. Indeed, none of the 9 explanatory variables has a significant impact on the random slope of *Leisure Satisfaction* and most explanatory variables have no significant impact on the average growth trajectory of *Material Satisfaction* from 2002 to 2005. Note that the 9 explanatory variables are modeled as time-invariant: only their 2002 values are considered in the model.

4.2. The effects of not taking into account the complex survey design

Table 4.1-2 displays the estimates (and corresponding standard errors) for the regression coefficients associated with the explanatory variables in the model. Two solutions are presented: without taking into account the survey design and taking the survey design into account. Values in bold correspond to significant estimates.

Table 4.1-2
Predictors of Life Satisfaction average levels in 2002

		LEISURE Random Intercept		MATERIAL Random Intercept	
	Estimate (Std Error)	without Survey Design	with Survey Design	without Survey Design	with Survey Design
Gender	Female	-0.04 (0.047)	-0.05 (0.05)	**0.095** (0.043)	0.083 (0.049)
Age Group	22-29	-0.087 (0.106)	-0.089 (0.112)	**-0.199** (0.092)	-0.126 (0.104)
	30-39	-0.203 (0.108)	-0.19 (0.114)	-0.152 (0.095)	-0.08 (0.099)
	40-49	**-0.338** (0.105)	**-0.335** (0.11)	**-0.208** (0.092)	-0.153 (0.095)
	>= 50	**-0.22** (0.106)	**-0.221** (0.107)	-0.061 (0.094)	0.002 (0.098)
Job Hours	16-29	**-0.211** (0.091)	-0.177 (0.096)	-0.043 (0.079)	-0.013 (0.085)
	30-40	**-0.273** (0.084)	**-0.216** (0.087)	-0.103 (0.076)	-0.092 (0.088)
	> 40	**-0.451** (0.104)	**-0.392** (0.105)	-0.049 (0.095)	-0.04 (0.116)
Qualif	Higher	**0.176** (0.063)	**0.146** (0.067)	0.068 (0.056)	0.056 (0.057)
	A levels	0.151 (0.082)	**0.161** (0.079)	0.037 (0.071)	0.069 (0.073)
	O levels	**0.259** (0.074)	**0.258** (0.083)	0.113 (0.064)	0.106 (0.064)
	Other	**0.344** (0.081)	**0.326** (0.077)	**0.251** (0.071)	**0.246** (0.078)
Marital Status	Alone	**-0.179** (0.056)	**-0.158** (0.066)	**-0.238** (0.051)	**-0.243** (0.053)
Social Class	Managerial	0.015 (0.101)	0.02 (0.114)	-0.069 (0.085)	-0.059 (0.098)
	Skilled	-0.048 (0.105)	-0.041 (0.114)	**-0.198** (0.089)	**-0.215** (0.105)
	Unskilled	-0.015 (0.113)	-0.018 (0.127)	-0.145 (0.095)	-0.152 (0.111)
Health Status	Good	**-0.314** (0.05)	**-0.336** (0.053)	**-0.314** (0.045)	**-0.309** (0.043)
	Fair	**-0.71** (0.064)	**-0.747** (0.068)	**-0.643** (0.058)	**-0.637** (0.065)
	Poor	**-0.859** (0.096)	**-0.887** (0.095)	**-0.787** (0.088)	**-0.81** (0.088)
Number of Children		**-0.133** (0.026)	**-0.155** (0.03)	0.003 (0.024)	-0.011 (0.028)
Income		-0.022 (0.107)	-0.039 (0.116)	**0.537** (0.093)	**0.576** (0.107)

Values in **bold** correspond to significant estimates

It is possible to conclude that, on average, in 2002 women are significantly more satisfied than men regarding *Material Satisfaction*, if the survey design is taken into account. Also, people aged 22-29 and 40-49 years old are, on average, less satisfied than those in the reference category (16-21 years old), as far as *Material Satisfaction* is concerned. One should note, however, that these differences are no longer significant if the survey design is not considered when estimating the model.

Working longer hours reduces the average *Leisure Satisfaction* in 2002. The perceived health status and the marital status are significant determinants of both average *Leisure* and *Material Satisfaction* in 2002. More children in the household implies a lower average *Leisure Satisfaction* level in 2002, whereas the average *Material Satisfaction* increases significantly with the household income.

5. Discussion

This paper has illustrated how to model perceptions of SWB using conditional LGCM. Modeling within the SEM framework has given us the flexibility to account for measurement error. Results have given us some evidence of the importance of considering the complex sampling design in the analysis of longitudinal data, even when highly sophisticated models are fitted. Indeed, failing to consider the sampling design features could have caused us to wrongly consider a series of categories of covariates as significant when in fact those should not be considered as such (e.g. the impacts of "females" and age groups "22-29" and "40-49" on the *Material Satisfaction* random intercept). We have therefore found variance-inflating impacts of

complex sampling schemes in the longitudinal analyses we have conducted. Future work includes dealing with missing data and including time-varying covariates in the model.

References

Asparouhov, T. (2005). Sampling weights in latent variable modeling, *Structural Equation Modeling*, 12(3), pp. 411-434.

Berrington, A., Hu, Y., Smith, P.W.F and Sturgis, P. (2008). A graphical chain model for reciprocal relationships between women's gender role attitudes and labour force participation, *Journal of the Royal Statistical Society, series A,* 171(1), pp. 89-108.

Bollen, K. A. and Curran, P. J. (2006). *Latent Curve Models – A Structural Equation Perspective*, New Jersey: Wiley.

Duncan, T. E., Duncan, S. C. and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues an Applications*, 2nd ed., New Jersey: Lawrence Erlbaum Associates.

Muthén, B. O. (1998-2004). *Mplus Technical Appendices*, Los Angeles, CA: Muthén and Muthén.

Muthén, L. K. and Muthén, B. O. (1998-2007). *Mplus User's Guide* (5th ed.), Los Angeles, CA: Muthén and Muthén.

Peasgood, T. (2007). Does well-being depend upon our choice of measurement instrument?, paper presented at the British Household Panel Survey Conference, University of Essex, Colchester, UK.

Preacher, K. J., Wichman, A. L., MacCallum, R. C. and Briggs, N. E. (2008). *Latent Growth Curve Modelling*, Thousand Oaks: Sage.

Taylor, M.F. (Ed.), with Brice, J., Buck, N. and Prentice-Lane, E. (2008). *British household panel survey user manual volume A: Introduction, technical report and appendices*. Colchester: University of Essex.

Acknowledgement

The research of M. de Fátima Salgueiro was supported by the Fundação para a Ciência e a Tecnologia, grant PTDC/GES/72784/2006.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

A Latent Transition Analysis Approach to Modeling Unobserved Population Heterogeneity over Time

Andy Ross [1]

Abstract

This paper explores the usefulness of latent transition analysis (LTA) for modeling unobserved population heterogeneity across time. Specifically, we use a latent class framework to capture subgroups of young people disengaged/engaged with education at ages 14 to 16, and model transitions across these various groups over time. Previous quantitative research in this area has often used narrow, single-dimension definitions of disengagement such as truancy or underachievement. Our study goes beyond this, using a statistical approach that captures the multidimensional nature of disengagement by drawing on information from a range of measures. Subgroups of disengaged/engaged young people are defined by their combined responses to questions measuring aspirations, attitudes and behaviour.

The analysis proceeds in stages. In a first step we estimate the latent subgroups for three waves of data. We then employ latent transitions analysis to test the stability of these subgroups and measure transitions across time. In a final step covariates (both fixed and time-varying) are added, measuring characteristics of the individual and their experiences within the home and school in order to explore when and why some young people disengage from education. Data for the study come from the Longitudinal Study of Young People in England (LSYPE) a contemporary panel study of 15,000 young people through years nine to eleven, and into the early years following post-compulsory schooling.

[1] Andy Ross, National Centre for Social Research, U.K. (Andy.Ross@natcen.ac.uk)

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Longitudinal Mixed-Membership Models for Survey Data on Disability

Daniel Manrique-Vallier, Stephen E. Fienberg[1]

Abstract

When analyzing longitudinal data we need to balance our understanding of individual variability with the production of meaningful and interpretable summaries of overall population tendencies. This is specially true when those in the target population are know to be heterogeneous in their ways of progressing over time due to unobserved individual traits. Additional complications arise when the data are discrete and multivariate so that the resulting contingency tables are very sparse. We propose a new family of models to analyze such data by combining features from a version of the cross-sectional Grade of Membership Model (Erosheva et al., 2007) and from the longitudinal Multivariate Latent Trajectory Model (Connor, 2006) and them to data the National Long Term Care Survey (NLTCS), a longitudinal survey with six completed waves aimed to assess the state and characteristics of disability among U.S. citizens age 65 and above. These models assume the existence of a small number of "typical" or "extreme" classes of individuals and model their evolution over time. We regard individuals as belonging to all of these classes in different degree, by considering them as convex weighted combinations of the extreme classes. In this way, we are able to describe distinct general tendencies (the extreme cases) while accounting for the individual variability. We propose a full Bayesian specification and estimation methods based on Markov chain Monte Carlo sampling. We illustrate the our methods using data from the NLTCS.

Key Words: Bayesian Hierarchical Model, Grade Of Membership Model, Latent Trajectories, Markov Chain Monte Carlo.

1. Introduction

In this paper we propose models and estimation procedures to deal with discrete multivariate longitudinal data obtained from a heterogeneous population. This work is motivated by the analysis of data arising from the National Long Term Care Survey (NLTCS), a longitudinal panel survey instrument aimed to assess chronic disability among the elderly population in the United States. Through the analysis of the NLTCS data, researchers seek to answer important questions related to the aging process and disability prevalence in the U.S.: How many elder Americans will live with disabilities? What is the of duration of disability episodes? What is the age of onset of disability? Is it changing for younger generations? (see e.g. Connor et al. (2006)). Answers to these questions are of great importance in public policy design due to, among other reasons, the increased public and private expenditure for disabled people in contrast with their able peers (Manton et al., 1997).

Many of the relevant public policy questions for which the NLTCS can potentially provide answers have to do with changes over time: changes during the life of an individual ("how is this individual likely to age?") or comparing people across different generations ("are people from later generations acquiring disabilities differently than people born 20 years before?"). To answer these questions we need to look at the data longitudinally. In addition, elderly American people are known to be a heterogeneous population, as not everyone could be expected to age the same way. Thus models for longitudinal disability data need to be capable of representing such heterogeneity.

In this paper we present models and methods for their analysis that seek to capture both the longitudinal nature of the NLTCS and the complexity in the heterogeneity of the human aging process, combining the ideas of mixed membership from the Grade of Membership model (Woodbury et al., 1978; Erosheva et al., 2007) and the longitudinal descriptions of the aging process from the Latent Trajectory family (Nagin 1999; Connor, 2006). We illustrate the methods using data from the six waves of the NLTCS.

2. Data - The National Long Term Care Survey

The National Long Term Care Survey (NLTCS) is a longitudinal panel survey aimed at assessing chronic disability among elderly population in the United States. Its target population are people aged 65 years and older that present functional

[1] Daniel Manrique-Vallier, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. USA; Stephen E. Fienberg, Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213. USA.

limitations lasting or expected to last 90 or more days (White, 2008). So far the survey has gone through six waves conducted in 1982, 1984, 1989, 1994, 1999 and 2004.

The sampling frame of the NLTCS is the Medicare record system, which provides a good representation of the elderly population of the U.S. since near 97% of Americans aged 65 or older are included in it (Corder and Manton, 1991). After an initial selection, according to a complex sample design, every individual in the sample is screened to detect if he or she presents a functional limitation. Those who are screened-in are then given a detailed questionnaire, and re-interviewed at each survey wave until they die. Those who were screened-out are re-screened on subsequent waves to check if their functional status has changed. At each wave, a new cohort (approx. 5,000 individuals) is sampled to replace those who died, so that the sample size for each wave is kept at around 20,000 individuals (Clark, 1998). So far 45,009 unique individuals have been interviewed, considering all waves.

The NLTCS approaches disability through the measurement of each individual's capacity to perform a set of six "Activities of Daily Living" (ADL) such as eating, bathing or dressing and ten "Instrumental Activities of Daily Living" (IADL) such as preparing meals or maintaining finances. Broadly stated, ADLs seek to measure a person's ability to take care of him or herself at a fundamental level, while IADLs measure the ability of living independently within a community (Connor, 2006). The survey instrument registers these measurements as a series of answers to triggering questions that are then summarized into a set of binary responses. These binary responses indicate the presence or absence of impairments to perform such activities.

3. Methods

The goal of our analysis is to characterize typical progressions of acquisition of disabilities over time, while taking into consideration and characterizing the heterogeneity of the population. We have proceeded by combining two previously employed methods.

The first method is the latent trajectory model (Nagin, 1999), which is specially well suited in applications where the researcher wants to understand typical evolutions over time and suspects that the population is heterogeneous but a small number of homogeneous classes might exist. Connor (2006) adapted this technique for the analysis of multivariate discrete data and applied it to the NLTCS analysis, identifying latent trajectory curves of probability of acquiring a disability over time. This tool provides a flexible and easy to interpret representation of the data that allows for latent heterogeneity in the population, handling it by clustering the population into *exclusive* classes. In Connor's formulation, this assumption essentially says that, within a class, every single individual responds to the exact same underlying aging process. All the response variability within class is thus attributed to random fluctuations within that class, disregarding the fact that these classes are ideals to which quite possibly no real individuals actually belong (Kreuter and Muthén, 2008).

The Grade of Membership (GoM) family of models (Woodbury et al., 1978; Erosheva et al., 2007) provides a conceptually attractive way relaxing this assumption. Instead of forcing every single individual into one and only one class, the GoM model seeks to identify *pure types* or extreme profiles and then assumes that every individual belongs to more than one of them in different degree. In this way it retains the interpretative power of specifying a reduced number of "typical" or "extreme" profiles but adds extra flexibility by not assuming exclusive membership.

3.1 Notation and setup

We will use the following notation and structure:

1. There are N subjects in the sample, indexed by i, and N_i measurements for each subject $i \in \{1,...,N\}$;
2. For each individual, we measure J binary variables simultaneously in each measurement event. The manifest response vector for individual i and question j is $y_{ij*} = (y_{ij1},...,y_{ijN_i})$ [2];
3. Each individual has an associated covariate vector $X_{i.}$. In this application we will only consider a vector of time dependent covariates $X_{i*} = (X_{i1},...,X_{iN_i})$, although time invariant and other more complicated structures of covariates can also be considered.

[2] Throughout this document we will use this notation when we want to refer to the vector that results from fixing a subset of the sub indexes of an indexed variable while letting the rest to vary. For instance if we have the collection of scalar variables λ_{jk} with $j \in \{1,..,J\}$ and $k \in \{1,...,K\}$ we can write the vectors $\lambda_{*k} = (\lambda_{1k}, \lambda_{2k}, ..., \lambda_{Jk})$, and $\lambda_{j*} = (\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jK})$.

3.2 A Grade of Membership multivariate trajectory model

We start modeling the marginal distribution of the response to question $j \in \{1,...,J\}$ at measurement time $t \in \{1,...,N_t\}$, y_{ijt}, for a full member of extreme profile k (i.e. an individual i such that $g_{ik}=1$ and $g_{ik'}=0$ for $k' \neq k$) as a function of some covariates registered at time t, X_{it},

$$Pr\left(Y_{ijt} = y_{ijt} \mid g_{ik} = 1, X_i\right) = f_{\theta_{j|k}}\left(y_{ijt} \mid X_{it}\right)$$

and model the same marginal distribution of response for a generic individual with membership vector $G_{i*} = (g_{i1},...,g_{iK})$ as the convex combination,

$$Pr\left(Y_{ijt} = y_{ijt} \mid G_{i*} = (g_{i1},...,g_{iK}), X_{i*}\right) = \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt} \mid X_{it})$$

Now, assuming that conditional on the membership vector, g_i, and the covariates, X_{i*}, the responses are independent between items and measurements,

$$Pr\left(Y_{i**} = y_{i**} \mid G_{i*} = (g_{i1},...,g_{iK}), X_{it}\right) = \prod_{j=1}^{J} \prod_{t=1}^{N_i} \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt} \mid X_{it})$$

which combined with the assumption of random sampling gives us the joint model

$$Pr\left(Y_{***} = y_{***} \mid g_{**}, X_{**}\right) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{t=1}^{N_i} \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijl} \mid X_{it})$$

This model is similar to the joint latent class trajectory model proposed by Connor (2006) where we are generalizing the clustering from full membership (i.e. $g_{ik} = 1$ for some k) to mixed membership.

Following Connor (2006), in this implementation we choose the distribution function $f_{\theta_{j|k}}(y \mid X_i)$ for the single response of pure-type individual of extreme profile k as $f_{\theta_{j|k}}(y_{ijt} \mid X_{i*}) = \lambda_{jt|k}(X_{it})^{y_{ijt}}\left(1-\lambda_{jt|k}(X_{it})\right)^{1-y_{ijt}}$ with $\lambda_{jt|k}(X_{it}) = logit^{-1}\left(\beta_{0j|k} + \beta_{1j|k} Age_{it}\right)$, where Age_{it} is the age of the ith individual at measurement time t. Note that under this specification, $\lambda_{jt|k}$ is actually a time dependent function. This specification has the advantage of being relatively simple, with just $2 \times J$ parameters per extreme profile and of representing the intuitively sound notion that the underlying probability of disability is a monotonic (increasing) function of age. Other specifications are certainly possible.

We regard the N membership vectors, g_{i*}, to be iid realizations from a common distribution, G_α, with support on the $K-1$ dimensional unit simplex, Δ_{K-1}. Similar to Erosheva et al. (2007), we model that distribution as $g_{i*} \mid \alpha \sim^{iid} Dirichlet(\alpha)$.

The Dirichlet distribution in this setting has some important properties. In the first place, it is conjugate to the multinomial distribution, facilitating a great deal the computations using Gibbs samplers; second, adopting the re-parametrization $\alpha = (\alpha_0 \cdot \xi_1,...,\alpha_0 \cdot \xi_K)$ with $\alpha_0 > 0$, $\xi_k > 0$ and $\sum_k \xi_k = 1$ we can interpret the vector ξ_* as the average proportion of the population in the k-th extreme profile and α_0 as a parameter governing the spread of the distribution: as α_0 approaches 0, the samples from G_α are more and more concentrated towards the vertices of Δ_{K-1} and; as α_0 increases they are more concentrated near the distribution's average.

As Erosheva et al. (2007) and Airoldi et al. (2007) discuss, a priori setting the parameters α for the Dirichlet distribution might be too strong an assumption to do realistic modeling. We will estimate these parameters from the data specifying hyper-priors and computing posterior distributions. For this purpose we use hyper priors for α_0 and ξ_* similar to the ones in Erosheva (2002) and Erosheva et al. (2007): $\alpha \sim Gamma(\tau,\eta)$, $\xi \sim Dirichlet((1,...,1)_k)$ (Uniform over Δ_{K-1}) and complete the specification with the priors $\beta_{0j|k} \sim^{iid} N(\mu_0, \sigma_0^2)$ and $\beta_{1j|k} \sim^{iid} N(\mu_1, \sigma_1^2)$, with β_0 independent from β_1.

4. Results

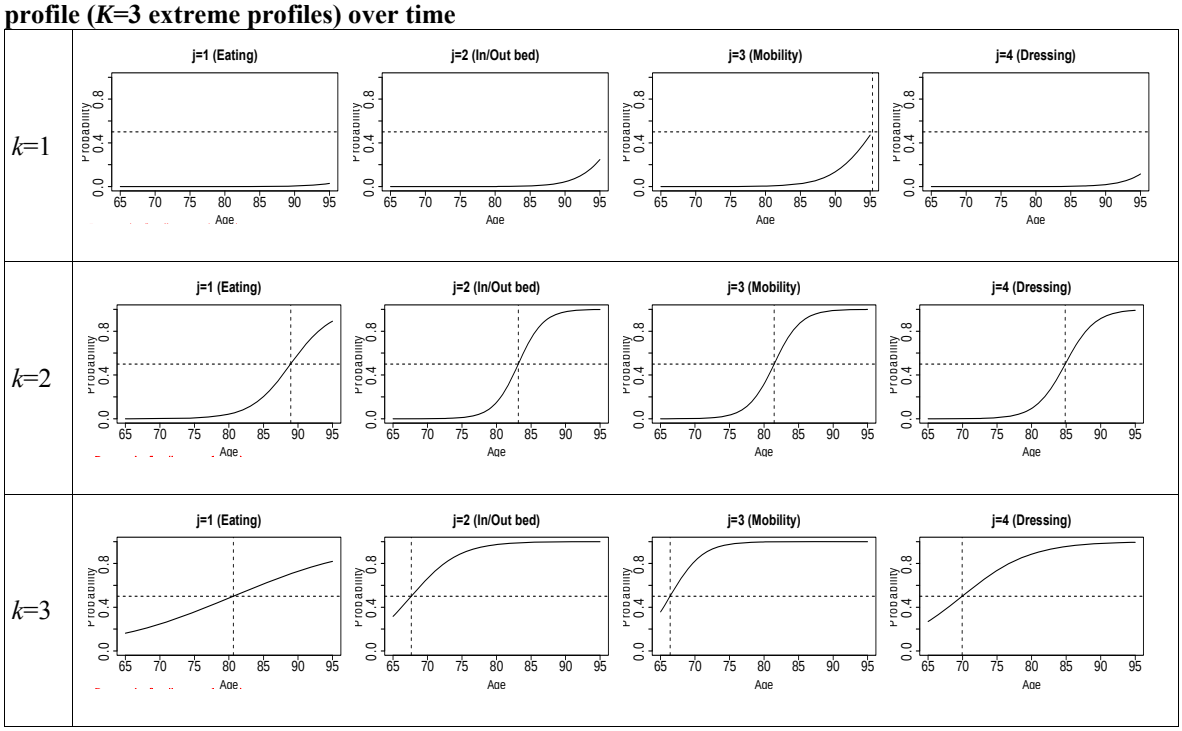
To test our methods, we have applied them to data from six waves from the NLTCS (N=45,009), analyzing the responses for the six ADLs[3] (J=6) using a total of K=3 extreme profiles. The posterior estimation has been performed using a custom Markov chain Monte Carlo (MCMC) algorithm based on a latent class representation first proposed by Erosheva (2002) and Erosheva et al. (2008). For this application we have chosen the priors $\beta_{0j|k}, \beta_{1j|k} \sim^{iid} N(0,100)$ and $\alpha_0 \sim Gamma(1,5)$.

Figure 4-1 shows the curves of the estimated posterior probability of acquiring a disability in each ADL (only four are shown) as a function of age, for the *ideal* members of the three extreme profiles, similar to the ones in Connor (2006). They appear reflect some desirable features: the slopes of all extreme profiles are positive, showing the expected increasing tendency of the probability of suffering a disability as time passes and the ages where the idealized individuals of the extreme profiles reach a probability of 0.5 of acquiring a disability are within reasonable ranges. Also, the posterior distribution of the parameters are quite concentrated around a central value (not shown), opposed to their prior specification. From the picture, we can see that the method has identified three well separated profiles that reflect quite different aging processes: a class of people that live relatively healthy until very late (k=1); a class of people that remain healthy until around the age of 85, when they experiment a sudden increase in their probability of acquiring disabilities (k=2) and; a class of people that have an early increase of the probability of getting disabled (k=3).

Table 4-1 shows point estimates (posterior means) for parameters α_0 and ξ, reflecting how particular individuals age, opposed to the ideal ones. Parameter estimate $\xi_* = (0.65, 0.25, 0.1)$ indicates the relative order of importance of each of the three extreme profiles, showing that more people are closer to profile k=1, followed by k=2 and k=3. Parameter estimate $\alpha_0 = 0.264$ indicates that the distribution over the population is quite concentrated towards the vertices of the simplex Δ_{K-1}, although not as much to make the model behave like a regular mixture model. Figure 4-2 shows an example of how the individual trajectories are formed from the extreme profiles. As can be seen, the model is quite flexible, allowing quite varied individual trajectories, but extracting just a few simple extreme curves that are easy to characterize and interpret.

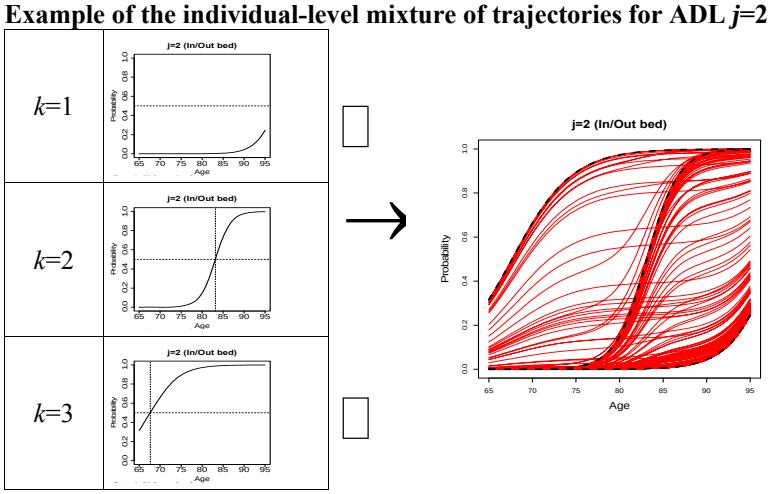
[3] The six ADLs are: Eating (j=1), getting in or out of bed (j=2), inside mobility (j=3), dressing (j=4), bathing (j=5), toileting (j=6).

Figure 4-1
Trajectories of the probability of acquiring a disability in the first four ADLs, for ideal members of each extreme profile (*K*=3 extreme profiles) over time



The intersection of the straight lines indicates the point where the probability of acquiring a disability for ADL *j* reaches 50%.

Figure 4-2
Example of the individual-level mixture of trajectories for ADL *j*=2 (getting in or out of bed)



The plot on the right superimposes the three extreme trajectories and presents a sample of 100 individual trajectories.

Table 4-1
Posterior means for population-level parameters for model with K=3 extreme profiles

Parameter	**Estimate [sd.]**
α_0	0.264 [0.00489]
(ξ_1, ξ_2, ξ_3)	(0.65 [0.004], 0.25 [0.003], 0.104 [0.002])

Numbers between brackets are posterior standard deviations.

5. Discussion

Our preliminary results are interesting because they show the potential of our methods. In our application, using longitudinal data from the NLTCS, we have been able to characterize well separated and intuitively sound extreme profiles that can be understood as typical ways of aging, while at the same time characterizing the heterogeneity of the population using a very simple device that allows to construct individualized curves from the extreme profiles. This way of handling heterogeneity, although slightly more complicated that the one proposed in Connor (2006), allows us to be able to keep the number of extreme profiles low and interpretable while avoiding the introduction of too strong in-class homogeneity considerations.

The model presented in this paper is a basic implementation of the general idea of combining complete time dependent trajectories using a mixed membership device. Depending on the problem at hand, there are a number of obvious extensions that can be worked out, some of which we are developing at the moment. In our application, the the NTLCS, some of these natural extensions are the inclusion of other covariates at the group membership level and at the extreme profile level and the joint formulation with survival models to study the relationship of disability and mortality. Many of these extension will be included in Manrique-Vallier (2010).

For purposes of illustration, we have chosen to illustrate the methodology with K=3 extreme profiles. More generally, we need to incorporate methodology for deciding on an optimal value of K. We have carried out full computation for a series of values of K, running from 2 through 5. While the fit of the model, as measured in terms of the posterior predictive responses, increases with K, we observed less separation of profiles for K=4 and K=5, and a less satisfactory interpretation of the shape and structure of the profiles. Choosing an appropriate value of K remains an open problem in our work that will be addressed in Manrique-Vallier (2010).

References

Airoldi, E., Fienberg, S., Joutard, C. and Love, T. (2007). Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models, *Data Mining Patterns: New Methods and Applications*, 240–275.

Clark, R. (1998). An Introduction to the National Long-Term Care Survey, Office of Disability, Agigng, and Long-Term Care Policy within the U.S. Department of Health and Human Services.

Connor, J., Fienberg, S., Erosheva, A. and White, T. (2006). Towards a restructuring of the national long term care survey: A longitudinal perspective, Tech. rep.

Connor, J. T. (2006). Multivariate Mixture Models to Describe Longitudinal Patterns of Frailty in American Seniors, Ph.D. thesis, Department of Statistics & H. John Heinz III School of Public Policy & Management. Carnegie Mellon University.

Corder, L. and Manton, K. (1991). National surveys and the health and functioning of the elderly: the effects of design and content, *Journal of the American Statistical Association*, 86, 513–525.

Erosheva, E. (2002). Grade of membership and latent structures with application to disability survey data, Ph.D. thesis, Department of Statistics. Carnegie Mellon University.

Erosheva, E., Fienberg, S. and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data, *Annals of Applied Statistics*, 1, 502–537.

Kreuter, F. and Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling, *J Quant Criminol*, 24, 1–31.

Manrique-Vallier, D. (2010). Longitudinal Mixed Membership Models With Applications, Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, expected May 2010.

Manton, K., Corder, L. and Stallard, E. (1997). Chronic disability trends in elderly United States populations: 1982-1994, *Proceedings of the National Academy of Sciences*, 94, 2593–2598.

Nagin, D. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach, *Psychological Methods*, 4, 139–157.

White, T. (2008). Extensions of Latent Class Transition Models with Application to Chronic Disability Survey Data, Ph.D. thesis, University of Washington.

Woodbury, M., Clive, J. and Garson Jr, A. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers in Biomedical Research*, 11, 277–98.

MEASUREMENT ERRORS

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Nonresponse and Measurement Error in Employment Research

Frauke Kreuter, Gerrit Mueller and Mark Trappmann[1]

Abstract

Survey methodologists are increasingly concerned with the interaction of multiple error sources. Particularly prominent are discussions about nonresponse and measurement error. One hypothesis that is often found among practitioners is that sample cases that are brought into the survey only after repeated attempts and alternated recruitment strategies, are more likely to provide low quality data (e.g. Groves and Couper 1998). Data quality is often internally assessed through the proportion of missing items, proportion of don't knows and the like (e.g. Fricker 2007). Rarely, in these studies, are external data available to evaluate the quality of respondents' answers (e.g. Cannell & Fowler 1963, Olsen 2006).

The panel study PASS (Trappmann et al. 2009) is a novel dataset in the field of labor market, welfare state and poverty research in Germany. With almost 19,000 interviewed persons in more than 12,500 households, PASS is currently one of the most comprising panel surveys in Germany. The first round of data collection started in 2006. In PASS, survey data on the employment and unemployment history, income and education of participants can be linked to corresponding data from respondents' administrative records.

Based on this study, we give an assessment of data quality as a function of contactability and response propensity. Only for some variables, the measurement error (variance or bias) assessed through the administrative records is increased with decreasing contactability and response propensity of the target persons. In particular, this is found in case of retrospective questions. Here, the differing length of time between date of interview and event explains a large part of the difference in measurement error between respondents with high vs. low response propensity.

[1] Frauke Kreuter, JPSM University of Maryland, U.S.A.; Gerrit Mueller (Gerrit.Mueller@iab.de) and Mark Trappmann, IAB Institute for Employment Research, Germany.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Inconsistencies in Reported Job Characteristics among Employed Stayers: Evidence from a Series of Two-Wave Panels from the Italian Labour Force Survey, 1993-2003

Francesca Bassi, Ugo Trivellato and Alessandra Padoan[1]

Abstract

In this paper we deal with measurement error, and its potentially distorting role, in information on industry and professional status. As a case study we consider two-wave panels one year apart collected by the Italian Quarterly Labour Force Survey in the period from April 1993 to April 2003. The focus of our analyses is on inconsistent information on employment characteristics – industry and professional status – resulting from yearly transition matrices for workers who reported that they were continuously employed over the year and did not change job.

First, we compute and comment upon some usual indicators of disagreement. We find clear evidence that there is sizable measurement error in both industry and professional status. Then, we test whether the consistency of repeated information significantly increases when the number of categories is collapsed. Aggregating categories improves agreement. For professional status the best level of aggregation is the binary one (Employee/Self-employed); for industry, two classifications minimize inconsistencies, with 5 or 6 classes. We further explore the patterns of inconsistencies among categories of variables by testing several specifications of Goodman's quasi-independence model. The model is almost always rejected, which points to the fact that even cross-section information is affected by non-random measurement error. Lastly, we consider and compare alternative 4-category classifications obtained by collapsing professional status and industry into a single variable. In this case the best level of aggregation is given by a non-standard 4-category classification, which distinguishes employees in the market services on one hand and in the industrial sector and private services on the other.

[1] Francesca Bassi and Ugo Trivellato (trivell@stat.unipd.it), University of Padova, Italy; Alessandra Padoan, ISTAT, Italy.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Challenges and Insights from Overlapping Seams in the HILDA Survey

Nicole Watson[1]

Abstract

An issue unique to longitudinal surveys is seam effects. These occur when there is a tendency for changes in the data to unusually concentrate in adjoining periods from different interviews. One component of the Household, Income and Labour Dynamics in Australia (HILDA) Survey subject to seam effects is the labour market activity calendar which asks respondents to recall spells over a 14 to 18 month period. As the interviews are conducted annually, this results in an overlap of 2 to 6 months that we can use to study inconsistencies in the two reports. The characteristics considered in modeling the likelihood of inconsistent reports include the various causes of errors in dating events, such as spell length, spell type, duration of the overlapping seam, recall ability of the respondent, and characteristics of the interview that may affect the respondent's recall. The overlapping seam also permits the study of measurement error over time to identify whether the same people continually make the same mistakes

Key Words: Measurement Error, Employment Calendar, Data Quality, Longitudinal Surveys.

1. Introduction

Longitudinal studies often incorporate questions that require the respondent to report activities over the intervening period between the current interview and the previous interview. Inconsistencies in their recall compared to the data from the previous interview gives rise to seam effects. These occur when there is a tendency for changes in the data to unusually concentrate in adjoining periods from different interviews (Tourangeau *et al.*, 2000). It is not uncommon for the size of transitions across adjoining periods (the seam) to be between two to eight times the size of the transitions that are not at the seam (for example, Burkhead and Coder, 1985; Tourangeau *et al.,* 2000; Lemaître, 1992).

This paper seeks to contribute to our understanding of why seam effects occur using data from the labour market activity calendar in the HILDA Survey. The calendar collects information from a fixed point in the previous year to the date of interview, resulting in two reports for the same portion of time (called the 'overlapping seam'). This overlapping seam gives us the opportunity to investigate a number of matching methods and study the differences in the two reports.

2. Factors resulting in inconsistent spell reports

During an interview, there are a number of errors that can occur which can result in inconsistent reports. These include omitting spells, misclassifying spells, bringing the spell forward or back in time (telescoping) especially at the beginning of the reference period, misunderstanding what is required or making incorrect inferences when the spell cannot be remembered clearly. These errors can occur in either the first or the second report of the spell, though we would expect more errors in the second report as this is recalled at greater distance from the spell.

The factors that affect these spell recall errors can be divided into five categories:

- *Spell characteristics* – Shorter spells are more likely to be dropped or misplaced (Paull, 2002; Jäckle, 2008). Events with clearly defined dates are reported more accurately than those with fuzzy dates, such as transitions out of the labour force (Jäckle, 2008).
- *Complexity of the recall* – The difficulty of the recall task and the accuracy of the inference strategies that respondents use when they cannot remember exactly affect the size of the seam effect (Callegaro and Belli, 2007).
- *Respondent characteristics* – While there are mixed results across the various studies, the likelihood of inconsistent reports tends to vary by age, sex, education levels (Hill, 1987; Jäckle and Lynn, 2007; Hill, 1987; Callegaro and Belli, 2007; Jäckle, 2008).
- *Interview characteristics* – Interviewer continuity may help reduce inconsistent reports, either by consistent application of interviewing techniques (Vick and Weidman, 1989) or because their very presence triggers the

[1] Nicole Watson, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (n.watson@unimelb.edu.au).

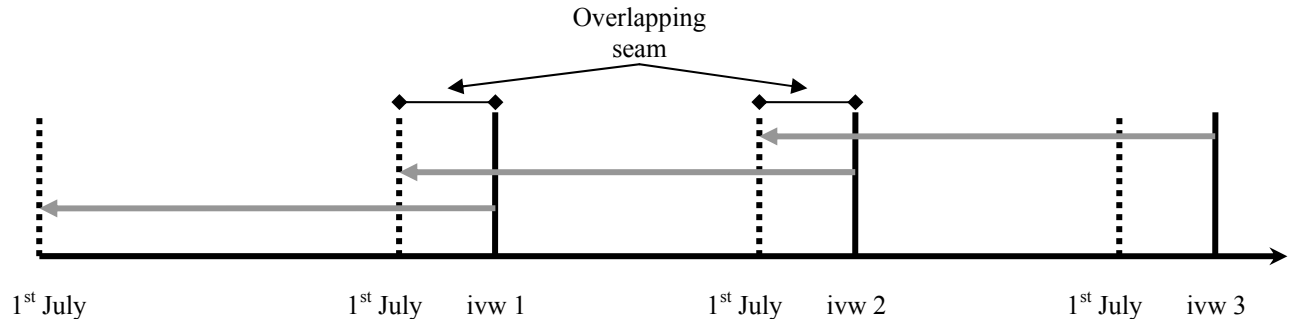
respondent's memory. The quality of the respondent's answers also depends on their understanding of the questions and their willingness to retrieve and interpret the required information from memory (Tourangeau *et al.*, 2000; Burkhead and Coder, 1985).

- *Survey process characteristics* – Seam effects can also result from changes or errors in the survey process. The question format, wording and order are important for consistent interpretation over time (Burkhead and Coder, 1985; Jäckle, 2008; Callegaro and Belli, 2007). Transcription or data entry errors may also occur (Burkhead and Coder, 1985; Lemaître, 1992).

3. The HILDA calendar

We now examine the inconsistencies in the labour market activity calendar of the HILDA Survey over time.[2] The HILDA Survey is an Australia-wide household panel survey, interviewing around 7000 households and 13,000 individuals each year (Watson and Wooden, 2004). The survey began in 2001 and, each wave, respondents are asked to recall the various jobs they have had, time spent in unemployment and time spent not in the labour force. In addition, spells of full-time and part-time education are also collected (though we do not discuss these spells in this paper for space reasons). Start and finish times of each spell are recorded by whether they are at the start, middle or end of each month. The calendar covers a 12 to 18 month period from 1st July of the preceding year to the date of the current interview (see Figure 3-1).[3] All job spells are included on the calendar (not just the main job). Where two or more jobs occur at the very beginning of the calendar, their start dates are also collected to assist matching across waves. As the calendar is administered every wave, an overlap of 2 to 6 months results, depending on when the respondent was interviewed (with the average being 2.9 months).

Figure 3-1
Overlapping seams in HILDA



4. Matching methods

To investigate the value of the overlapping seam, three matching methods are evaluated:

i) Match the jobs at 1st July;
ii) Match the jobs at the date of last interview (giving precedence to the information collected closest to the period being recalled);
iii) Reconcile spell reports between the two versions of events reported approximately one year apart.

In the first two matching methods, the alternative report of the same period is ignored. Where multiple spells from one wave could be matched to one spell in another (as occurs with job spells), a match was randomly chosen. Spells are matched within each type of spell (being job, unemployment, and not in the labour force).

The third matching method uses the two reports to produce a reconciled view of the overlapping period. The first report (recalled closest to when the events occurred) is taken in precedence over the second. Nevertheless, spells recorded in the second report are matched with those of the same type in the first report to identify (and remove) spells that were misplaced. This method cannot resolve situations where a spell in the first report is not recalled in the second or visa versa. A match score was created to identify which spells should be matched. The match score is the sum of whether the spell starts match exactly within the overlapping seam, the spell starts match within one month, the spell ends match exactly, the spell ends match within

[2] The dataset used for the analysis in this paper is the In-Confidence Release 7 HILDA data.
[3] Interviews are conducted from mid August to March, though the last period on the calendar is in December (96 to 98 per cent of the interviews completed by then).

one month and 3 times the ratio of the length of the spells (with the longest as the denominator). The maximum match score is, therefore, 7. Spells with the highest match score are matched first. The remaining spells are then matched in the same way until no more spells can be matched. Spells with a match score of less than 3 are not matched.

5. Results

5.1 Size of the seam effect

The number of job transitions at the seam for spells matched at the start of July is around eight times those off-seam. For the other two methods, the seam is at the date of interview of each respondent, so spreads from mid-August to December resulting in an elevation in the number of job transitions over this period. Figure 5-1 shows the (unweighted) number of job starts centred at the seam for those interviewed all seven waves. The black line shows the number of job starts when the spells are matched at the start of July (LFY method). The light grey line shows the results for when the spells are matched at the date of last interview (DOLI method) and the (barely visible) medium grey line is when the spells are reconciled (OLAP method, which for the most part, behaves very much like the DOLI method). Further, for clarity, the counts of job starts (and ends) at the seam for each method are also reported in Table 5-1.

Several observations from this figure and table are:

- Respondents tend to report fewer job starts at the beginning of the calendar (this can be seen in the results for the LFY method in the first few months after the seam), and as a result, the DOLI method has slightly more job starts at the seam than the LFY method. This suggests either a decay in the memory over time or there is backwards telescoping of events to the boundary of the reference period (a similar effect was found by Kalton and Miller (1991) in an analysis of the US Survey of Income and Program Participation). When spells are taken from the respondent's first report of the events (as in the DOLI and OLAP methods), this difference is eliminated.
- The OLAP method has more job starts at the seam than the DOLI method. When comparing the spells in the two versions of the overlapping seam, it is apparent that some of the spells matched in the DOLI method are actually not the same, so the spells are not connected in the OLAP method, resulting in more job ends and job starts at the seam.

Figure 5-1
Number of job starts, centred at seam between each wave

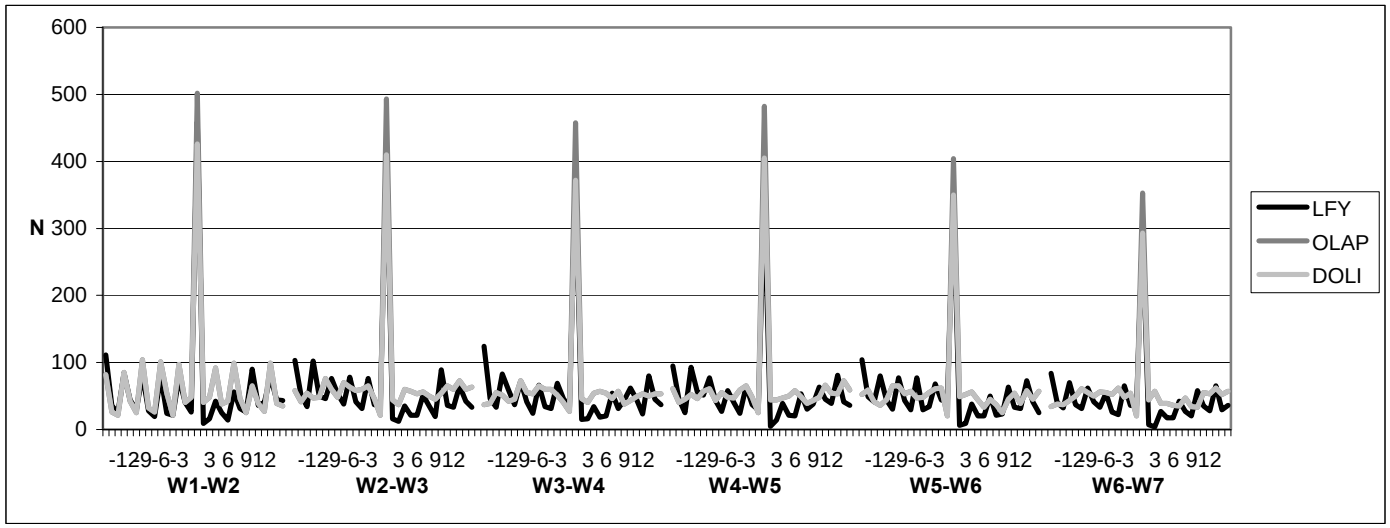


Table 5-1
Number of job starts and ends at the seam, by wave and method

	Start			*End*		
	LFY method	*DOLI method*	*OLAP method*	*LFY method*	*DOLI method*	*OLAP method*
W1-W2 seam	459	426	502	485	432	526
W2-W3 seam	407	410	493	465	416	512
W3-W4 seam	368	372	458	467	387	474
W4-W5 seam	347	405	482	437	389	469
W5-W6 seam	342	350	404	512	423	495
W6-W7 seam	298	293	353	480	421	495
Total	2,221	2,256	2,692	2,846	2,468	2,971

- The number of job starts at the seam appears to decline over time, but there is not a corresponding decline in the number of job ends at the seam. It is not apparent what might be causing this.

It is clear that the DOLI and OLAP methods are better than the LFY method as they avoid the drop in the number of spells that start or end at the beginning of the reference period. The OLAP method is better than the DOLI method as it helps identify spells that should not be matched in the overlapping period.

5.2 Characteristics associated with recall errors

The overlap period is used to identify dropped spells (in the first report but not the second), misplaced spells, and added spells (that were in the second report but not the first). The spell-level dataset is pooled across the first seven waves of the HILDA Survey and is restricted to individuals interviewed all waves. Only those spells occurring in the six overlapping seams are included in the analysis. Three separate logistic models were used to estimate the likelihood that a spell would be dropped, misplaced or added. The pooled dataset of the first report contains 64,085 spells on 8409 individuals and is used to estimate the likelihood of dropping or misplacing spells. A second pooled dataset of the second report contains 59,981 spells on 8409 individuals and is used to estimate the likelihood of adding spells. To allow for the repeated measures on the same individuals, the estimated standard errors assume the spell outcomes are correlated across observations on the same individual but are independent across individuals.

As per section 2, the probability that a spell is dropped, misplaced or added is assumed to be a function of five groups of characteristics. The *spell characteristics* include spell length and whether the same type of spell was current at the second interview date. The *complexity of the recall* was estimated by the number of spells between interview dates and the time between interviews. The length of the overlap was included as a control. The *respondent characteristics* include sex, age (15-24, 25-54, 55 and over), education level, financial year personal income, and country of birth. The *interview characteristics* incorporated the interviewer's assessment of the respondent's understanding of the questions, whether it was a telephone interview, whether the same interview conducted both interviews and the experience of the interviewer (measured by the number of interviews completed that wave). The *survey process characteristic* included was whether the seam was between waves 1 and 2 (as there was a change in the calendar design between these waves).

The mean marginal effects estimated from the models are presented in Table 5-2 with the following observations:

i) The single most important factor in reducing the likelihood of providing an inconsistent report is if the same type of spell was current in the subsequent interview. For job spells, it reduces the chance of dropping a spell by 23%, misplacing a spell by 2% and adding a spell by 10% (all other things being equal).

ii) Consistent with previous research, shorter spells are more likely to be dropped, added or misplaced.

iii) Generally, the more complex the recall task the greater chance of an inconsistent report. Contrary to expectations, we find that unemployment and not in the labour force spells are *less* likely to be dropped if many spells are recalled between interview dates.

iv) Some age and sex effects were present – inconsistent reports were more likely with spells more unusual for the respondent's stage in the life cycle. Education and country of birth often had no effect.

v) Individual incomes had a mixed effect. With higher individual incomes, the likelihood of dropping or adding job spells decreased, but for spells not in the labour force it increased.

vi) Some support was found for face-to-face interviewing and continuity of the interviewer to reduce the likelihood of inconsistent reports. The experience of the interviewer was not significant.

vii) Respondents who seemed to have a good understanding of the questions asked in the entire interview were less likely to provide inconsistent reports.

viii) Unexpectedly, we found that job spells were *less* likely to be dropped or misplaced in the overlapping period between waves 1 and 2. It is not clear why this might be the case.

The above estimated models do not let us determine whether respondents make the same mistakes over time. To test this, the pooled dataset was restricted to the overlapping seams for waves 2 and 3 through to waves 6 and 7 and an additional variable indicating whether the respondent had made the same mistake (of dropping, misplacing or adding a spell) in the previous seam. We find that dropping a spell in the previous overlapping period increases the risk of dropping a spell in the following overlapping period – for job spells, the chance of dropping a spell increased by 2%, for unemployment spells it increased by 5% and for spells not in the labour force it increased by 2%. Similarly, if a spell is added in the previous overlapping period, the chances of adding a spell in the next overlapping period is higher for most types of spells, however, misplacement of spells in one overlapping period does not seem to have any bearing on whether spells are misplaced in the next overlapping period.

Table 5-2
Mean marginal effects for characteristics associated with inconsistent reports in labour force spells

	Job spells			*Unemployment spells*			*Not in labour force spells*		
Variable	*Drop*	*Misplace*	*Add*	*Drop*	*Misplace*	*Add*	*Drop*	*Misplace*	*Add*
Spell characteristics									
Spell length (thirds of a month)	-0.003***	-0.003***	-0.001***	-0.005***	-0.006***	0.001	-0.002***	-0.002***	-0.001***
Same type of spell at date of interview	-0.231***	-0.018***	-0.105***	-0.343***	0.051**	-0.152***	-0.476***	0.005*	-0.322***
Complexity of recall									
Number of spells btw ivws	0.004***	0.003***	0.018***	-0.032***	0.010*	0.021*	-0.014***	0.003***	-0.002
Time between last and current ivw date	0.002***	-0.001**	0.001	-0.002	-0.002	0.003	0.001	0.000	0.002*
Length of overlap	0.005***	0.004***	0.003***	0.001	0.005	0.018**	0.002	0.003***	0.003***
Respondent characteristics									
Sex and age (base = younger males aged 15-24)									
Prime males (aged 25-54)	-0.011*	0.009**	-0.004	-0.013	-0.005	-0.023	0.020	-0.002	-0.042
Older males (aged 55 or older)	0.019*	0.000	0.029***	0.084*	-0.045	0.021	-0.005	0.000	-0.069***
Younger females	0.007	0.009	-0.016	0.081*	-0.029	0.067	-0.010	0.030	-0.029
Prime females	-0.005	0.003	0.007	0.066*	-0.006	0.071	-0.011	0.015	-0.086***
Older females	0.011	0.011	0.034***	0.183***	-0.055*	0.182***	-0.008	0.001	-0.070***
Education level (base=year 11 or below)									
Year 12	0.002	0.001	0.000	0.053	-0.018	0.054	0.003	0.003	0.002
Certificate	0.009*	0.001	-0.005	0.022	-0.032*	0.013	0.004	0.007***	0.009*
Diploma	0.007	-0.003	-0.002	0.054	-0.045*	0.062	-0.001	0.001	-0.003
Graduate	0.015***	0.000	-0.001	-0.010	-0.021	0.037	0.003	0.001	0.011**
Financial year income ($/10^5$)	-0.039***	0.005	-0.034***	-0.206	0.109	0.043	0.062***	-0.001	0.069***
Financial year income squared ($/10^{10}$)	0.004**	-0.002	0.003***	0.065	-0.037	0.008	-0.010***	-0.004	-0.015*
Country of birth (base = Australia)									
Main English-speaking country	-0.009*	-0.001	-0.005	0.007	0.026	0.034	0.001	0.002	0.001
Other overseas country	0.009	-0.005	0.007	0.004	0.014	-0.024	-0.003	0.006	0.000
Interview characteristics									
Understanding of questions was excellent or good	-0.014	0.002	-0.028***	0.011	-0.017	0.001	-0.020***	0.004	0.003
Telephone interview	0.013*	-0.001	0.000	0.026	0.022	-0.024	0.008	-0.005	0.022*
Continuity of interviewer	-0.007**	-0.004*	-0.002	-0.018	0.000	-0.030	-0.003	0.000	-0.009**
Experience of interviewer (calendars done) ($/10^2$)	-0.003	0.003	-0.004	0.018	-0.002	-0.030	0.002	-0.001	-0.007
Survey process characteristics									
Seam between wave 1 and 2	-0.012**	-0.010***	0.007	-0.046	0.001	0.067*	0.001	0.003	0.001
Pseudo R^2	0.134	0.357	0.120	0.123	0.137	0.039	0.518	0.362	0.355
N spells	36,383	36,383	34,529	1,771	1,771	1,412	16,843	16,843	16,744
Proportion of spells with inconsistency	0.119	0.049	0.080	0.644	0.119	0.495	0.117	0.029	0.097

Note: * significant at 10 per cent, ** significant at 5 per cent, *** significant at 1 per cent.

6. Conclusion

The overlapping seam collected in the HILDA labour market activity calendar has not resolved the problem of the seam effect. By matching the spells at the date of last interview, each respondent effectively has their own seam and this spreads the seam effect out across the interviewing period. This analysis has shown that matching at the date of last interview is better than at the last financial year as this avoids the problems of low recall of spell starts and ends in the first few months of the calendar. The method which attempts to reconcile the spells across the seam identifies spells which are not actually the same and, even though it produces a higher number of dropped or added spells, the resultant spell file is more accurate.

The spells which are most subject to recall error are spells that are unlike those reported at the current date of interview, short spells, and those spells which are part of a complex history. Some limited support was found in this study for reduced recall error when the interview is conducted face-to-face and the interview is the same between waves. The effect that the respondent characteristics had on the likelihood of recall errors varied by the type of spell recalled. We also found evidence that respondents tend to make the same recall mistakes over time in terms of dropping or adding spells, but not in misplacing spells.

Collection of overlapping spells permits the analysis of inconsistent reports which can help users of the data better understand the data's limitations. This would not have been possible had we employed dependent interviewing techniques which may sometimes force consistency over accuracy.

References

Burkhead, D. and Coder, J. (1985). Gross Changes in Income Recipiency from the Survey of Income and Program Participation, *Proceedings of the American Statistical Association, Social Statistics Section*, Washington, DC, pp. 351-356.

Callegaro, M. and Belli, R.F. (2007). Impact of Event History Calendar on Seam Effect in PSID: Lessons for SIPP, paper presented at Event History Calendar Methods Conference, Dec 5-6, Washington, DC.

Hill, D.H. (1987). Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods, *Proceedings of the American Statistical Association*, Section on Survey Research Methods, Washington, DC, pp. 210-215.

Jäckle, A. (2008). The Causes of Seam Effects in Panel Surveys, Institute for Social and Economic Research Working Paper Series, No. 2008-14.

Jäckle, A. and Lynn, P. (2007). Dependent Interviewing and Seam Effects in Work History Data, *Journal of Official Statistics*, Vol. 23, No. 4, pp. 529-551.

Kalton, G. and Miller, M.E. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation, *Journal of Official Statistics*, Vol. 7, No. 2, pp. 235-245.

Lemaître, G. (1992). Dealing with the Seam Problem for the Survey of Labour and Income Dynamics, in SLID Research Paper Series, Statistics Canada, Ottawa.

Paull, G. (2002). Biases in the Reporting of Labour Market Dynamics, The Institute for Fiscal Studies Working Paper Series, No. 02/10.

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). *The Psychology of Survey Response*, Cambridge: Cambridge University Press.

Vick, S. and Weidman, L. (1989). Reporting on Income Recipiency by Self and Proxy Respondents in SIPP, *Proceedings of the American Statistical Association*, Section of Survey Research Methods, Washington, DC, pp. 315-319.

Watson, N. and Wooden, M. (2004). The HILDA Survey four years on, *Australian Economic Review*, Vol. 37, pp. 343-349.

IMPUTATION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Donor-Imputation and Weighting in Presence of Non-response Under a Model-based Approach

Roberto Gismondi[1]

Abstract

In this paper we formalise an estimation strategy based both on non responses' deterministic donor imputation and optimal model-based estimation in a sampling survey frame. We develop the mean squared error of the estimator based on donor imputation and carry out a theoretical comparison of its efficiency with respect to the estimator based on re-weighting of respondent units without imputation. We discuss the conditions – depending on both the theoretical sampling rate and the response rate – for which donor imputation may improve the ordinary pseudo-optimal model based predictor. Finally, we present outcomes of an empirical comparison among donor imputation, model based prediction and calibration based on real turnover data

Key Words: Donor, Imputation, Nearest Neighbour, Non-Response, Weighting.

1. Non-response treatment in business surveys

In a sampling survey frame, adjustments for tackling non responses are aimed at reducing the potential *non response bias* (Billiet *et al*. 2007). It often depends on a model misspecification, for instance because respondents and not respondents follow different patterns. Late experiences (Gismondi, 2008) showed that in many real business surveys contexts the non response bias is not systematic, but could happen for some survey occasions and/or for some domains only. Performances of traditional strategies for reducing non response bias are often poor, for instance because too few auxiliary variables are available at the estimation stage (Rizzo *et al*., 1996). In particular, the most part of imputation techniques do not reduce bias enough to balance the increase of variance due to imputation (Copeland and Valiant, 2007).

In this context, according to a model-based approach (Valiant *et al*., 2000), first we resume the basic rules for carrying out prediction of the total in presence of non responses through re-weighting of respondents, under the hypothesis that non response bias can be neglected (section 2). Afterwards (section 3) we formalise an estimation strategy based both on non responses' deterministic donor imputation and optimal model-based estimation through re-weighting of sample units. Donor imputation is frequently used in surveys, but very few variance estimations have been developed (Beaumont and Bocci, 2009). We derive the mean squared error of the estimator based on donor imputation and carry out a theoretical efficiency comparison with respect to re-weighting of respondent units without imputation. We point out the conditions – depending on both sampling rates and response rates – for which donor imputation may improve the ordinary pseudo-optimal re-weighting. Finally, we present main outcomes of an empirical study (section 4) where donor imputation, model based prediction and calibration have been compared.

2. Model-based imputation and weighting

2.1 Estimation without non responses

For each unit i of a population U including N units we suppose the model $y_i=\beta x_i+\varepsilon_i$, with $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2 v_i$, $Cov(\varepsilon_i,\varepsilon_j)=0$ if $i\neq j$, β is unknown, x is known for all the units in U and v must be specified.
If we also suppose a sample S with theoretical size n drawn from U, the *BLU* predictor – which minimises the *Mse*: $E(T-N\bar{y})^2$ – is (Cicchitelli *et al*., 1992, pp.385-387):

$$T^* = N[f\,\bar{y}_S + (1-f)\,\bar{x}_{\bar{S}}\,\hat{\beta}^*] \qquad \text{where:} \qquad \hat{\beta}^* = \left(\sum_S \frac{x_i y_i}{v_i}\right)\left(\sum_S \frac{x_i^2}{v_i}\right)^{-1} \tag{1}$$

[1]Roberto Gismondi, ISTAT, Via Tuscolana 1788, 00173 Roma, Italy, gismondi@istat.it

with $f=n/N$ and $\bar{y}_S$, $\bar{x}_{\bar{S}}$ are, respectively, the sample y-mean and the x-mean referred to the not observed units. If we label with j a generic not observed unit, the model *Mse* of T^* based on n observation is:

$$Mse(T^*) = \sigma^2 \left[\left(\sum_{\bar{S}} x_j \right)^2 \Big/ \left(\sum_{S} (x_i^2 / v_i) \right) + \sum_{\bar{S}} v_j \right] \qquad (2)$$

The previous estimation strategy can be generalised, since it is equivalent to carry out a deterministic imputation of each not observed unit through the rule:

$$\hat{y}_j = \hat{\beta} x_j \qquad \text{with:} \qquad \hat{\beta} = \sum_S a_i y_i \qquad \text{for each unit } j \in \bar{S} \qquad (3)$$

where the n coefficients a_i must be specified. If X_Δ is the sum of x_i extended to the domain Δ, the imputation rule defined by (3) is equivalent to the weighting rule $w_i = 1 + a_i X_{\bar{S}}$. If we put:

$$a_i = a_i^* = \frac{x_i}{v_i} \left(\sum_S x_i^2 / v_i \right)^{-1} \qquad (4)$$

we come back to the optimal unbiased predictor (1) with mean squared error (2). The use of imputation rules (or particular re-weighting systems) leading to predictors alternative to T^* is justified when this predictor may be biased, for instance because: a) all and only the units not included in the observed sample follow a different model; b) the observed sample includes some units which follow the supposed model, but also some units following another model to be specified. We can define a general predictor of the total as:

$$T = \sum_S y_i + \sum_{\bar{S}} \hat{y}_j = \sum_S y_i + \sum_{\bar{S}} \hat{\beta} x_j \,. \qquad (5)$$

Its mean squared error can be written as:

$$Mse(T) = \sigma^2 X_{\bar{S}}^2 \sum_S a_i^2 v_i + \sigma^2 \sum_{\bar{S}} v_j + Bias^2(T) = \sigma^2 X_{\bar{S}}^2 \sum_S a_i^2 v_i + \sigma^2 \sum_{\bar{S}} v_j + \beta^2 X_{\bar{S}}^2 \left(\sum_S a_i x_i - 1 \right)^2 . \qquad (6)$$

Under a model based approach, the case of sampling surveys without non responses is formally similar to the case of census surveys referred to populations including N units with $(N-n)$ non responses.

2.2 Estimation with non responses

We suppose $(n-n_R)$ non responses. The estimator of the total based on respondent units only is given by:

$$T_R = \sum_{S_R} y_i + \sum_{\bar{S}_R} \hat{y}_j = \sum_{S_R} y_i + \sum_{\bar{S}_R} \hat{\beta}_R x_j \qquad \text{with:} \qquad \hat{\beta}_R = \sum_{S_R} a_{Ri} y_i \quad \text{and} \quad \hat{y}_j = \hat{\beta}_R x_j \,. \qquad (7)$$

The label R means that estimation is based on respondent units only, without any preliminary imputation of non respondents. In particular, if $a_{Ri} = a_{Ri}^*$ - where a_{Ri}^* is given by (4) putting S_R in place of S – it also means $\hat{\beta}_R = \hat{\beta}_R^*$ and we have the *BLU* predictor based on n_R respondent units, which *Mse* is given by (2), putting respectively S_R in place of S and $\bar{S}_R$ in place of $\bar{S}$. A basic result is that, if the formal structure of a_i and a_{Ri} is the same – as it happens if they derive from the *BLU* estimation process under the model – imputation of non respondent units before final estimation is not influential. On the other hand, it may be useful if the choice of a_{Ri} derives from a logic different with respect to that driving the choice of a_i. For instance, in presence of non responses the estimation strategy may be based on calibration (Lundström and Särndal, 1999), which use is due to the risk of a bias derived from a wrong model specification. One may avoid imputation using coefficients a_{Ri} derived from calibrated weights, or impute non responses on the basis of $\hat{\beta}_R$ and then use calibration coefficients a_i: Results of the two strategies will be different (section 4). The general form of any estimator based on imputation of non respondents and weighting is:

$$T_w = \sum_{S_R} w_i y_i + \sum_{S\bar{R}} w_j \hat{y}_i \qquad (8)$$

3. Deterministic donor imputation and weighting

We analyse an estimation strategy based on two steps: 1) donor imputation procedure applied to non respondent units; 2) joint use of respondent and non respondent units data in order to estimate units not belonging to the theoretical sample on the basis of the same coefficients a_i defined in (3). When donor imputation is used, for each $j \in S_{\bar{R}}$ we select the particular donor labelled with $i(j)$ among the n_R respondents labelled with i. We say also $d_j=i$. The donor imputation process can be expressed through:

$$\hat{y}_{jd} = \hat{\beta}_j x_j \quad \text{where:} \quad \hat{\beta}_j = y_{i(j)} / x_{i(j)} = a_{i(j)} y_{i(j)} \quad \text{for each } j \in S_{\bar{R}} \tag{9}$$

The imputation rule (9) leads to model unbiased estimates. The estimator derived from (9) is given by:

$$T_{I(d)} = \sum_{S_R} y_i + \sum_{S_{\bar{R}}} \hat{y}_{jd} + \sum_{\bar{S}} \hat{y}_j = \sum_{S_R} y_i + \sum_{S_{\bar{R}}} \hat{\beta}_j x_j + \sum_{\bar{S}} \hat{\beta}_{I(d)} x_j \quad . \tag{10}$$

where $\hat{\beta}_j$ is defined by (9) and $\hat{\beta}_{I(d)} = \sum_{S_R} a_i y_i + \sum_{S_{\bar{R}}} a_j \hat{y}_{jd} = \sum_{S_R} a_i y_i + \sum_{S_{\bar{R}}} a_j \hat{\beta}_j x_j$. The estimator (10) is based on donor imputation of the (n-n_R) non respondents and imputation of units not belonging to S through the original coefficients a_i. It is *not linear* if the choice of donors – and, as a consequence, of coefficients $a_{i(j)}$ – depends on the knowledge of labels in S. A common technique used for selecting donors through a deterministic approach is based on the *nearest neighbour* method, according to the rule:

$$D_{i(j),j} = \left| x_{i(j)} - x_j \right| = \underset{i \in S_R}{Min}(D_{i,j}) \qquad \text{for each unit } j \in S_{\bar{R}} \tag{11}$$

where D is a distance operator. On the basis of (11), we can suppose that $x_{i(j)} \approx x_j$ for each unit $j \in S_{\bar{R}}$. We have: $Var(\hat{y}_{jd}) = Var(y_{i(j)}) = \sigma^2 v_{i(j)}$. Moreover, we obtain:

$$Mse(T_{I(d)}) = \sigma^2 \sum_{S_{\bar{R}}} v_{i(j)} + \sigma^2 \left(\sum_{\bar{S}} x_j \right)^2 \left\{ \sum_{S_R} a_i^2 v_i + \sum_{S_{\bar{R}}} a_j^2 x_j^2 a_{i(j)}^2 v_{i(j)} + 2 \sum_{i \in S_R} \left[a_i v_i \left(\sum_{j \in S_{\bar{R}}: d_j = i} a_j a_{i(j)} x_j \right) \right] \right\} +$$

$$+ 2\sigma^2 \left(\sum_{\bar{S}} x_j \right) \left\{ \sum_{i \in S_R} \left[a_i v_i \left(\sum_{j \in S_{\bar{R}}: d_j = i} a_{i(j)} x_j \right) \right] + \sum_{i \in S_R} \left[v_i \left(\sum_{j \in S_{\bar{R}}: d_j = i} a_j a_{i(j)}^2 x_j^2 \right) \right] \right\} + \sigma^2 \sum_{\bar{S}_R} v_j \, . \tag{12}$$

The meaning of each term included in squared brackets is that, fixed the label i in the first sum, the term into round brackets is a sum including a number of terms equal to the number of not respondent units j whose donor is given by that particular unit i. Given the formulas (2) when $S=S_R$ and (12), we can verify under which condition one may have: $Mse(T_{I(d)}) \le Mse(T_R)$. In particular, if $T_{I(d)}$ and T_R are based, respectively, on the optimal coefficients a_i^* and a_{Ri}^*, we can focus on the inequality:

$$Mse(T^*_{I(d)}) \le Mse(T^*_R) \, . \tag{13}$$

Even though T^*_R is the *BLU* predictor by construction, $T^*_{I(d)}$ may have a lower *Mse* because it is based on a larger number of sample units, of which n_R are real respondents and (n-n_R) have been imputed.

We can suppose that: $x_i=v_i$ for each unit i in the population. That simplifies formula (12), where we can put $v_{i(j)} = v_j = x_j$. From (2) when $S=S_R$ we know that:

$$Mse(T^*_{R/v=x}) = \sigma^2 (X^2_{\bar{S}_R} / X_{S_R} + X_{\bar{S}_R}) \, . \tag{14}$$

We have also: $a_i^* = X_S^{-1}$, $a_{i(j)} = x_i^{-1}$. Taking into account that $S = S_R \cup S_{\bar{R}}$ and $\bar{S}_R = S_{\bar{R}} \cup \bar{S}$, we obtain:

$$Mse(T^*_{I(d)/v=x}) = \sigma^2 \left[X_{S_{\bar{R}}} (1 + 2 X^2_{\bar{S}} X_S^{-2} + 4 X_{\bar{S}} X_S^{-1}) + (X^2_{\bar{S}} X_S^{-1} + X_{\bar{S}_R}) \right] . \tag{15}$$

From (14) and (15), we can also verify that:

$$Mse(T^*_{I(d)/v=x}) \le Mse(T^*_{R/v=x}) \qquad \leftrightarrow \qquad (X_U / X_S)^2 \le (X_S / X_{S_R})(X_U / X_S - 0{,}5) \, . \tag{16}$$

For instance, if X_U=100 and X_S=20, the condition (16) is satisfied if $X_{S_R} \le 3{,}6$. From (16), putting $f_{xS} = X_S / X_U$ and $f_{xS_R} = X_{S_R} / X_U$, we derive the condition for which the donor strategy may improve T_R:

$$(X_{S_R}/X_S) \le (X_U/X_S - 0{,}5)(X_U/X_S)^{-2} \quad \leftrightarrow \quad (f_{xS_R}/f_{xS}) \le (f_{xS}^{-1} - 0{,}5) f_{xS}^2 \tag{17}$$

In the Table 3-1 we have calculated the ratio between the mean squared errors of the estimators $T^*_{I(d)}$ and T^*_R for some levels of the ratios X_S/X_U and X_{S_R}/X_S when *v*=*x*. Ratios lower than one correspond to conditions for which (17) is satisfied (boxes with a grey shade). The usefulness of donor imputation is guaranteed if the weighted sampling rate is at least 60% and the weighted response rate is not larger than 40%: that can happen in strata characterised by a large sampling rate and a bad survey outcome (respondents are less than half of the original sample size). In particular, in a census survey it is necessary that the weighted sampling response rate is lower than 50%. If the sampling response rate is low (not larger than 40%), in the most part of cases estimation without donor imputation should be preferred: for instance, with a 20% weighted sampling response rate we should carry out a donor imputation only if the original weighted sampling rate is 30% at least. For weighted sampling response rates larger than 40% we should avoid donor imputation. In the last column we have calculated the level of the weighted sampling response rate beyond which donor imputation may be used: for instance, when the weighted sampling rate is equal to 40%, donor imputation may be preferred if the weighted sampling response rate is lower than 32%. In short, a weighted sampling rate lower than 50% is a *necessary* but not sufficient condition such that the donor imputation strategy may improve the model based one. If S=S_R and donor imputation is carried out on the units belonging to $\overline{S}$, then the previous condition would be *sufficient* as well.

Table 3-1
Ratio between *Mses* of estimators $T_{I(d)}$ and T_R when *v*=*x* for some levels of the weighted sampling rate and the weighted response rate and conditions for which $T_{I(d)}$ improves T_R (relation (17))

$100 \cdot f_{xS}$	$100 \cdot (f_{xS_R}/f_{xS}) = 100 \cdot F_{xS_R}$										$Mse(T^*_{I(d)}) \le Mse(T^*_R)$
	5,0	10,0	20,0	30,0	40,0	50,0	60,0	70,0	80,0	90,0	if $100 \cdot F_{xS_R} \le$
5,0	1,03	2,05	4,10	6,15	8,21	10,26	12,31	14,36	16,41	18,46	4,88
10,0	0,53	1,05	2,11	3,16	4,21	5,26	6,32	7,37	8,42	9,47	9,50
20,0	0,28	0,56	1,11	1,67	2,22	2,78	3,33	3,89	4,44	5,00	18,00
30,0	0,20	0,39	0,78	1,18	1,57	1,96	2,35	2,75	3,14	3,53	25,50
40,0	0,16	0,31	0,63	0,94	1,25	1,56	1,88	2,19	2,50	2,81	32,00
50,0	0,13	0,27	0,53	0,80	1,07	1,33	1,60	1,87	2,13	2,40	37,50
60,0	0,12	0,24	0,48	0,71	0,95	1,19	1,43	1,67	1,90	2,14	42,00
70,0	0,11	0,22	0,44	0,66	0,88	1,10	1,32	1,54	1,76	1,98	45,50
80,0	0,10	0,21	0,42	0,63	0,83	1,04	1,25	1,46	1,67	1,88	48,00
90,0	0,10	0,20	0,40	0,61	0,81	1,01	1,21	1,41	1,62	1,82	49,50
100,0	0,10	0,20	0,40	0,60	0,80	1,00	1,20	1,40	1,60	1,80	50,00

4. Empirical attempt and conclusions

Data used for the empirical attempt derive from the quarterly wholesale trade sample survey carried out by ISTAT. Even though its main purpose is the estimation of quarterly turnover indexes with base 2005=100 – based on turnover data picked up quarterly – in this context we focus on the estimation of the yearly total turnover available from the business register for each active enterprise, for the years 2003-2007. The choice is due to the need of using the true total amount of the target *y*-variable as a benchmark for assessing precision of estimates got using different estimation criteria. Moreover, we have supposed that:

1) the population (size *N*) is given by the whole theoretical sample;
2) the sample (size *n*) is given by the sample of units that respond within 180 days from the end of the reference quarter (in the real survey context, final estimates are just released after 6 months);
3) the sample of respondents (size n_R) includes the quick respondent units (responding within 90 days).

For each year we have considered as "respondents within 180 days" (the *n* units defined in 2)) those responding within 180 days in *all* the 4 quarters, and as "respondents within 90 days" (the units defined in 3)) those responding within 90 days in *all* the 4 quarters. Separate analyses have been carried in each of the 8 following domains: 1) Wholesale on a fee or contract basis; 2) Agriculture raw materials and live animals; 3) Food, beverages and tobacco; 4) Household goods; 5) Non-agriculture intermediate products; 6) Machinery, equipment and supplies; 7) Other products; Total wholesale trade. If *Y* is the current reference year, we have considered as *y*-variable the total turnover referred to (*Y*-1) – since the business register is updated with a year delay – while *x* is the total turnover of (*Y*-2). On the average 2003-2007 (Table 4-2) for the total

wholesale trade we have N=7.572, n=5.636 and n_R=4.631, so that $100 \cdot n/N$= 74,4 and $100 \cdot n_R/n$=82,2. According to the theoretical issues derived from Table 3-1, we should not have large efficiency gains using donor imputation. The 8 estimation strategies compared are described in the Table 4-1.

Table 4-1
Compared estimation strategies

Code	Model	Definition	Details
I	v=1	Optimal prediction - no imputation	Predictor (7), β is given by the second formula (1) with S=S_R
I	v=x		
II	v=1	Model-based imputation of non responses and calibration	Predictor (8), non responses estimated by the second and third formulas (7), β given by the second formula (1) with S=S_R, weights w from calibration
II	v=x		
III	v=1	Calibration - no imputation	Predictor (7), weights w from calibration
III	v=x		
IV	v=1	Donor imputation and optimal prediction	Predictor defined by (10) where coefficients a are optimal under the model
IV	v=x		

Main results have been summarised through the mean of absolute percent estimation errors (*MAPE*): figures in Table 4-2 are arithmetic means of 20 quarterly estimation errors covering the period 2003-2007. The use of the true response rates (upper part of the table) lead to the following conclusions:

a) on the average of eight domains, the best strategy (bold) is (1) with v=x (*MAPE*=1,70%), that is also the best one for six domains (all but 4 and 6) and the second best (underlined) for domain 6. This outcome implies that imputation is scarcely useful. That is also confirmed by the second best result got by strategy (III) with v=x (*MAPE*=2,00%), which is based on calibration without imputation. This strategy is the second best in four domains as well (1, 2, 7 and total).
b) However, strategy (IV) - based on donor imputation - with v=x turns out to be the best one for domain 6 (*MAPE*=2,28%) and the second best for domains 3 and 4. Moreover, the alternative model based imputation defined by strategy (II) with v=x leads to the lowest *MAPE* (1,64%) for domain 4.
c) Finally, as a matter of fact the option v=1 is quite always less realistic than v=x: that implies refusal of the homoskedastic pattern for unit model variances.

In order to assess steadiness of the previous results with a lower response rate, a blanking of the 50% of responses was replicated at random 1.000 times. Results summarised in the bottom part of the table lead to similar conclusions: the only significant exception concerns the enforcement of calibration without imputation (strategy (III) with v=x) with respect to strategy (I) with v=x, since the former becomes the best one for domains 1, 2 and 3 and on the average (*MAPE*=2,98%), while the latter gets the second best performance. No relevant changes characterise performances of strategies based on donor imputation, even though, on the average, that should still be preferred to model based imputation (strategy (II)).

Even though the previous results suggest that, in presence of non responses, donor imputation may improve precision of estimates in a few circumstances only, additional efforts should be spent towards two directions: 1) theoretical improvements of the donor selection mechanism (for instance, donor's selection may be based on more than one variable); 2) further empirical attempts referred to different variables (not continuous, as the number of job vacancies) and frameworks characterised by different response rates.

Table 4-2
***MAPE* by domain and strategy (real data and 1.000 random replications, average 2003-2007)**

Domain	N	n	n_R	Strategy (I) v=1	(I) v=x	(II) v=1	(II) v=x	(III) v=1	(III) v=x	(IV) v=1	(IV) v=x
True response rates											
1	1.205	789	605	1,95	**1,15**	3,70	1,96	2,89	1,31	2,68	1,55
2	570	417	331	1,67	**1,58**	2,11	1,85	2,13	1,62	2,43	1,65
3	961	743	605	2,39	**1,63**	4,19	3,18	3,60	1,96	2,87	1,75
4	2.227	1.671	1.387	3,15	1,95	**1,64**	1,72	1,82	2,02	2,23	1,72
5	1.248	997	853	3,31	**2,16**	3,11	3,22	3,49	2,90	2,84	3,17
6	863	671	570	3,29	2,60	3,29	3,03	3,44	3,14	4,66	**2,48**
7	498	348	281	1,68	**1,41**	3,17	1,86	2,40	1,50	2,25	2,71
Total	7.572	5.636	4.631	1,89	**1,16**	2,10	2,21	2,66	1,56	1,84	1,82
Average				2,42	**1,70**	2,91	2,38	2,80	2,00	2,72	2,11
1.000 random replications with a 50% response rate											
1	1.205	789	405	4,48	3,16	6,07	4,85	4,75	**3,01**	4,97	3,34
2	570	417	214	5,73	3,41	4,77	4,22	4,44	**2,97**	5,09	4,08
3	961	743	385	3,98	2,50	5,95	4,81	5,20	**2,43**	5,09	2,71
4	2.227	1.671	857	5,51	3,69	3,98	4,87	4,06	3,57	5,54	**3,02**
5	1.248	997	512	4,68	**3,60**	4,60	4,20	4,94	3,78	5,37	3,78
6	863	671	339	4,84	**4,02**	6,19	5,32	7,12	4,23	4,88	4,14
7	498	348	175	2,10	**1,73**	4,17	3,57	4,37	1,87	3,59	2,78
Total	7.572	5.636	2.887	2,91	**1,91**	3,74	3,54	4,18	1,98	4,40	2,31
Average				4,28	3,00	4,93	4,42	4,88	**2,98**	4,86	3,27

References

Beaumont, J.F. and Bocci, C. (2009). Variance Estimation when Donor Imputation is Used to Fill in Missing Values, *Canadian Journal of Statistics*, on line publication at the address: http://www3.interscience.wiley.com/cgi-bin/fulltext/122466454/PDFSTART.

Billiet, J., Philippens, M., Fitzgerald, R. and Stoop, I. (2007). Estimations of Non-response Bias in the European Social Survey Using Information from Reluctant Respondents, *Journal of Official Statistics*, Vol.23, 2, pp.135-162.

Cicchitell,i G., Herzel, A. and Montanari, G.E. (1992). *Il campionamento statistico*, Bologna, Il Mulino.

Copeland, K.R. and Valliant, R. (2007). Imputing for Late Reporting in the U.S. Current Employment Statistics Survey, *Journal of Official Statistics*, Vol.23, 1, pp.69-90.

Gismondi, R. (2008). Reducing Revisions in Short-term Business Surveys, Statistica, anno LXVIII, 1, pp.85-116, Bologna, Clueb.

Lundström, S. and Särndal, C.E. (1999). Calibration as a Standard Method for Treatment of Non-response, *Journal of Official Statistics*, Vol.15, 2, pp.305-327.

Rizzo, L., Kalton, G. and Brick, M.J. (1996). A Comparison of some Weighting Adjustment Methods for Panel Non-response, *Survey Methodology*, 22, 1, pp.43-53.

Valliant, R, Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference – A Prediction Approach*, New York, J.Wiley & Sons.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

On Balanced Random Imputation in Surveys

David Haziza, Guillaume Chauvet and Jean-Claude Deville[1]

Abstract

Random imputation methods are often used in practice because they tend to preserve the distribution of the variable being imputed, which is an important property when the goal is to estimate quantiles. A special case of random imputation, random hot-deck imputation, is often used in practice if the variable being imputed is categorical because it eliminates the possibility of impossible values. Also, it is used when it is desired to impute more than one variable at the time because the same donor can be used to impute all the missing values, which helps preserving the relationships between variables.However, random imputation methods introduce an additional amount of variability, called the imputation variance, due to the random selection of residuals. In this presentation, adapting the Cube method (Deville and Tillé, 2004) for selecting balanced samples, we propose a class of random balanced imputation methods which reduce/eliminate the imputation variance while preserving the distribution of the variable being imputed. The proposed class of imputation methods can be applied for both categorical and continuous variables. Also, it can be used for any type of sampling design. The results of a limited simulation study will be presented.

[1] David Haziza, Université de Montréal, Canada (david.haziza@statcan.gc.ca); Guillaume Chauvet and Jean-Claude Deville, Laboratoire de Statistique d'Enquête (CREST/ENSAI), France.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Testing New Imputation Methods for Earnings in the Survey of Income and Program Participation

Martha Stinson and Gary Benedetto, U.S. Census Bureau, U.S.A.[1]

Abstract

This paper explores the feasibility and effectiveness of three significant changes to standard Census Bureau methods of imputing earnings in the Suwey of Income and Program Participation (SIPP). Currently imputation is performed by stratifting the data based on a set of analyst-chosen characteristics, randomly sorting within each sub-group, and choosing a donor based on the nearest neighbor. We investigate the possibility of using a model-based approach, supplementing survey*collected job and demographic characteristics with administrative earnings data, and using multiple imputation as proposed by Rubin. We will model monthly earnings from January 2A04 6 December 2005 using the SIPP 2004 panel linked to W-2 tax records extracted from the Social Security Master Earnings file. We will use linear regression techniques to estimate a posterior predictive distribution that is the distribution of earnings conditional on all observed characteristics (including administrative earnings). From this distribution, we will take four draws to create four imputed values per case with missing earnings. After thus "completing" the missing data, we will compare results using original versus new imputed values from several standard analyses in order to assess the impact of our new method. In particular, we will look at coefficients in a classic eamings regression, trends in earning changes over time, the moments of the cross-sectional eamings distribution for a particular month, and poverly levels as based on family income, of which eamings are an important component. The four imputed values will allow us to calculate variance estimates using Rubin's multiple imputation variance formulae and to assess the impact of imputation on the significance of regression coefficients, the shape of the earnings distribution, and the margin of error on poverty estimates.

[1] Martha Stinson (martha.stinson@census.gov) and Gary Benedetto (gary.linus.benedetto@census.gov), U.S. Census Bureau, U.S.A.

EDIT AND IMPUTATION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

EU-SILC in Slovenia – Experiences so far

Rudi Seljak[1]

Abstract

The Survey on Income and Living Conditions (SILC) is a European harmonized survey aiming at providing the data on living conditions in which the household members and selected individuals live and how they integrate themselves in the society. Since the survey is output harmonized, the output variables are prescribed by the European Regulation, while the method of collecting the input micro-data is more or less left to the decision of each particular country.

In Slovenia the micro-data for the EU-SILC survey are gathered from three types of sources: »classical« survey, other statistical sources, administrative sources. The exhaustive use of the administrative sources has an obvious advantage of response burden reduction but can also cause certain disadvantages, most obviously the increased extent of data editing work. In the paper we summarize the four-year »EU-SILC experience« on merging the data from different sources and point out the main advantages and disadvantages of such an approach.

Key Words: Survey On Income And Living Conditions, Editing, Administrative Data.

1. Introduction

The European Survey on Income and Living Conditions (EU-SILC) is the project aiming at setting up the European harmonized survey for gathering comparative statistics on income distribution and social exclusion from EU Member States, Norway and Iceland. The project was launched in 2003 (at that time still on the basis of a gentlemen's agreement) in 6 European Member States, widened in 2004 to 12 "old" Member States, Estonia and Iceland, and then in 2005 including all (at that time) Member States, Norway and Iceland.

From 2004 on the survey has been carried out on the basis of the European Parliament Regulation. The Regulation defines the EU-SILC as the output harmonized survey, meaning that the Member States should provide the output variables prescribed by the Regulation, whereas the method of collecting the data is left to the particular country. So, some countries still collect all the data in the "classical" way, with the field survey, while other use a combination of the "classical" survey data and data derived from the different administrative sources. Since the EU-SILC data should serve for the purposes of cross-sectional as well as longitudinal analyses, it is strongly recommended to design it as a panel survey.

In Slovenia the EU-SILC was first carried out in 2004 as the pilot survey and then in 2005 as the "regular" survey. In the planning and setting-up phase we tried to follow the Eurostat's recommendation that as many already existing data sources as possible should be used in order to reduce the response burden and to consequently increase the response rate. Hence, in Slovenia the micro-data for the EU-SILC are gathered from three types of sources. The first part of the data is collected by the »classical« survey; the second part comes from other statistical sources and the third part from registers and administrative sources. Among others, all the income-related variables (which are usually considered as highly sensitive ones) are gathered from the different administrative sources.

Although the exhaustive use of the administrative sources has many advantages, especially in the field of response burden and survey costs reduction, such an approach can also cause certain disadvantages. The most outstanding disadvantage is certainly the increased extent of data editing work. When we first carried out the survey, we were not fully aware of the complexity of the editing process. Therefore, at that time all the editing procedures were more or less ad-hoc made computer programs, with lack of systematic and long-term perspective. Since such an approach caused a lot of problems in the processing phase, we started to build a more generic system, which would enable easier management and better control of the data processing.

In the paper we describe the experiences with the four-year execution of the EU-SILC survey in Slovenia. In the first part of the paper some general information will be given and the application for data processing will be described. In the second

[1] Rudi Seljak, Statistical Office of the Republic of Slovenia, e-mail: rudi.seljak@gov.si, Tel.:+386 1 2415 294

part we will summarize main advantages and disadvantages of the exhaustive use of the administrative data as the direct data source. At the end some conclusions and some directions for the future development will be pointed out.

2. EU-SILC in Slovenia

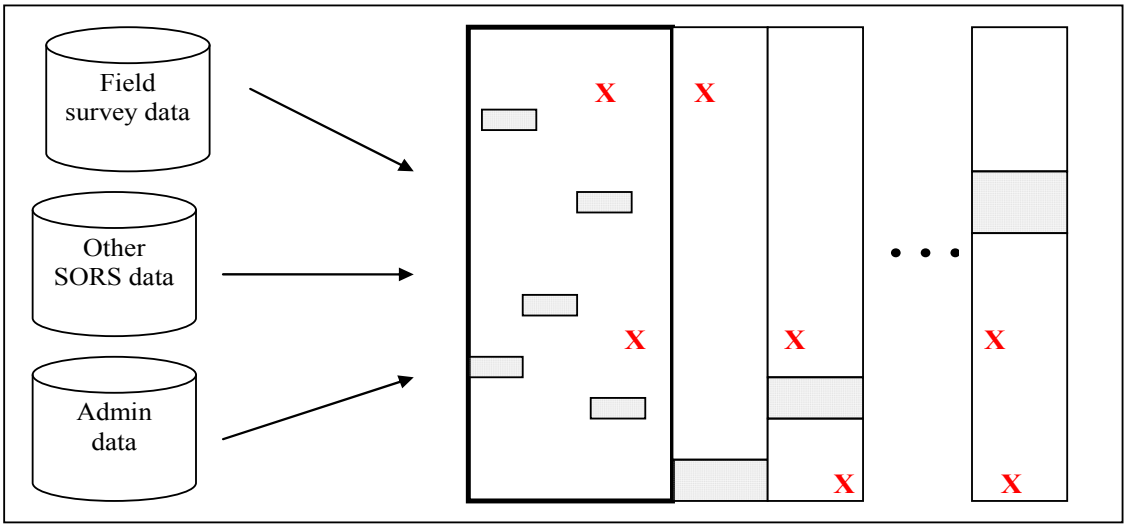
As already mentioned in the introduction, in Slovenia the micro-data for the EU-SILC survey are gathered from three different types of sources:

- Classical field survey using CAPI and CATI questionnaires. The data set of completed questionnaires for approximately 9,000 households (approx. 30,000 persons) per year is the basis for the data integration.
- Other sources inside the Statistical Office. The data from the Employment Register and the Survey on Scholarships are used.
- Data from the external institutions. The data from the following institutions are used: Tax Authority; Ministry of Labor, Family and Social Affairs; Pension and Disability Insurance Institute; Employment Service of Slovenia; Health Insurance Institute; Ministry of the Interior.

Since in Slovenia the Register of Households and the Register of Dwellings are still in the establishment phase, they are at the moment not used for the sampling purposes. This means that we select a random sample of selected persons and then all the persons living in the same household as the selected person are interviewed. Since we are not allowed to obtain the administrative Personal Identification Number (PIN) directly from the respondent, these identification numbers are available for the selected persons only. Therefore, before the integration phase, we use the indirect record linkage approach, using the base information (name, surname, address, birth date, etc.) collected in the field to find the corresponding PIN in the Central Population Register. With the combination of the computer application and some manual work, PINs for approx. 99% of the persons are determined and only 1% of PINs are imputed.

After getting the PINs for all the units, the data integration of different sources is carried out. Because of the imperfection of the sources, data for some units are missing and some data become inconsistent after the integration phase. The situation after the integration phase is presented in Figure 1, where the grey parts of the integrated data present the missing data and the red crosses logically inconsistent data.

Figure 2-1
Integration process



3. Data processing

In 2005, the first year of the execution of the survey, all the processing was still done in a "classical" way by using the custom made (SAS) computer programs. In practice, subject-matter and general methodology personnel prepared the instructions and on the basis of these instructions the computer programs were prepared by the IT personnel. After the first execution of the programs, eventual needs for corrections and improvements were provided and then the new version of the program was prepared. Due to the complexity of the EU-SILC data, such an approach was very time consuming, wasting a great deal of work-time of many employees and was justifiably designated as very inefficient. Therefore, we started to consider the feasibilities for constructing a new, more general and more flexible system which would in the final stage serve also for the purposes of other surveys.

The general idea was to prepare an application which would be metadata driven (MDD), meaning that all information that determines the parameters for the execution of the processing for a specific survey and a specific reference period should be provided outside the core computer code. No information referring to a specific survey execution should be incorporated in the program code, but should be provided by the subject-matter personnel through the special metadata tables. A more detailed description of these metadata for the example of three particular processes is given below.

3.1 Logical checks

Two groups of logical checks could be performed: cross-sectional and longitudinal ones. When performing the cross-sectional checks only the data from the current year can be controlled, whereas in the case of longitudinal checks also the data from the previous reference year could be part of the checks. The user has also the option to choose which "version" of the data is to be controlled. By the term "version" we here refer to data sets after different parts of the process, such as raw data, data after the transfer of previous year data, data after logical corrections, etc.

The following metadata are required in order to perform the process of logical checks:

- Denotation of the logical check
- Name of the table in the database
- Logical expression, which should always be written in the form that determines the error
- Comment or description of the check (arbitrary)

After the execution, the user gets immediately on the screen the number of the failed records for each of the defined checks. The set of failed units is written into the standard (SAS and Access) table. In this table for each of the failed records there is also the information which check(s) the record failed. These outputs are then the basis for the preparation of the "correction metadata" which are needed later in the process.

3.2 Data corrections

With this process the values of the variables that have formerly been designated as erroneous can be corrected. Two different ways of selecting the units for which the particular variable should be corrected and three different ways of determination of the new value can be chosen. The first way to determine the units for which the value should be corrected is to provide the unique identification of the unit. We call these corrections individual corrections. The second way is to correct the values for the certain variable for all the units that satisfy the certain condition. We call these corrections systematic corrections.

The new (corrected) values can be determined in three different ways. The user can provide the exact new value, provide the arithmetic expression which would determine the new value or provide the lower and upper limit (range) for the new value. The value in the latter case is estimated by using the Banff imputation procedure proc donorimputation, where the edit constraints are determined with the given range.

The metadata that are to be provided slightly differ with regards to the chosen method of unit selection and to the chosen method of new value determination:

- Name of the variable
- Name of the table in the database
- Identification of the unit (in the case of individual corrections)
- Expression which determines the set of units for which the corrections should be performed (in the case of systematic corrections)

- Table in the database from which the above expression is to be calculated. It is usually the same table as the one which contains the value of the variable, but not necessary.
- Fixed value if the first method of corrected value determination is chosen
- Arithmetic expression if the second method of corrected value determination is chosen
- Lower and upper limit for the new value if the third method of corrected value determination is chosen

3.3 Imputations for missing data

So far, 7 different imputation methods can be used and these methods can be divided into two different groups: parametric and non-parametric ones. The parametric method in fact represents a group of methods and the user can use the arbitrary number of different parameterizations to create an arbitrary number of different methods. On the other hand, the non-parametric method needs no parameterization, meaning that the user can not create "the custom made" methods, but one still needs to insert the required metadata.

For the illustration how the system works, we here present how the group of hot-deck methods is to be managed. When one decides that a certain variable will be imputed by using the hot-deck method, a two-step procedure has to be carried out. In the first step the parameterization of the method has to be accomplished. In the case of the hot-deck method the parameterization means that the user chooses the (up to 5) stratification variables and the matching variable. These variables must of course be part of the incoming data set. When the parameters are determined the user designates the method with the denotation, for example HD1. If the suitable parameterization of the hot-deck method has already been defined before for the purposes of some other variable imputation, the first step of the procedure can of course be omitted. In the second step the metadata for the imputation of the particular variable have to be inserted.

The required metadata are:
- Name of the variable to be imputed
- Name of the table in the database
- Denotation of the imputation method to be used (in our case HD1)
- Logical expression which determines the units for which the imputations should be performed
- Logical expression which determines the units which could be treated as the potential donors
- The step (from 1 to n) in which this particular method for this particular variable will be executed. The system enables the possibility of the execution of the imputation procedure in several consecutive steps.

4. Use of administrative data – pros and cons

The Statistical Office of the Republic of Slovenia (SORS) has a long history of successful usage of the data coming from registers and other administrative sources and has always declared itself as a register-oriented statistics. Therefore the decision for the exhaustive usage of different administrative sources in the case of the EU-SILC survey was somehow a natural choice in the phase of planning the design of the survey. Although such usage has many advantages, it also causes some "side effects", which make the statistical process more demanding and presents quite a challenge for the designers of the statistical process. Most of these challenges are related to the integration and editing part of the process. In the continuation, we present the main advantages and main disadvantages of the administrative data usage in the EU-SILC survey, as detected through the discussion with the survey personnel.

The following main advantages have been pointed out:
- Because of the administrative data usage, the field questionnaire is much shorter and the interview duration is significantly shortened. We estimate that in the case of "full" questionnaire the duration of the interview would be around 60 minutes, while now the average duration is around 25 minutes.
- Because of the shorter questionnaire it is easier to maintain the reasonably low level of item and unit non-response. We can justifiably assume that in the case of "full" questionnaire it would be difficult to have (approx.) 25% of unit non-response as it is the case at the moment.
- The most difficult and sensible questions (income related) are skipped, because all these items are gathered from the administrative sources. Therefore the risk for the item non-response is much lower and fewer imputations for the missing values procedures are required.
- There is essential cost reduction due to the administrative data usage. Namely, if all the data had to be gathered with the questionnaire, no telephone interviewing would be possible and the survey costs would be much higher.

The main drawbacks, which present the challenges for the future development, are:

- Because data are coming from many different sources, the statistical process is much more demanding and time consuming. When data are coming from different sources, we are inevitably faced with many linkage and consistency problems, making the editing phase the crucial stage in the statistical process.
- As usually in the case of administrative data usage, there are problems with inconsistency of variable definitions. Therefore all the sources should be carefully and constantly studied to detect the possible concept differences.
- There is a timelines problem, caused by the late arrival of some (especially tax data) data sources. The results are disseminated approximately 9 months later as they would be in the case of the full field survey.

5. Conclusions

The EU-SILC is for SORS (as probably for most of the European statistical offices) one of the most demanding, time and cost consuming surveys. A large amount of (harmonized) statistical results has to be produced and published on the national as well on the aggregated European level. While the output results of the survey are fully prescribed by the Regulation, the method of collecting the needed data is left to the decision of each particular country. In the case of Slovenia the micro-data are gathered from three different sources: a part of the data is gathered from the field survey, a part from other statistical sources and a (large) part from different administrative sources.

Although the usage of different administrative sources has many advantages it also causes some "side effects", which make the statistical process more demanding and presents quite a challenge for the designers of the statistical process. Most of these challenges are related to the integration and editing part of the process. To ease the very demanding data processing phase, a new generic application, based on the metadata driven approach, was built. The main change with regards to the previous practice is that the subject-matter personnel have much more influence on the decisions about how the data should be processed. This fact is usually also pointed out as the main advantage of the new system.

Since the application has been developed recently, there is still a lot of space for improvements and completions. The main goal of the future development is to upgrade the application in such a way that it could be used also for the purposes of other surveys. To achieve this goal, we firstly have to build a more integrated tool which would enable easier management of the whole process and would be the main challenge for the forthcoming years.

References

Banff Support Team. *Functional Description of the Banff System for Edit and Imputation System*, Statistics Canada, Quality Assurance and Generalized Systems Section, Technical Report.

European Parliament and Council of the European Union (2003). *Regulation of the European Parliament and of the Council of 16 June 2003 concerning Community Statistics on Income and Living Conditions (EU-SILC).*

Eurostat (2004). *Imputation procedures*, EU-SILC Documents 136/04, European Commission, Eurostat.

Eurostat (2004). *Description of Target Variables: Cross-Sectional and Longitudinal*, EU-SILC document 065/04.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, March 1976, Volume 71, No. 353, 17-35.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Longitudinal data editing for the Italian LFS

Simona Rosati and Barbara Boschetto[1]

Abstract

Although the main product of the Italian Labour Force Survey is cross-sectional data, longitudinal data is an aim of great consequence. Nevertheless the time dimension makes it more difficult to develop a strategy for dealing item non-response as well as unit non-response. This paper is mainly devoted to the longitudinal imputation method applied to correct item non-response on single questions. Special emphasis is given to the issues regarding data inconsistencies and difficulties that arise when longitudinal data are used. The paper concludes with a discussion about the main outcomes. An ample explanation of the process for the record linkage is also reported. All the results refer to the period 2007(1)-2008(1).

Key Words: Item Non-Response, Data Editing, Longitudinal Imputation.

1. Introduction to the Labour Force Survey

The Italian Labour Force Survey (LFS) is a continuous survey carried out during every weeks of a year, designed to produce annual estimates, quarterly and monthly estimates of labour market main indicators.

Thanks to the design structure it is also possible to reconstruct longitudinal microdata. Survey design is in fact based on a rotating sample, where households are interviewed four times over a fifteen-month period, following a 2-2-2 rotation scheme: households participate for two consecutive quarters, then they temporally exit for the following two quarters, and then come back in the sample for other two consecutive quarters until they definitively exit from the survey. After four interviews the household is replaced by another in the same or a similar area.

In particular, the sample of households for each quarter consists of four rotation groups or panels, which conduct the first, second, third and fourth interview respectively. Each round of interviews for a panel is designated a wave. As displayed in the example in Figure 1-1, sampling households that are interviewed for the first time in the 1st quarter of 2007, identified with the group G1, will be interviewed again in the 2nd quarter of 2007 (G2), then, after a break of two quarters, will be interviewed in the 1st and 2nd quarter of 2008 (G3 and G4, respectively). This rotation scheme implies that half of the households interviewed in a quarter are re-interviewed after three months, a half after twelve months, a quarter after fifteen months.

Figure 1-1
Rotation scheme of the survey design (LFS)

Quarter	Panel								
2006(4)	A4	B3			E2	**F1**			
2007(1)		B4	C3			**F2**	**G1**		
2007(2)			C4	D3			**G2**	H1	
2007(3)				D4	E3			H2	I1
2007(4)					E4	**F3**			I2
2008(1)						**F4**	**G3**		
2008(2)							**G4**	H3	

Interviews of the same individuals in different time periods represent the longitudinal information of survey. It is important to remember that the LFS was not designed to be a longitudinal survey, thus persons which move out of the selected households, or households which move out of the municipality, are not re-interviewed. This excludes the part of the population that moves in the territory from the longitudinal analysis of labour market. Consequently, longitudinal population is defined as the population that is resident in the same municipality at a given time frame.

[1]Simona Rosati and Barbara Boschetto, Italian National Institute of Statistics, Via Adolfo Ravà 150, Roma, Italy, 00142 (e-mail: sirosati@istat.it; boschett@istat.it)

An additional important factor is that survey design involves the change of household in case of exit from the sample due to different reasons (refusal, wrong address, etc.), thus the sample of households is actually composed by "base household" and its three "replacing households" (the so-called "quartet"). Obviously in case of replacement of household, the longitudinal population will be less than the cross population.

Interviews are carried out through the combination of different Computer Assisted Interviewing (CAI) techniques by professional interviewer network. General rule is that Computer Assisted Personal Interviewing (CAPI) technique is used for the first interview, while Computer Assisted Telephone Interviewing (CATI) technique is used for the interviews following the first, wherever is possible.

Use of CAI techniques ensures a high quality level of data collection, furthermore it is particularly useful to support the panel component of the survey. The electronic questionnaire is able to manage information given by respondents in previous interviews, so it is possible to conduct the interviews following the first asking to the sampling person to confirm or not the information given in the previous wave (only if there is no confirmation, the question will be asked again).

This system offers many advantages (when the labour status or other characteristics of the respondent do not have substantially changed): it not only significantly reduces the time of interview and the possibility of errors due to interviewer, but also allows an immediate correction in case of some errors recorded in a previous interview. This factor, as will be explained below, provides a strong support to the process of correction of longitudinal data.

Use of electronic questionnaire also greatly facilitates the matching of same individual information for the construction of longitudinal microdata. The assignment of a unique key, identifier for each individual in the household, assure that information collected in different periods are effectively linked to the same person. The electronic questionnaire also records all the changes occurring within the family, due at the exit or entry of a component, between an interview and the next. This immediately gives the number of people that are not present in each moment of the survey, and to take it into account during the record linkage process.

2. The linkage process

The rotating system of the LFS implies that 50 per cent of the quarterly sample is interviewed again after three and twelve months. Although the goal is to obtain labour market flows at twelve months, we chose to link all four waves of survey, for each rotation group of quarterly sample, to be able to use all available information during correction and thus ensure maximum consistency information. This means that the record linkage procedure was applied initially on individuals who participated in two waves at twelve months, and later information on two other waves (at three and fifteen months) were attached.

The longitudinal population in the LFS is defined as population that is resident in the same municipality at a given time frame. For this reason, chosen the reference period for the linkage at twelve months (in this work, the 1st quarter of 2007 and the 1st quarter of 2008), two source datasets were compared to identify the reference population of longitudinal file. In fact, initial population can change during one year for several reasons:

- Non-responding households in a "quartet" (i.e. all families included in the "quartet");
- Non-responding household which is replaced by another family within the same "quartet", therefore it will not be present at the next time;
- Individuals who are no longer members of the household due to death or transfer to another municipality;
- New household members due to birth or transfer from another municipality.

The presence of a unique individual identification code within the family (individual-key, composed by six numbers), makes the record linkage easier, but it can be affected by errors, so that a deterministic verification procedure was applied, using specific variables, which allows us to identify on the one hand the individuals that are correctly matched, on the other hand the "false negatives" (when one person has two different individual-keys) and "false positives" (when a single individual-key is associated with two different people).

Considering only the households that were interviewed in both matched quarters, a file composed by "matchable" individuals (i.e. individuals who could be linked) was obtained. In this file couples of individuals linked through individual-key were distinguished from not linked individuals.

Record linkage is controlled by a deterministic process that verifies the coincidence of some important personal identifiers: the date of birth, sex and name. Following this control a specific code is given to identify the exact correspondence of these variables or otherwise.

In this way individuals who were matched correctly were immediately identified, because after a year they had the same individual-key and the same personal identifiers.

Individual-key affected by error is a possible event even if very rare, thus the individuals that were not matched by individual-key or those that had discordant personal data in two quarters, were compared with each other within the same family, through the only personal identifiers (date of birth, sex and name). Individuals who showed a perfect coincidence of all variables (except individual-key) were then recovered.

At this point, initial dataset of "matchable" individuals (composed of 79,151 individuals in households interviewed in 2007(1) and 2008(1)), was divided as follows (Table 2-1):

- Total linked individuals (76,985, representing 97.3 per cent), of which:
  - Individuals who were correctly matched by individual-key and consistent with personal information (75,854, representing 95.8 per cent of total "matchable" individuals);
  - Individuals with wrong individual-key (1,131 at 1.4 per cent of total), of which:
    - Linked individuals who were recovered from the initial amount of "false positives" (individuals who had the same individual-key, but conflicting values of personal variables, retrieved through key exchange: 740 people, 0.9 per cent of cases);
    - Unlinked individuals or "false negatives" (individuals who had different individual-key but consistent values of personal variables, retrieved through the exchange of key: 391 individuals, 0.5 per cent of cases);
- Total no-matched individuals (2,166 amounted to 2.7 per cent of total matchable individuals), of which:
  - "False positives" (individuals with the same individual-key but different personal information: 567 individuals, 0.7 per cent);
  - Individuals that were present in a single wave of the period for analysis (1,599 individuals, 2.0 per cent of total).

Table 2-1
Record linkage rates

Individuals in households interviewed in 2007(1) and 2008(1)	number	percent
Matchable records	79,151	100.0
- Linked records	76,985	97.3
-- Records correctly matched	75,854	95.8
-- Records with wrong individual-key	1,131	1.4
---- Linked records recovered from false positives with exchange of key	740	0.9
---- Unlinked records recovered from false negatives with exchange of key	391	0.5
- Total no-matched records	2,166	2.7
-- False positives	567	0.7
-- Individual records that are present in a single wave of the period for analysis	1,599	2.0

The results of this match revealed first the high degree of reliability of individual-key attribution (75,854 records matched with the correct key, equal to 98.5 per cent of 76,985 matched individual records). Secondly they showed the efficiency of deterministic control procedure and record linkage procedure described above. Indeed on the one hand, it allows to validate the correctness of individual-key, on the other hand to recover individual records which have a wrong key. Although this procedure is not automatic, it has the advantage of being divided into distinct and replicable phases, so it is applicable for record linkage of other datasets.

To complete the reconstruction of the longitudinal file, the same record linkage procedure was used to attach the information from the other waves to obtain all information for each rotation group, and use them in the correction phase. To do so the sample of twelve-month matched individuals was divided in two sub-samples and each of them was associated with information from different quarters. Indeed, as can be seen from Figure 1-1, the sample of matched individuals after twelve months is composed of two rotation groups, F and G: for group F the 2nd and 4th waves were matched, for group G the 1st and 3rd waves were matched. The remaining waves for each group came from different quarters: thus, to the group F was attached the 1st interview took place in the 4th quarter of 2006 and 3rd took place in the 4th quarter of 2007, to the group G was attached the 2nd interview took place in the 2nd quarter of 2007 and 4th took place in the 2nd quarter of 2008.

After the record linkage process two longitudinal data files with the information from each time of the survey are obtained; those files will be used separately in the correction phase, and finally will be rejoined for analysis of all individuals interviewed in the 1st quarter of 2007 and in the 1st quarter of 2008.

3. Longitudinal imputation

3.1 Longitudinal non-response

In longitudinal survey two types of non-response may arise: item non-response and wave non-response. Wave non-response happens when unit failed to provide data for one wave or more of the survey waves. Item non-response occurs when data items are missing at any wave; values for two or more variables which are logically inconsistent between two waves and response that is out of an accepted range of values are also included. Weighting adjustments are generally used to compensate for wave non-response and imputation is used for item non-response. Nevertheless, the time dimension makes it more difficult to develop a strategy for handling item non-response as well as unit non-response.

The present work is concerned with imputation of item non-response in longitudinal data surveys. It is important to underline that in this case longitudinal item non-response consists only of inconsistent data between two waves as a consequence of the imputation strategy. In fact, the LFS longitudinal data file is obtained from single cross-sectional data files, each of which were previously corrected for item non-response in cross-sectional context (Ceccarelli and Rosati, 2005). In this way the imputation process is divided in more steps in order to reduce computational costs and to ensure timeliness of estimates.

According to the linkage process panel members can be divided into three different groups:
(a) Complete respondents, i.e. units that provided data for all the four waves;
(b) One wave non-respondents, i.e. units that failed to provide data for one out of four waves;
(c) Two wave non respondents, i.e. units that failed to provide data for two out of four waves.

Units which failed to provide data for at least one wave of the reference period for analysis are not included (in this study 2.0 per cent of total "matchable" records as shown in Table 2-1).
In Table 3.1-1 are reported the permissible patterns of the LFS resulted from the linkage process. As the Table 3.1-1 shows, a substantial number of panel members responded in all the four waves of survey (in percentage terms the values for the groups F and G were found to be 81.3 and 85.9, respectively). Those that were respondents in three out of four waves were around 17 per cent for the group F, and 11 per cent for the group G, while only a small proportion participated in two out of four waves (approximately 2.4 per cent considering both of the rotation groups).

Table 3.1-1
Response patterns in four waves for the groups F and G

Response status	wave 1	wave 2	wave 3	wave 4	number	percent
Complete	F1	F2	F3	F4	31,936	81.3
One wave non-respondents	F1	F2	---	F4	2,589	6.6
	---	F2	F3	F4	4,058	10.3
Two wave non-respondents	---	F2	---	F4	704	1.8
Complete	G1	G2	G3	G4	32,413	85.9
One wave non-respondents	G1	G2	G3	---	3,116	8.3
	G1	---	G3	G4	1,050	2.8
Two wave non-respondents	G1	---	G3	---	1,119	3.0

--- non-respondent

It is worth noting that the LFS response rate is quite high (ranging from 76 per cent to 83 per cent in the reference period), denoting that the LFS longitudinal data can be deemed reliable for analysis, especially if three or twelve-month longitudinal data are analyzed, which are about 50 per cent of quarterly sample data.

3.2 Imputation strategy

The literature proposed several classification schemes for imputation methods. However, as Kalton and Kasprzyk discussed (1982), two main categories of imputation can be distinguished: deterministic and probabilistic. *Deterministic imputation* assigns only one value, a priori determined, on the base of other values considered "true" by experts. On the contrary,

probabilistic imputation assigns a value according to a stochastic model (e.g. a regression model) or using a donor unit which is similar to unit to be imputed. Nevertheless, as the same authors noted, most imputation methods fall within a general multiple regression framework (Kalton and Kasprzyk, 1986).
It is well known that many considerations must be examined in order to choose an appropriate imputation methodology. For our purpose we decided to develop a longitudinal deterministic method for several reasons. Considering that interview can be conducted with the actual respondent or with another household respondent, the interview status (proxy, non-proxy) can be used as auxiliary variable for assessing reliability of inconsistent responses across the waves. Secondly, given that CATI is a dependent interviewing method, that is answers from the previous wave are used in the formulation of the question in order to remind the respondent of previous responses or to ask for clarification about inconsistencies between variables, the variable related to confirmation of previous information was also incorporated as auxiliary variable. Furthermore, the entire work history for each individual, from the first wave to the fourth, was investigated in order to derive general deterministic rules of longitudinal imputation. These rules allowed us to deduce the imputed value from data available in current or previous wave on the base of the relation between the auxiliary variables and item non-response.

More exactly, the imputation strategy provides that the errors related to the first two waves can be corrected by processing an algorithm which imputes the incorrect variable changing the value either in the first wave or in the second wave according to the following hierarchical rules:

(1) If CATI technique was used in the second wave, hence "confirm" question was filled in, variable in the first wave was imputed on the base of the response given in the second wave if the respondent denied the response given in the first wave;

(2) If CATI technique was not used in the second wave, variable in the first wave was imputed using the information given by the non-proxy respondent in the second wave; vice versa variable in the second wave was imputed with the value provided by the non-proxy respondent in the first wave;

(3) If neither of the above conditions could be applied, variable in the first wave was imputed assigning the value of the corresponding variable in the second wave. Such a rule was deduced from the examination of consistency of responses to a repeated item across all the four waves.
In the subsequent waves longitudinal imputation consists of using data from the previous wave to impute inconsistent data in the current wave (*conditional imputation*). Although this rule adds further constraints to the entire process, it ensures a basic requirement for the LFS that is to obtain final estimates that not need to be revised when a new wave is available.

3.3 Results

The impact of imputation is briefly summarized in Table 3.3-1. Here only the results concerning the rotation group F are reported. As shown in the Table, the imputation rate associated with the first two waves of interview was equal to 9.8 per cent. In other words, the deterministic rules described above produced 3,838 of 39,287 records with at least one variable changed either in the first wave or in the second wave. As far as the conditional imputation is concerned, two different calculations need to be computed. One refers to the records which represent respondents who participated in three waves or more and it was 9.3 per cent (wave 3 conditioning on wave 2); the other regards the records of complete respondents (i.e. they responded in all the four waves) and it was found to be 13.8 per cent (wave 4 conditioning on wave 3).

It should be noted that many variables involved in imputation concern retrospective questions (e.g. month and year when you started the job, month and year when you lasted the job), which are most affected by errors over time, such as problem of recalling and of panel conditioning (Kalton *et al*., 1989).

Table 3.3-1
Records per type of imputation (group F)

Type of imputation	number	percent
wave 1 or wave 2	3,838	9.8
wave 3 conditioning on wave 2	3,592	9.3
wave 4 conditioning on wave 3	4,415	13.8

3.4 Conclusion

The imputation methodology, that we developed, revealed some good properties, but further research need to be addressed to evaluate the effects of imputation. More analysis is needed in order to assess the impact of imputation on distributions of data.

Even though we based our method on deterministic approach, it seemed to be appropriate for the current LFS. Auxiliary variables were indeed used in order to deduce proper imputation rules, and to reduce bias of imputation. Moreover, the method handled with item non-response for a wide range of variables, and with different patterns of respondents. Nevertheless, conditional imputation was required in order to assure final longitudinal estimates for each quarter.

It is expected that the variables which were mostly changed by imputation were related to dates. For this reason, we suggest introducing a set of longitudinal edit rules into the questionnaire with the aim of reducing inconsistencies during the interview.

We must also mention that another two longitudinal imputation strategies were implemented for the old LFS, which was carried out until the first quarter of 2004. One of them was a combination of deterministic and probabilistic imputation (Rosati, 2004). The experience revealed that generalized methodologies can be adapted to correct longitudinal data, although they could require considerable computational resources. However, this solution seemed to be unsuitable for the new LFS.

We can conclude that our method provided a useful tool for imputation of longitudinal data in large survey which made microdata internally consistent. Although the entire system of imputation, including the linkage process which is an integral part of it, is completely deterministic, it has the advantage of being reproducible for any dataset of the LFS and therefore for any similar survey data.

References

Ceccarelli, C. and Rosati, S. (2006). Data Editing for the Italian Labour Force Survey, Statistical Data Editing: Impact on Data Quality, Vol. No. 3, *United Nations Economic Commission for Europe Work Sessions on Statistical Data Editing*, pp. 291-300.

Kalton, G. and Kasprzyk, D. (1982). Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 22-31.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data, *Survey Methodology*, 12, pp. 1-16.

Kalton G., Kasprzyk D. and Mc Millen, D.B. (1989). Nonsampling Errors in Panel Surveys, in *Panel Surveys*, New York: Wiley, pp. 249-270.

Rosati, S. (2004). Longitudinal Imputation for the Quarterly Labour Force Survey, *Atti della XLII Riunione Scientifica*, *Socieà Italiana di Statistica*, pp. 335-338.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Imputation and Derivation of Longitudinal Registers: The Households Case

Laan, D. J. van der, and Kuijvenhoven, L.[1]

Abstract

Registers are potentially a rich source for longitudinal analyses. However, many of the editing and imputation strategies focus on cross-sectional analyses, creating longitudinal inconsistencies at a micro-level. At Statistics Netherlands households are derived from municipal population registers. However, the derivation and imputation procedure only take information of the current time period into account, making longitudinal analyses impossible. We will present modifications to the present methodology making the households suitable for longitudinal analyses. Special care is take to ensure that both cross-sectional estimates and change estimates are accurate.

Key Words: Registers, Longitudinal, Imputation, Households.

1. Introduction

Registers are potentially a rich source for longitudinal analyses. However, many of the editing and imputation strategies focus on cross-sectional analyses, creating longitudinal inconsistencies at a micro-level. In order to make the data suitable for longitudinal analyses, it is necessary to take into account information from other time periods in the derivation and imputation of variables.

Wallgren and Wallgren (2007) distunguish three quality aspects for statistical registers: cross-sectional quality, time series quality and longitudinal quality. If a register has cross-sectional quality comparisons can be made within the register. Comparisons over time on a aggregated level can be made if the register has time series quality. If a register has longitudinal quality the comparisons can be made at a microlevel over time.

At Statistics Netherlands households are derived from municipal population registers since 2000. Using the (family)relations present in these registers, for approximately 93% of the addresses the household composition can be uniquely determined. The remaining 7% of the addresses are imputed using a stochastic imputation model that takes background properties of the persons living at the addresses into account. However, as the derivation and imputation procedure only take information of the current time period into account, the households are not suitable for longitudinal analyses. We will present modifications to the present methodology that will result in households suitable for longitudinal analyses. Special care is take to ensure that both cross-sectional estimates and change estimates are accurate.

A complication is that in statistical offices one is periodically supplied with new data, while at same time one has to publish periodically. This new data might contain new or better information about previously derived data. We will show that in general it is not possible to obtain a longitudinal consistent register without corrections to previously derived data. Plans on how to handle this at Statistics Netherlands are presented.

2. Deterministic derivation of households

The figures on households as published by Statistics Netherlands are based on the household keeping definition. This definition implies that only those persons who reside on the same adress and who provide for their own daily needs are considered as one household.

The Dutch population and household statistics compiled by Statistics Netherlands are based on the automated municipal population registers. This registration system is known as the GBA system, which stands for ‘Gemeentelijke Basis

[1]D.J. van der Laan, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, (dj.vanderlaan@cbs.nl); L. Kuijvenhoven, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, (l.kuijvenhoven@cbs.nl)

Administratie persoonsgegevens', the municipal basic registration of population data. 'Basic' refers to the fact that the GBA serves as the basic register of population data within a system of local registers. For more detailed information on the GBA system, see Prins (2000).

The GBA has information about family ties. Every personal record contains information on parent(s) and of all children born, irrespective of their present residence. There is also information about the partner of the person. Together with the detailed address information it is possible to identify all traditional nuclear families. This leads to the following deterministic rules:

D1 If two persons belong to the same family and live on the same address, then they belong to the same household.
D2 If two persons are related to each other and live on the same address, then they all belong to the same household.
D3 If two persons move to a new address at same date then they belong to the same household.
D4 If a person aged 15 or younger without an identifiable parent lives at an addresses with an other household, then he belongs to that household.

3. The use of additional sources

One of the modifications to the current derivation is the use of additional sources in order to derive more households deterministically. The most important sources come from the department of Finance. From this department information is obtained on the use of (1) tax support systems on medical care, (2) on housing, and (3) on general tax-schemes. All three of these systems have in common that the applicant indicates with whom he/she forms a household. This information can be linked to the population registers. An additional rule can now be created:

D5 If it is known from a source other than the GBA that two persons living on the same address are a couple, then they belong to the same household.

Using the rules D1-D4 with the population registers, of about 92.6% of the addresses the household position and household composition can be derived. The use of additional sources and applying rule D5 leads to 94.0% directly determined households (see table 3-1). Especially the number of couples in the categories 'two unattached persons' and 'lone parent family + unattached person' befenit a great deal from the use of thes additional sources.

Table 3-1
Reduction in the number of addresses that need to be imputed when using the additional sources. The data concerns 2006.

Source	Number of persons in source	Number to impute	Reduction
Initial situation	-	7,4%	-
Health support	2,075,460	6,6%	11,3%
Rent support	886,802	7,3%	2,0%
Provisional tax refund	3,065,367	6,7%	9,7%
All three sources	-	6,0%	19,5%

4. Stochastic imputation

As was discussed in the previous section, for 7.4% (6.0% using the additional sources) of the addresses it is not possible to completely derive the household composition using deterministic rules. This means that at the end of the deterministic derivation, we are left with more than one 'partially derived household' at these addresses, and we have decide whether or not we want to further combine these households into larger households.

It was decided that the composition of these addresses should be imputed. Unfortunately, this is a special group of addresses. For example, a large group in these addresses are the unmarried couples. Therefore, we can not use the information of the fully derived addresses, and we need additional data that gives information about the remaining addresses in order to estimate the model that is used for the imputation. For this the Labour Force Survey (LFS) is used. The LFS uses a random sample of addresses. All households at the address are approached using interviewers. Therefore, in principle the complete household composition is known of these addresses. The LFS is coupled to the remaining addresses and the household composition of these addresses is derived.

For the imputation a model is estimated that predicts the household composition using background properties of the address (urbanisation) and the persons living at the address (age, gender, ethnicity). Separate models are estimated for each of the following groups: two persons, family and a person, lone parent and a person, three persons, and four to nine persons. These groups are sufficiently large for accurate modelling, and it is expected that the probabilities for the different possible compositions will differ.

In most of these cases there are two possible compositions: either they form one single household, or two separate households. Therefore, a logistic regression model was used to predict the probabilities for each of the two possible compositions. For three persons at an address, two logistic models were used, one that predicts the probability of two or three households versus one household and one that predicts the probability of three households versus one or two households. The group of four to nine persons at an address is too small to estimate a model that uses background properties. Therefore, the observed frequencies were used as probabilities.

Using the estimated models, for each address that needs to be imputed probabilities for each of the possible household compositions can be calculated. These were used to stochastically imputed the composition at these addresses. Stochastic imputation was chosen over deterministic imputation, to guarantee that the aggregates are unbiased, although deterministic imputation (imputation of the most probable composition) would lead to more correct imputation at a micro level. For the stochastic imputation we used the method of Fellegi (1975) that gives the same totals as following from the model.

5. Longitudinal consistency

There are some problems with the current methodology as described above when the derived households from different periods are used for longitudinal analyses. As the methodology is applied for each period separately, households that are imputed in different periods are not guaranteed to be consistent with each other. For example, with the current methodology it is possible that two persons living at an address are one year imputed as two single person households and the next year as a couple. This leads to an overestimation of change when using the data for longitudinal analyses.

In order to avoid longitudinal inconsistencies data from multiple time periods need to be combined in the derivation and imputation. It is reasonable to assume that the relation between two persons (belonging or not belonging to one household) does not change as long as these persons live at the same address. This assumption leads to the following longitudinal rules:

L1 If two persons A and B belong to the same household at time t_j, and are living at the same address at time ti ($i=j$-1 or $i=j$+1), then A and B belong to the same household at time t_i.
L2 If two persons A and B do not belong to the same household at time t_j, and are living at the same address at time ti ($i=j$-1 or $i=j$+1), then A and B do not belong to the same household at time t_i.

Suppose we have a complete set of longitudinal consistent households for time periods t_1 and t_2. Using the longitudinal rules, the algorithm for the derivation for time period t_3 then is the following:

1. Apply deterministic rules (see section 2) for time period t_3.
2. Transport deterministically derived information from period t_2 to period t_3 (=forward) using longitudinal rule L1.
3. Transport deterministically derived information back to time periods t_2 and subsequently t_1 (=backward) using longitudinal rule L1.
4. Correct stochastic imputations of time period t_1, and transport stochastically imputed information from time period t_1 to time period t_2 using rules L1 and L2.
5. Correct stochastic imputations of time period t_2, and transport stochastically imputed information from time period t_2 to time period t_3 using rules L1 and L2.
6. Impute remaining addresses.

In the algorithm we also use the new information to correct the derivation of previous periods. This is necessary to maintain longitudinal consistency. For example, two persons living at an address imputed as two single person households get married. From the moment the got married we know that they form one household. To maintain longitudinal consistency with earlier periods, it is necessary to change the imputation of two single person household. It is important to remark that in step 4 and step 5 the logistic regression models are re-estimated with the newly required data. These models are subsequently used for new imputations.

An other benefit of using information form multiple time periods is an increased number of deterministically derived households. Table 5-1 shows the reduction in the number of addresses that need to be imputed when using information from other periods. For example, the number of addresses that need to be imputed in 2006 is reduced by 2.5% when we use information from 2002–2005, and the number of addresses that needs to be imputed in 2002 is reduced by 23.4% when using information from 2003–2006. These are the results when we do not require that the persons A and B in the rules L1 and L2 live at the same address at both time periods. When we do require that the addresses are the same the reduction is smaller. From the results it is clearly visible that new data contains more information about previous periods than the other way around.

Table 5-1
Reduction in the number of addresses that need to be imputed when using the longitudinal rules. The data spans the period 2002–2006.

	Forward [a]		Backward [a]	
	Unrestricted [b]	Restricted [b]	Unrestricted [b]	Restricted [b]
2002			23.4%	10.5%
2003	1.1%	0.7%	21.7%	10.0%
2004	1.8%	1.1%	18.1%	8.4%
2005	2.2%	1.3%	12.2%	5.9%
2006	2.5%	1.4%		

[a] Forward/backward: deterministic information from one time period is used in (later/earlier) time periods to deterministically derive households
[b] Unrestricted/restricted: when transporting information persons to other time periods address changes (are allowed/are not allowed)

6. Discussion

In this paper we have presented a new approach for deriving longitudinal consistent households. This new methodology is based on the current methodology which has time-series quality and not longitudinal quality. We plan to investigate additional sources besides the three sources investigated here in order to increase the number of deterministically derived households. A disadvantage of an increased number of deterministically derived households, is that the amount of data that can be used to estimate the models for the stochastic imputation decreases, making these models less accurate. This is something that has to be investigated further. This question is related to the stochastic imputation method used, which is now cross-sectionally based, but could be based on longitudinal analyses methods. This also needs to be investigated further.

Another important point of discussion is how far does one transport new information back to earlier periods thereby changing (already published) previously derived data. Besides complications with publications (there are now different versions of data for the same period), this also complicates data storage and version control. This discussion has not been decided yet.

References

Fellegi, I. P. (1975). Controlled Random Rounding. *Survey Methodology,* 1, pp. 123–133.

Van der Laan, D.J. and Kuijvenhoven, L. (2009). Improving the Derivation of Households. (Vernieuwing Afleiding Huishoudens), Internal Report, CBS, The Hague.

Prins, C.J.M. (2000). Dutch population statistics based on population register data, *Maandstatistiek van de bevolking*, 48(2), CBS, Heerlen/Voorburg.

Wallgren, A. and Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*, Chichester: Wiley.

APPLICATION: LONGITUDINAL ANALYSIS OF HEALTH AND BUSINESS DATA

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

The Children of Older First-time Mothers in Canada: a Longitudinal Analysis of their Health and Development

Tracey Bushnik and Rochelle Garner[1]

Abstract

Using a sample of 3,382 first-born children from the National Longitudinal Survey of Children and Youth (NLSCY), this study examined the relationship between late childbearing (having a first child at age 35 or older) and the health and development of children aged 0 through 5. Several methodological issues were identified and addressed including decreasing sample sizes at older age groups; selection of appropriate survey weights; impact of the study's sample selection; and cohort effects. The results suggested that while children of older mothers were more likely to be exposed to developmental risk factors such as preterm birth, with few exceptions their outcomes were not significantly different from those of children in the reference group.

Key Words: Longitudinal Analysis, Late Childbearing, Children's Health, Children's Development.

1. Introduction and objectives

In Canada, it is increasingly common for women to delay childbearing. Not only are more women over thirty giving birth, but the proportion of first births occurring among women over thirty has increased steadily over the past 20 years (Statistics Canada, 2007).

It is important to understand the possible implications of these trends for human health and development. While many studies have examined the pregnancy and health-related outcomes of late childbearing for women, less is known concerning the potential consequences for their children.

To help fill this gap, data from the National Longitudinal Survey of Children and Youth (NLSCY) were used to examine the relationship between late childbearing among first-time mothers and three facets of children's development: (i) physical health and development, (ii) behaviour, and (iii) cognitive development. Late childbearing was defined as giving birth to a first child at or after age 35.

This proceedings document is based on a lengthier report entitled "The Children of Older First-time Mothers in Canada: Their Health and Development" that was published by Statistics Canada in September 2008 (Bushnik and Garner, 2008). The main purpose of the 2008 report was to present the results of a detailed analysis of the relationship between late childbearing and an extensive list of children's outcomes. The present document will focus primarily on the methodological challenges of the analysis – providing some detail that was not included in the 2008 report – and will touch only briefly on the analytical results.

2. Methods

2.1 Data source and participants

The National Longitudinal Survey of Children and Youth (NLSCY) is a long-term study of Canadian children that follows their development from birth to early adulthood. The NLSCY began in 1994 and is conducted by Statistics Canada and sponsored by Human Resources and Skills Development Canada. The survey covers a broad range of topics including health, physical development, learning, behaviour, and the social environment.

For children under 16 years of age, most of the information in the survey is provided by the person most knowledgeable about the child (known as the PMK), usually the mother. She provides information about herself, the household and family, and the child. Direct measures of the child's abilities are also taken, depending on the child's age.

[1]Tracey Bushnik, Health Analysis Division, Statistics Canada, K1A 0T6; Rochelle Garner, Health Analysis Division, Statistics Canada, K1A 0T6.

The full NLSCY sample consists of the original longitudinal cohort, who have been followed biennially since they were 0 to 11 years old in 1994, and several early child development cohorts, who are interviewed at ages 0 to 1 and followed every two years until the ages of 4 and 5. For the present study, children who were part of the early child development cohorts recruited in Cycles 3 through 6 (1998 through 2005) were selected and pooled together, resulting in a sample of 18,907 children 0 to 1 years of age. From this pooled sample, only first-born children whose biological mother completed an interview at each cycle were retained for this study. This resulted in 3,382 children 0 to 1 years of age in the study sample. Of these children, 2,365 were re-interviewed at ages 2 to 3, and 1,705 were interviewed a third time at ages 4 to 5. Table 2.1-1 shows that, while all cohorts of first-born children in this study had data for characteristics measured at ages 0 to 1, only cohorts from Cycles 3 through 5 had data at ages 2 to 3, and only those cohorts introduced in Cycles 3 and 4 had data at ages 4 to 5.

Table 2.1-1
Age of child at each cycle of NLSCY interview

Cohort of entry into sample	Child's age at cycle of interview			
	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Cycle 3	0 to 1	2 to 3	4 to 5	...
Cycle 4	...	0 to 1	2 to 3	4 to 5
Cycle 5	...	...	0 to 1	2 to 3
Cycle 6	...	...	...	0 to 1

… not applicable

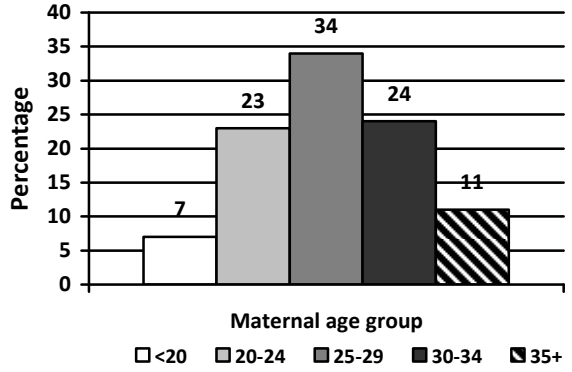
2.2 Measures

2.2.1 Explanatory variables

Maternal age group was the explanatory variable of interest, and was based on a woman's age when she had her first child. The following groups of women and their children were defined: (i) *teenaged mothers* (under age 20 at first birth); (ii) *young mothers* (aged 20 to 24 at first birth); (iii) *reference mothers* (aged 25 to 29 at first birth); (iv) *middle mothers* (aged 30 to 34 at first birth); and (v) *older mothers* (aged 35 or older at first birth). The weighted distribution of children by maternal age group is presented in Figure 2-1.

Additional explanatory variables included: socio-demographic characteristics (household low-income status, maternal level of education, number of parents in the household); perinatal and postnatal measures (child's gestational age, child's birth weight, duration of breastfeeding); and parenting practices, family functioning, and maternal mental health.

Figure 2.2-1
Distribution of children aged 0-1 by maternal age group



2.2.2 Outcome variables

The outcome variables of interest included health measures at various ages (receipt of specialized care at birth; general health status; prevalence of diagnosed asthma) and development (developmental milestones; motor and social skills), behaviour at ages 2 to 5 (physical aggression; emotional disorder and anxiety, hyperactivity and inattention, and positive behaviour), and cognitive development measures at ages 4 to 5.

2.3 Methodological issues

Given the study's sample selection strategy, several issues had to be addressed including: decreasing sample sizes (fewer cohorts) at older age groups; identification of appropriate survey weights; assessment of impact of study's sample selection; and cohort effects.

2.3.1 Decreasing sample sizes

The pooling of the four cohorts created a sample sufficiently large to examine children's characteristics over time, by maternal age group. However, not all cohorts had data measured at all time points. While all cohorts had data measured at ages 0-1, the sample size and number of cohorts decreased at older ages (Table 2.1-1). This had two main consequences. First, there were increased variances of estimates at older ages, which affected the statistical significance of certain differences. Second, inferences that were made about 2 to 3-year-olds or 4 to 5-year-olds pertained to a more limited target population. Specifically, the 2 to 3-year-olds in this study represented the population of children aged 0-1 in January of the years 1999, 2001 and 2003. The 4 to 5-year-olds represented the population of children aged 0-1 in January of the years 1999 and 2001.

2.3.2 Selection of survey weights

A single survey weight was selected per respondent. All weights were taken from the *last* cycle in which the child participated (Table 2.3.2-1). This permitted inferences to be drawn about the population of Canadian children represented by the sample that entered the survey at ages 0 to 1. The corresponding bootstrap weights were used for variance estimation and significance testing.

Table 2.3.2-1
Selected survey weights

Cohort of entry into sample	Survey weight
Cycle 3	Longitudinal weight from cycle 5
Cycle 4	Longitudinal weight from cycle 6
Cycle 5	Longitudinal weight from cycle 6
Cycle 6	Cross-sectional weight from cycle 6

2.3.3 Assessment of study sample selection effects

The study's sample selection criteria resulted in the exclusion of certain respondents. In effect, these excluded children became non-respondents. Given that non-response has the potential for biasing results if non-respondents have significantly different characteristics from respondents, it was important to assess the impact of the selection criteria on the analysis.

There were 625 first-born children whose biological mother was the PMK during the first interview who were lost to attrition, and thus excluded from the study. There were also 396 first-born children whose biological mother was the PMK during the first interview, but whose PMK changed over the course of subsequent interviews. These children were also excluded from the study. The former group of children were referred to as 'lost to attrition' while the latter group were referred to as 'PMK changed over time'.

The characteristics of the children in these two subgroups were compared to those of the 3,382 children retained in the study. These characteristics had been measured at ages 0 to 1 and included the socio-demographic, prenatal, perinatal, and health characteristics of interest to the main analysis.

The non-response analysis indicated that children excluded from the study because the PMK changed over time did not differ significantly from the final sample on most of the examined characteristics. There were, however, several significant differences between the children lost to attrition and children in the study sample. For the most part, all of these differences reflect the fact that children who do not continue their participation in the NLSCY over time were more likely to be from lower socio-economic (SES) backgrounds than those who continued to participate. There are two main reasons why possible bias resulting from the loss of these children was not of concern to the main study. First, the longitudinal weights provided with the data had been adjusted to address this type of attrition. Second, the children born to older mothers were, on average, from relatively high SES backgrounds and therefore were less likely to drop out of the survey. Bias may have

been more of a concern if the focus of this paper had been on children of teenaged mothers rather than children of older mothers. Given their lower SES backgrounds, the former had an increased likelihood of being under-represented in the selected sample.

2.3.4 Cohort effects

Due to the number of cohorts that were pooled for the analysis, it was necessary to check for cohort effects. Characteristics and outcomes were compared across cohorts by maternal age group. The results suggested that there was little systematic variation by cohort. The limited variation that was observed was further analyzed to determine whether differences in the underlying age-in-months distributions between the cohorts were responsible for the minor differences. This did not seem to be the case. Regardless, a cohort dummy variable was included in all multivariate analysis to control for possible cohort effects.

3. Analysis and results

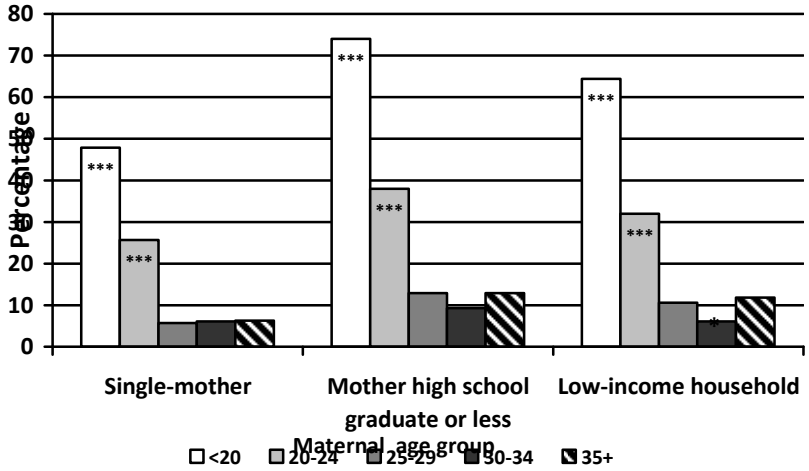
For each outcome measure, means or proportions were generated overall and by maternal age group. In all analyses, children whose mothers were aged 25 to 29 at the child's birth formed the reference category (referred to as *reference mothers*). Subsequently, multivariate regression models were generated for each outcome. Included in each model was the mother's age at the child's birth and a group of socio-demographic characteristics. In addition, depending on the outcome examined, certain maternal characteristics or behaviours shown in the literature to be associated with that particular outcome were included in the model. As mentioned above a cohort control variable was also included in each model.

Logistic models were fit to dichotomous outcomes, while linear models were fit to continuous outcomes. All analyses were weighted (see section 2.3.2) to generalize to the Canadian population. To account for the complex survey design of the NLSCY, bootstrap weights were used to produce variance estimates. All analyses were conducted using SAS-callable SUDAAN.

3.1 Socio-demographic characteristics

Children of older mothers shared a similar socio-demographic profile with children born to reference mothers (Figure 3.1-1). Children born to teenaged and young mothers, however, were significantly more likely to live with a single-mother, to be in a low-income household, and to have a mother with no more than a high school education compared to children of reference mothers.

Figure 3.1-1
Socio-demographic characteristics at ages 0-1 by maternal age group

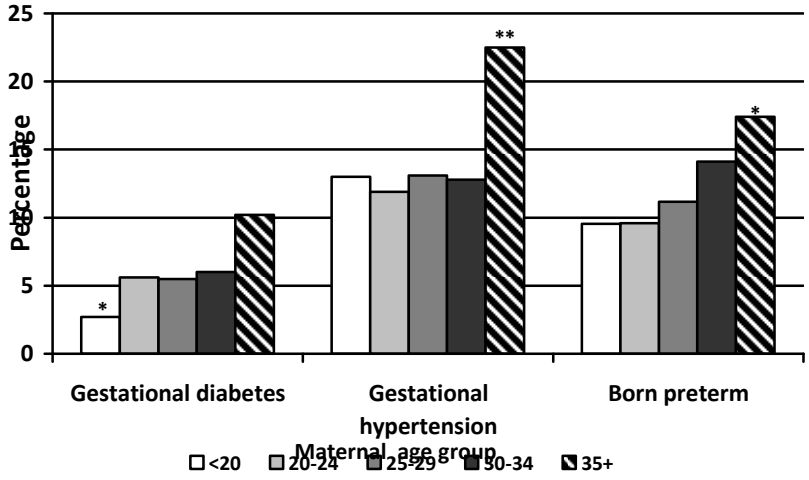


*significantly different from reference group at p<0.05
***significantly different from reference group at p<0.001

3.2 Prenatal risk factors and birth-related characteristics

Children of older mothers were more likely to have a mother who suffered from gestational hypertension and to be born preterm (Figure 3.2-1) than children of reference mothers. Caesarean deliveries were more common among children of older mothers: about 40% of children of older mothers were delivered by caesarean compared to 23% of children of reference mothers.

Figure 3.2-1
Prenatal risk factors and preterm birth by maternal age group



*significantly different from reference group at p<0.05
**significantly different from reference group at p<0.01

3.3 Late childbearing and children's outcomes

While not presented here, the results of the analysis showed that children of older mothers were no different from children of reference mothers with respect to many developmental outcomes. For example, children of older mothers were as likely as children of reference mothers to have received special care at birth, to be in excellent or very good health during early childhood, or to be diagnosed with asthma. They shared similar timing with respect to saying their first word and taking their first step and had similar averages scores in physical aggression, emotional disorder and anxiety, and hyperactivity and inattention. Children of older mothers also had similar scores in cognitive development as children of reference mothers.

However, advanced maternal age was significantly associated with other outcomes. Even after controlling for a number of characteristics, a higher proportion of children of older mothers were considered as late achievers in sitting up by themselves compared to children of reference mothers. In addition, children of older mothers scored lower on the motor and social development scale at ages 0 to 1 and 2 to 3. They also had lower positive behaviour scores at ages 4 and 5.

4. Discussion

Using the NLSCY, this study assessed the relationship between late childbearing and children's outcomes over time. The results suggested that while children of older mothers were more likely to be exposed to developmental risk factors such as preterm birth, with few exceptions their outcomes were not significantly different from those of children in the reference group.

The analysis was not without its methodological challenges. Attrition and its effect on sample sizes and inference, the appropriate use of survey weights, and the potential for cohort effects needed to be acknowledged and addressed.

There were some limitations to the study. They included: a possible bias towards healthy children because the NLSCY is a household survey and does not sample the institutionalized population; the use of maternal-report for certain outcomes; and the relatively short period of children's development (ages 0 to 5) that was examined.

Despite these limitations, this study offered novel insight into the relationship between late childbearing and children's development in the Canadian context.

References

Bushnik, T. and Garner, R. (2008). *The Children of Older First-time Mothers in Canada: Their Health and Development*, Ottawa, Statistics Canada. Catalogue 89-599-M, no. 005.

Statistics Canada (2007). *Births: Live Births 2005*, Statistics Canada Catalogue no. 84F0210XIE. Ottawa.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Life Course BMI and Height Trajectories: A Comparison of Two British Birth Cohorts

Leah Li, Rebecca Hardy, Diana Kuh and Chris Power[1]

Abstract

Obesity continues to increase worldwide. The development of BMI trajectories may have changed across recent generations experiencing the obesity epidemic at different life stages. Other components of physical development may also have changed. We compared child-to-adult growth trajectories across two British birth cohorts, born in 1946 (n=5,300) and 1958 (n=17,000), followed-up to ages 53y and to 45y respectively.

Individuals born in 1958 were not heavier at birth than the 1946 cohort, but were taller in early childhood by 1cm, grew faster and were 3-4cm taller by adolescence. The 1958 cohort achieved adult height earlier, were taller by 1cm, an increase entirely due to their longer leg length. We adopted linear spline models to repeated BMI measures (at 7, 11, 15, 20, 26, 36, 43, and 53y for the 1946 cohort, at 7, 11, 16, 23, 33, and 45y for the 1958 cohort) corresponding to distinct BMI trajectories for "childhood'' and "adulthood''. BMI trajectories diverged from early adulthood, with a faster growth rate in the 1958 cohort than the 1946 cohort, although mean BMI at 7y and rate of childhood gain had not shown an increase between two cohorts. By mid-adulthood the 1958 cohort had a greater BMI (1-2 kg/m2), larger waist (7-8cm) and hip (5cm) circumferences, and a higher prevalence of obesity (24% vs 12%). These changes over a relatively short period of 12 years suggest the likelihood of opposing trends of influences on later disease risk in these populations.

[1] Leah Li (l.li@ich.ucl.ac.uk), Rebecca Hardy, Diana Kuh and Chris Power, University College London, U.K.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Impact of training on the productivity of Canadian businesses in a longitudinal context: Comparison of an additive model and an interactive model

Amélie Bernier et Jean-Michel Cousineau[1]

Abstract

Using employer data from the Workplace and Employee Survey from 1999 to 2005, we estimate the effects of training expenditures on business productivity. Our findings show that investments in training have positive effects extending over a period of three years.[2] The interaction between investments in physical capital and training serves to verify the hypothesis that investments in physical capital and investments in human capital are complementary and mutually supportive.

Key words: Longitudinal study, recursive model, training, productivity, firms

1. Introduction

The face of the Canadian workplace is diversifying and changing. Owing to factors such as new technologies, globalization, the knowledge-based economy and demographic change, Canadian firms are undergoing a transformation, and the hallmark of this transformation is a skilled, adaptable and high-quality workforce. If one examines the formal training provided by the Canadian firms in our sample for the period 1999 to 2005, it may be seen that on average, 32.5% of firms offer this type of training. A number of firms appear to question the necessity and value of participating in continuous training (Goldenberg, 2006). Is this because they doubt that their investment in training will affect their productivity in the short term and their relevance in the medium term?

This study uses data from the Workplace and Employee Survey (WES) to estimate the effects of training expenditures on short- and medium-term productivity. To the best of our knowledge, our study is the only one that uses a recursive model to estimate returns on investment in formal training on the basis of data from Canadian firms. Our contributions are as follows. First, to be part of the current trend in research on returns on training investment, it is essential to use the longitudinal nature of data, so as to measure the possible effects of such investments beyond the current year of the investments made by firms. Next, to adequately measure the returns on training investment in firms, it is first necessary to consider the technical problems related to the idiosyncratic characteristics of the firms and the simultaneity of training with the dependent variable used. Our lagged-effect model serves to capture most of the problem of endogeneity. Lastly, it is relevant to validate empirically the hypothesis that training investment is complementary with other types of investment within firms, since the results of the longitudinal studies consulted are not always conclusive (Zwick, 2006).

2. The problem

First, when a firm invests in training, it hopes to obtain either monetary gains or an improvement in the quality of its workforce. With reference to the theory of human capital and the cost sharing model of Becker (1964), additional training for workers is potentially a means of increasing the firm's productivity, via the marginal value of post-training productivity as well as the monetary gains of the employees themselves. Although the theory of human capital has largely shown its empirical usefulness in explaining the returns on post-secondary education, we are inclined to believe that it can, simply and convincingly, shed light on the lagged effects of training, in particular by applying a recursive model to firm-level data.

This approach does not propose clearly identified solutions for dealing empirically with the potential problem of the endogenous[3] nature of training. The effect of this is to generate a huge debate on the empirical aspect. A review of the

[1] Amélie Bernier, Candidate au doctorat (Ph.D.), École de relations industrielles, Université de Montréal, Canada, courriel : amelie.bernier@umontreal.ca ; Jean-Michel Cousineau, professeur titulaire, École de relations industrielles, Université de Montréal, Canada, courriel : jean.michel.cousineau@umontreal.ca
[2] Although the research and analyses are based on Statistics Canada data, the opinions expressed are those of the authors only.
[3] When we speak of endogeneity, we are referring to so-called independent variables that are correlated with the error term.

literature (Colombo and Stanca, 2008; Zwick, 2006) revealed that the possibly endogenous nature of training is often cited to explain differences between estimation results.

This paper also opens the way to a deeper understanding of the possible links between the different types of investments, making it possible to verify the hypothesis that some factors of production are complementary and mutually supportive in improving productivity.

Empirically, the challenges are many. In the various works consulted, there seems to be no clearly defined relationship between the costs and benefits of training, and there appears to be no consensus as to returns on investment (Ballot, 2006). The fact is that the links between training and productivity are not so obvious, as we shall show. Although a growing number of studies have attempted to measure the impact of training on various corporate performance indicators using representative data at the firm level (Colombo and Stanca, 2008; Dostie and Pelletier, 2007), the results are not always conclusive. In the various studies reviewed, there is no unanimity regarding 1) the actual measurement of the "training" variable, 2) the type of training to be considered for analysis purposes, 3) the results obtained and 4) the estimation techniques used. What is also missing, in our view, is a way to determine the link between firms' past investments in training and their current productivity. It is necessary to establish this link if we want to understand the effects of investment in training for Canadian firms.

3. Data and method

To estimate our models on the impact of training on the productivity of Canadian firms, we work with a sample consisting of 1,621 firms and 11,347 observations. The data used are drawn from the WES questionnaire of employers for the period from 1999 to 2005. In summary, our sample consists of all for-profit firms which have at least one employee and have reported positive financial returns, and for which we have information regarding their training decisions. In the next section we will present the results obtained.

To estimate the returns on training at the corporate level, we favour a Cobb-Douglas function[4] presented as a production process (Q_{it}) serving to relate value added[5] to training expenditures within organizations. A Cobb-Douglas function with production (Q_{it}) as the dependent variable can be used to study the impact of training on productivity, since employment (or the labour factor) is an independent variable with constant returns, which in our view imposes a more flexible and less constraining solution in estimating the effect of training expenditures on productivity.

Following the example of Almeida and Carneiro (2006) and Barrett and O'Connell (2001), our basic model, characterized by a firm i in year *t*, takes the following form:

$$\ln Q_{it} = \ln A_{it} + \alpha \ln L_{it} + \beta \ln I_{it} + \gamma \ln F_{it} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \qquad (1)$$

Where i = 1,…, N; t = 1,…, T; α; β; γ; δ; η are parameters to be estimated.

Q_{it} represents the value added,[6] which is a function of three main factors (L_{it}; I_{it}; H_{it}) and of a parameter of scale (A_{it}). First, labour (L_{it}) is measured by the total workforce of the firm. Next, we use investments in physical capital (I_{it}) measured by total expenditure on equipment[7] within organizations (Barrett and O'Connell, 2001; Colombo and Stanca, 2008). Investments in human capital (F_{it}) are represented by a ratio of expenditures on formal training to the total workforce of the firm.[8] The WES employer questionnaire yields data on two major training categories: formal training and on-the-job

[4] The main characteristic of the Cobb-Douglas function is that the elasticity of substitution is equal to 1 and remains constant along the isoquant. A change in the marginal rate of substitution leads to a proportional change in the factor quantities ratio. With reference to the present analysis, this elasticity of substitution equal to 1 assumes that the factors are as much complements as they are substitutes.

[5] Value added is determined by the gross value of output, from which we have subtracted the cost of the inputs (Dostie and Pelletier, 2007).

[6] To compensate for the effect of inflation on nominal variables such as gross operating revenue, equipment expenditures and expenditures on formal training, we have deflated the values by the annual Consumer Price Index for Canada (where 1992=100), as reported by Statistics Canada (CANSIM table no. 326-0002).

[7] The equipment expenditures variable was constructed using gross operating expenditures, from which we subtracted gross payroll, expenditures on employee benefits and expenditures on formal training.

[8] Although training expenditures are little used in the earlier empirical literature, they are a reliable indicator of the intensity of the training provided within business locations. The reliability of our data is based in particular on the proven methodology of Statistics Canada, which enables us to obtain high response rates and ensure data quality. To cite an example, the response rate of employers in 2005 was 77.7%. Another way to measure data reliability is to

training. Thus, regardless of whether the training is general or specific, the information available in the WES instead suggests a differentiation as to the predetermined nature of the content and the objectives that can be evaluated rather than the theoretically prescribed division.

Although imperfect, a technology variable (T_{it}) representing the proportion of employees using a computer in their daily work was added to the model. Finally, the equation estimated on the basis of specification (1) also includes a set of control variables estimated by a vector (X_{it}).

To get an overview of the possible effects of training on firms' performance, we must first estimate equation (1) by the ordinary least squares (OLS) technique. Also, without entirely calling into question the idea that differences in results in longitudinal surveys are due to the fact that training is endogenous to the dependent variable, we propose to verify the predetermined or endogenous nature of training in the context of our estimates. In this case, the following are some of the options considered for treating the violation of the assumptions underlying the OLS: the use of instrumental variables (IVs), random or fixed effects (respectively REs and FEs), correction for first order autocorrelation (AR1), the development of dynamic models using the generalized method of moments, and analysis of longitudinal data.

As already pointed out, the lag of one year (or more) is necessary since a problem of endogeneity would arise if decisions regarding training investments and corporate productivity were determined simultaneously. Thus, since we assume that there may be a considerable time lag between the decision to invest in training and its return on investment in terms of productivity growth, we propose a model in which the time lag is specified (t = 4). The variable on investments in physical capital (I_{it}) will also be subject to a time lag. To apply these time lags, we must rewrite equation (1) as follows:

$$\ln Q_{it} = \ln A + \alpha \ln L_{it} + \sum_{j=1}^{4} \beta_j \ln I_{it-j} + \sum_{j=1}^{4} \gamma_j \ln F_{it-j} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \qquad (2)$$

Where j varies from 1 to 4 years; $b_1 = \sum_{j=1}^{4} \beta_j$; $b_2 = \sum_{j=1}^{4} \gamma_j$

To validate the hypothesis that investments in training (F_{it}) and investments in physical capital (I_{it}) are complementary in their effects on productivity, we include a multiplicative term represented by an (*) symbol between F_{it} and I_{it} which we can see in equation (3):

$$\ln Q_{it} = \ln A_{it} + \alpha \ln L_{it} + \sum_{j=1}^{4} \beta_j \ln I_{it-j} + \sum_{j=1}^{4} \phi_j \ln I_{it-j} * \ln F_{it-j} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \qquad (3)$$

Where j varies from 1 to 4 years; $b_1 = \sum_{j=1}^{4} \beta_j$; $b_3 = \sum_{j=1}^{4} \phi_j$

The following section presents the results obtained for our various estimates.

4. Results

4.1 The endogenous nature of training

When the Nakamura-Nakamura endogeneity test (Nakamura and Nakamura, 1998) is conducted, it shows that the results are on the margin of the threshold for accepting the null hypothesis (absence of endogeneity). A second estimator available to us for testing the presence of endogeneity is the use of the Hausman test. Once again, the results obtained suggest that we should reject the null hypothesis of exogeneity of the independent variables, which leaves the possibility that the training variable is endogenous to productivity.[9] To correct this potential bias, training should be estimated over a long period of

calculate the coefficient of variation. If the value obtained is less than 16.5%,[8] the estimate of the variable F_{it} will be considered reliable. In our case, the coefficient of variation is 3.57%, which indicates homogeneity in the data used.

[9] The results of these estimates are available on request.

time, to document all possible returns; hence the value of using a recursive model to measure returns on investments in training with a time lag of more than one year. For example, Model 2 (Table 4.1-1) shows that the exogenous part of the training variable has a positive and significant effect on productivity. Thus, it may be seen that a 10% increase in expenditures on formal training per employee generates a 1.7% increase in the productivity of the firm for the following year.

Table 4.1-1
Results of estimates of the impact of training expenditures on the productivity of firms in Canada from 1999 to 2005

Dependent variable: ln Production (value added)	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	OLS	**IV**	**RE**	**FE**	**AR1, RE**
Independent variables					
Ln (total number of employees)	0.8306***	1.0081***	0.8069***	0.5611***	0.8096***
	(0.0253)	(0.0183)	(0.0295)	(0.0592)	(0.0289)
Proportion of employees using a computer (percentage)	0.0040***	0.0025***	0.0025***	0.0012*	0.0026***
	(0.0007)	(0.0004)	(0.0006)	(0.0006)	(0.0006)
ln (investment in physical capital)		0.0346***			
		(0.0117)			
ln (expenditures on formal training/total number of employees) $_{t-1}$	-0.0119	**0.1656*** **	-0.0006	-0.0081	0.0002
	(0.0253)	**(0.0203)**	(0.0136)	(0.0157)	(0.0138)
ln (expenditures on formal training/total number of employees) $_{t-2}$	0.0047		0.0175	0.0029	0.0178
	(0.0283)		(0.0131)	(0.0151)	(0.0132)
ln (expenditures on formal training/total number of employees) $_{t-3}$	**0.0340**		**0.0374*** **	**0.0206**	**0.0380*** **
	(0.0270)		**(0.0124)**	**(0.0141)**	**(0.0126)**
ln (expenditures on formal training/total number of employees) $_{t-4}$	0.0341		0.0094	-0.0086	0.0107
	(0.0238)		(0.0124)	(0.0137)	(0.0126)
ln (investment in physical capital) $_{t-1}$	0.0177		0.0659***	0.0282	0.0683***
	(0.0304)		(0.0165)	(0.0205)	(0.0167)
ln (investment in physical capital) $_{t-2}$	0.0538*		0.0510***	0.0051	0.0508***
	(0.0330)		(0.0159)	(0.0190)	(0.0161)
ln (investment in physical capital) $_{t-3}$	0.0397		0.0338***	-0.0049	0.0338***
	(0.0286)		(0.0128)	(0.0146)	(0.0130)
ln (investment in physical capital) $_{t-4}$	0.0542**		0.0260**	-0.0007	0.0279***
	(0.0218)		(0.0109)	(0.0116)	(0.0111)
Total number of observations (N)	1555	4477	1555	1555	1555
Total number of firms (n)		1080	620	620	620
R square of the model	0.7814	0.7649	0.7728	0.6839	0.7738

Notes:
1. Robust standard deviations in parentheses. *** significant at 1% ** significant at 5% * significant at 10%
2. Also included in the models are a constant, dichotomous variables for industry (14), time and control variables. Complete results are available on request.

4.2 The deferred effects of training on productivity

In the lagged-effect models, the lag structure is t-1 to t-4. Time "t" for the current year is not used, so as to avoid potential problems of simultaneity. When one looks at the lagged effects of training, the results also suggest that the shortest time periods exhibit a weak impact, confirming the hypothesis of the theory of human capital, with an increase then occurring over time. In models 1, 3, 4 and 5 (Table 4.1-1), the estimated coefficient at t-3 suggests that there is a period of adjustment between the training expenditure and the profitability of the investment. This explanation, associated with returns on training investments, holds for the various recursive models estimated. Also, the results presented in models 3 to 5 show that the results for the training variable exhibit a profitability structure that differs from those for investments in physical capital. The lag structure (t-1 to t-4) for explaining the impact of training on the productivity of Canadian firms has an upside-down U shape: it first shows a period of growth, reaches a peak and then declines, whereas for investment in physical capital it is strictly declining. The addition of control variables in no way changes this finding. The usual conception that applies to physical investments does not necessarily apply to investments in human capital.

4.3 Complementarity of factors of production

The results of estimates taking account of the interaction between capital investments and training investments are reported in Table 4.3-1. The interaction variable $\ln F_{it}$* $\ln I_{it}$ has a positive and significant coefficient for models 6 and 7, indicating complementarities between F_{it} and I_{it} for the productivity of firms. As an example, the interpretation of the overall effect of I_{it} values and the interaction between F_{it} and I_{it} in Model 6 means that a 10% increase in investments in physical capital,

when boosted by expenditures on formal training per employee, will on average lead to a 0.6% increase in the productivity of the firm the next year, compared to a firm that has not incorporated practices complementary to training, such as investments in physical capital.

Table 4.3-1
Results of estimates of the combined effect of expenditures on training and expenditures on physical capital on the productivity of firms in Canada from 1999 to 2005

Dependent variable: ln Production (value added)	**Model 6**	**Model 7**
	AR1, RE	**AR1, RE**
Independent variables	**(t-1)**	**(t-4)**
Ln (total number of employees)	0.9080***	0.8091***
	(0.0167)	(0.0289)
Proportion of employees using a computer (percentage)	0.0026***	0.0026***
	(0.0004)	(0.0006)
ln (investment in physical capital) $_{t-1}$	0.0598***	0.0691***
	(0.0090)	(0.0173)
ln (investment in physical capital) $_{t-2}$		0.0441***
		(0.0167)
ln (investment in physical capital) $_{t-3}$		0.0205
		(0.0136)
ln (investment in physical capital) $_{t-4}$		0.0244**
		(0.0120)
ln Training * ln Investments in physical capital $_{t-1}$	**0.0013****	-0.0000
	(0.0006)	(0.0008)
ln Formation * ln Investments in physical capital $_{t-2}$		0.0012
		(0.0008)
ln Formation * ln Investments in physical capital $_{t-3}$		**0.0022*****
		(0.0008)
ln Formation * ln Investments in physical capital $_{t-4}$		0.0006
		(0.0008)
Total number of observations (N)	4487	1555
Total number of firms (n)	1088	620
R square of the model	0.7738	0.7736

Notes :
1. Robust standard deviations in parentheses. *** significant at 1% ** significant at 5% * significant at 10%
2. Also included in the models are a constant, dichotomous variables for industry (14), time and control variables. Complete results are available on request.

Just as in some earlier empirical studies (Zwick, 2006), we cannot conclude, on the basis of our estimations, that the interaction between investments in physical capital and expenditures on formal training per employee is superior. This is the conclusion to be drawn from a comparison of the results obtained (such as the value of R squares) in additive model 5 with the combined effects of F_{it} and I_{it}, in model 7. On the other hand, this information is complementary. Thus, in Model 5 for example, if we sum the time-lagged effects for investments in physical capital with those for the variable $F_{it,}$ we obtain an overall effect of 0.2475 for the four years of lag. This means that a 10% increase in total investments for a given year will lead to a total increase of approximately 2.5% in the firm's productivity over a four-year period. Also, considering that the overall combined effect of F_{it} and I_{it}, in Model 7 is 0.1621, we can expect a lesser impact of the effect of interaction between F_{it} and I_{it} on firms' productivity. In other words, a 10% increase in the product of investments in physical capital and expenditures on formal training per employee, for one year, will result over a four-year period in an overall productivity gain of approximately 1.6%.

5. Conclusion

Our results definitely justify the use of longitudinal data in the study of returns on investments in training. Indeed, there are several advantages to using a longitudinal data bank, as is the case with Statistics Canada's Workplace and Employee Survey. Using a model with longitudinal data enhances the ability to carry out dynamic research. Accordingly, a study of training in firms can be better documented, especially with regard to a causal effect; but use of longitudinal data also makes it possible to identify effects that are not usually detectable with the use of cross-sectional data, including effects that are ambiguous in the short term but largely significant in the medium term.

Despite the empirical support for returns on training investments in the medium term for Canadian firms, the question arises as to why these firms do not provide more training if the returns on investment in training are substantial. One possible reason for this reluctance is the fear of seeing employees leave for better opportunities outside the firm. In complementary studies, we will attempt to shed additional light on how investments in formal training per employee affect different aspects of the turnover rate in Canadian firms.

References

Almeida, R.andt Carneiro, P. (2006). The return to the Firm Investment in Human Capital, *Discussion Paper Series*, no 1937, Institute for the Study of Labor (IZA), 24 pages.

Ballot, G., Fakhfakh, F. and Taymaz, E. (2006). How Benefits from Training and R&D, the Firm or the Workers?, *British Journal of Industrial Relations,* vol. 44, no 3, p. 473-495.

Barney, J. (1991). Firm Resources and Sustained Competitive Advantage, *Journal of Management,* vol. 17, no 1, p. 99-120.

Bassi, L., Harrison, P., Ludwig, J. and McMurrer, D. (2001). Human Capital Investments and Firm Performance, Working Paper, Human Capital Dynamics, Washington.

Becker, G. (1964). Human capital : A theoretical and empirical analysis, with special reference to education, New York, National Bureau of Economic Research.

Colombo, E. and Stanca, L. (2008). The impact of Training on Productivity : Evidence from a Large Panel of Firms, Working Papers 134, University of Milano-Bicocca, Department of Economics.

Dostie, B. and Pelletier, M.-P. (2007). Les rendements de la formation en entreprise, *Canadian Public Policy/Analyse des Politiques.* Vol. XXXIII, no 1, 21-40.

Goldenberg, M. (2006). *Investissements des employeurs dans l'apprentissage en milieu de travail au Canada* , Document de recherche préparé par les Réseaux canadiens de recherche en politiques publiques au nom du Conseil canadien sur l'apprentissage, 67 pages.

Nakamura, A. and Nakamura, M. (1998). Model specification and endogeneity, *Journal of Econometrics*, vol. 83, 213-237.

Zwick, T. (2006). The impact of training intensity on establishment productivity, *Industrial Relations*, vol. 45, no. 1, 26-46.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Workers' Mobility: A Review and Some New Results from the Workplace and Employee Survey (WES)

Yves J. Decady[1]

Abstract

In this paper, data from the employee portion of the Workplace and Employee Survey (WES) are used to first describe the incidence of job mobility. Then, the paper combines instrumental variable estimation with first differencing to produce estimates of returns to job mobility. This approach helps eliminate workers' unobserved effects on the one hand and tackle the endogeneity of the job mobility variable on the other hand. The effect of sample selection on the estimates is also examined. The research results indicate a positive return to job mobility. At the same time, this research provides evidence in support of the view that the impact of labour mobility on wage depends on the position of the worker in the wage distribution.

Key Words: First Differencing, Fixed Effects, Instrumental Variable, Sample Selection, Job Mobility.

1. Introduction

1.1 Theoretical framework

Workers' mobility and wage relationships have been studied in the field of labour economics using a wide array of theoretical perspectives. From a reading of this literature, four competing views stand out.

Various job mobility models share the predominant view that job mobility is voluntary and therefore it has a positive effect on wages. For example, job matching models predict a positive effect of job mobility on wages since workers initiate jobs separations in order to find better job matches (Jovanovic, 1979). In a similar vein, the job-shopping theory (Stigler, 1962) and the training approach in Mortensen (1988) explain job mobility. When skills are transferable across occupations, occupational mobility between jobs and employers also results in higher wage increases for movers than for stayers (Sicherman and Galor, 1990).

A second set of models contends that job mobility has a negative effect on wage (Blumen *et al.*, 1955). Several hypotheses anchored in human capital theory indicate that displaced employees may suffer wage losses with job mobility. If employers take past unemployment experience as an indication of productivity (Vishwanath, 1989), they may offer lower wages to workers re-entering the labour market after spells of unemployment (Pissarides, 1992). Labour market segmentation theorists stress that workers moving from the internal/primary labour market into the external/secondary labour market suffer wage cuts. However, mobility between primary segment jobs is expected to result in wage increases.

The third view on the wage-job mobility relationships posits that job mobility is not an important explanatory variable. To name only one school of thoughts, human capital theory, which can be traced back to the seminal work of Becker (1962), advocates that wages grow with the accumulation of specific human capital. If wages pay for worker's productivity, one can argue that job mobility is not an important explanatory variable.

Finally, some researchers rooted in job search models assume that workers use job and occupational mobility as a way to maximize lifetime income. These models suggest that voluntary mobility will generate positive wage gains (Burdett, 1978). Job mobility will lead to wages mobility, giving rise to a bi-directional relationship between job mobility and wages. Consequently, job mobility can be seen as an endogenous variable.

Considering that each view provides a plausible alternative theoretical justification for the wage-job mobility relationships, it is needless to say that the expected sign of the net effect of job mobility on wage is, all in all, ambiguous. Hence, the primary objective of this paper will be to obtain consistent estimates of returns to job mobility and to contribute to this discussion by providing additional empirical evidence.

[1]Yves J. Decady, Statistics Canada, 170 Tunney's Pasture, Ottawa, Canada, K1A 0T6 (yves.decady@statcan.gc.ca)

The remainder of the paper is organized as follows. In the next section, the data, basic definitions as well as some descriptive evidence is presented. The estimation strategy is developed in Section 3. Results are discussed in Section 4. Finally, Section 5 concludes with some remarks on a future research agenda.

2. Data, basic definitions and descriptive evidence

2.1 The WES survey data

The WES is an annual longitudinal survey sampling Canadian workplaces and employees from selected workplaces that started in 1999. The WES tracks employees for a period of two years and employers for a period of eight years. Throughout the survey cycle three two-year employee panels were gathered: the 1999 to 2000, the 2001 to 2002 and the 2003 to 2004 panel. By comparing employment status in the first year, with employment status in the second year, job mobility rate is defined as the number of employees who have changed jobs across or within workplaces between the two years of the panel divided by the total number of individuals employed in the first year of the panel.

In 2003, 20,834 employees were interviewed and in 2004; 16,804 of these employees were re-interviewed, 990 of whom were no longer in paid employment. Of those 990 employees, 113 who were self-employed were excluded. We obtain a final sample of 15,814 workers, 14,938 of whom were in paid employment in 2003 and in 2004. It appears that the wage offer was not observed for 877 individuals in 2004, either because they were unemployed or inactive. In all analysis presented in this paper, the survey weights and the bootstrap weights were used in order to properly account for the complex nature of the survey design.

2.2 Descriptive analysis

Overall, job mobility decreased between 1999 and 2004. The proportion of workers changing jobs decreased by 5 percentage points in the 2003 to 2004 employee panel, compared to the 1999 to 2000 employee panel (14.6 % versus 19.7%). Particularly noticeable was the decline in the proportion of workers changing jobs within their workplaces when the three panels were compared. From 10.9% in the 1999 to 2000 employee panel, intra-workplace mobility decreased by almost 5 percentage points in the 2003 to 2004 employee panel (10.9% versus 6.1%). Also noticeable, the outflow from paid employment was particularly higher for female workers than for their male counterparts in the third employee panel.

In the 2003 to 2004 employee panel, workers who stayed with their employers and in the same job had a smaller increase in hourly earnings (5.9%) compared to those who changed jobs while remaining with the same employer (4.0%), or changed employers (12.2%). There was indeed a wage premium associated with job mobility. This was true for both intra-workplace and inter-workplace job mobility. However, mobility was not always rewarding.

The empirical analysis will focus on the following issues. Firstly, the extent to which employees entering the panel who change workplaces in the following year has a higher wage return relative to the immobile workers or stayers. Secondly, special attention is paid to mobility wage premium differences between men and women. Finally, the effect of an eventual sample selection on the wage premium associated to job mobility is investigated. To this end, data from the 2003 to 2004 WES employee panel is used.

3. Estimation strategy

3.1 Three major methodological problems

As stated before, the main focus of this study lies in obtaining consistent measures of the effect of job mobility on wages in the 2003 to 2004 employee panel. However, three major methodological problems threaten consistency and need to be tackled: unobserved heterogeneity, the endogeneity of job mobility and the eventual sample selection bias.

Job mobility is likely to be correlated with characteristics that are unobservable such as ambition, ability and motivation, which also influence wages. Therefore, using Ordinary Least Squares (OLS) will produce biased estimates of the return to job mobility. These unobserved factors or unobserved heterogeneity, have been frequently treated via fixed effects (FE) estimations (Light and McGarry, 1998; Arulampalam 2001; Gregory and Roberts, 2001; Munasinghe and Sigman, 2004). In this paper, First Differencing (FD) is used. FD provides estimates and inference identical to FE when the number of time

periods is equal to two. By taking differences, the person-specific effects are eliminated as well as time invariant worker and job characteristics.

As for the endogeneity problem, a handful of strategies have been put forward for dealing with it, notably structural models (Flinn, 1986; Antel, 1991) and endogenous switching models. A commonly used method is instrumental variables (IV). Examples of papers using IV include Topel (1991) and Altonji and Williams (2005), to name only a few. The estimation strategy in this paper uses instrumental variable estimation (Angrist and Krueger, 2001) in order to control for the endogeneity of the of the binary job mobility variable.

The third problem emerges because wage changes are observed only for workers in paid employment in both 2003 and 2004. Self-selection bias may be an issue if the unobserved wage determinants also affect individuals' decisions to participate into the labour force. The effect of an eventual sample selection will be examined based on a new methodology introduced by Semikyna and Wooldridge (2006).

3.2 First-differenced and instrumental variable regression models

A twofold modelling strategy is adopted. The focus is put first on eliminating the unobserved heterogeneity and accounting for the endogeneity of job mobility on the estimation of the return to job mobility while ignoring sample selection. Then, the effect of an eventual sample selection bias on the estimates is assessed. Consider the unobserved effects model:

$$Y_{it} = X'_{it}\beta + (jobmob)_{it}\gamma + \delta_i + u_{it}, t = 1,2.$$

where Y_{it} is employee i's gross real hourly wage[2] at time t; X_{it} is a vector of characteristics of worker i at time t; $(jobmob)_{it}$ is a binary variable equal to 1 if a worker changes employer in the second year, and 0 otherwise; β and γ are regression parameters to be estimated; δ_i captures the unobserved effects; and u_{it} is the idiosyncratic error term. The coefficient γ is the return to job mobility. The explanatory variables include age, age squared, job tenure, position in the wage distribution prior to the workplace change, educational level, permanent versus temporary employment contract, occupation, union status, income of the rest of the household, classroom training and non-wage benefits. Consider now the following first-differenced equation:

$$\Delta Y_{it} = \Delta X'_{it}\beta + (jobmob)_{it}\gamma + \Delta u_{it}$$

where Δ indicates the difference between the second and the first year values. By taking differences, the person specific effects are eliminated as well as time-invariant worker and job characteristics. The time-varying covariates that remain in our estimated specifications are the changes between t and $t+1$.

In order to address the endogeneity problem, the first-differenced model was re-estimated using number of job changes in the past five years minus its mean, the number of job vacancies minus its means as instruments and satisfaction with the job. Regarding the use of past job changes, the intuition is that the number of job changes in the past five years approximate past investments in job search and therefore accounts for past shocks affecting current mobility status. At the same time, number of job changes should not have an effect on wages other than through current mobility status. It is assumed that workers change employers more often when there are more job openings, which positively affects the likelihood of a voluntary job change. The use of the deviation from the mean number of job changes as well as the deviation from the mean number of job vacancies causes orthogonality to take place. As for the job satisfaction instrument, given a greater level of job satisfaction, people seem to be more willing to take lower wages than if they were happier in their jobs. Perhaps one can argue that satisfied workers demand fewer wage increases to compensate them for a poor work situation than workers who are very unhappy with their jobs.

It is assumed that workers are more likely to leave the job involuntarily if a workplace's employment was decreased significantly due to layoffs, dismissals or voluntary workforce reduction. Hence, the total number of involuntary separations (number of layoffs plus number of dismissals) minus its mean is used as instrument for involuntary job changes instead of the number of job openings minus its mean. These instruments are highly correlated with the workers' likelihood to change employers, making them strong instruments. Moreover, it seems plausible to argue that they are uncorrelated to unobservable individual characteristics affecting wages.

In order to account for sample selection, a Mundlak (1978) probit selection equation was used and inverse Mill's ratios were computed. The sample selection equations reflect the fact that part of the population receives no wages either because they

[2] The nominal hourly wage is deflated with the Consumer Prices Index for each province to obtain real hourly wages using 2002 as the base year.

are unemployed or inactive.The first-differenced models were augmented with the inverse Mill's ratios and re-estimated using the same set of instruments. The tests for sample selection bias were carried using Fixed-effects-2-Stage Least Squares (FE-2SLS) on the inverse Mill's ratios, which were derived from three separate probit participation equations (all workers, male and female workers). In each case, the test did not reject the hypothesis of no selection. Therefore, the FE-2SLS approach provides consistent estimates of the returns to job mobility.

4. Estimation results

4.1 Empirical findings on job mobility

Column 1 of Table 4.1-1 provides a selected list of explanatory variables for two separate regressions. Regression 1 uses as main explanatory variable workplace movers versus stayers and Regression 2 uses voluntary movers versus stayers instead. Wage growth equations are estimated for all workers and for men and women separately. Columns 2, 4 and 5 of Table 4.1-1 contain the first-differenced regression results of the wage growth analysis for the 2003 to 2004 employee panel. Job mobility entails a positive wage return in the year after the change has taken place relative to the group of similar immobile workers. Compared to the group of immobile workers, workplace changers earn a premium of 7.0%. This return is significantly different from zero for all workers, and for both women and men. Second, the return is higher for women than for men in the FD models. Voluntary mobility is rewarded with a higher return (9.6%). As mobile female workers earn a notable wage premium, job mobility may help mitigate, over time, gender wage differences.

Columns 3, 6 and 7 of Table 4.1-1 provide results for the instrumental variable regressions. Before analysing these results, the validity of the instruments is first examined. The Anderson-Rubin Wald test, in all IV models, rejects its null hypothesis, indicating that all three instruments are highly significant (1.0%) and strongly correlated with job mobility. The Hansen *J* statistic does not reject the null hypothesis that all instruments are uncorrelated with the disturbance process in all cases. Therefore, the validity of the instruments is not called into question. The null hypothesis of the endogeneity of the job mobility variable was also rejected using the Hausman test, confirming the bi-directional nature of the wage-job mobility relationships as put forward by various job mobility theories.

Table 4.1-1
First-differenced and instrumental variable regression models results

Selected explanatory variables (1)	All workers		FD		IV	
	FD (2)	IV (3)	Men (4)	Women (5)	Men (6)	Women (7)
1) All movers relative to stayers	0.0700***	0.0847***	0.0519***	0.0841***	0.0664***	0.1059***
Position in the wage distribution						
Low-paid workers						
Medium-paid workers	0.0321***	0.0315***	0.0366***	0.0312***	0.0386***	0.0310***
High-paid workers	0.0562***	0.0550***	0.0638***	0.0541***	0.0667***	0.0544***
Hansen J Chi-sq(2) P-value Wald test Chi-sq(3) P-value		3.595 (0.1657) 26.82 (0.0000)			3.971 (0.1373) 11.80 (0.0081)	3.601 (0.1652) 19.41 (0.0002)
2) Voluntary movers relative to stayers	0.0956***	0.1135**	0.0920***	0.0993***	.0995***	0.1257***
Position in the wage distribution						
Low-paid workers						
Medium-paid workers	0.0314***	0.0311***	0.0335***	0.0316***	0.0316***	0.0315***
High-paid workers	0.0491***	0.0490***	0.0523***	0.0549***	0.0502***	0.0561***
Hansen J Chi-sq(2) P-value Wald test Chi-sq(3) P-value		2.242 0.3260 32.71 (0.0000)			3.686 0.1583 18.37 (0.0004)	2.746 0.2533 18.50 (0.0003)

Notes : Regression coefficients signifiant at : ***= 1%; **=5%; *= 10%

Turning now to the analysis of the IV estimates of the returns to job mobility, one notes that these are higher than the first-differenced estimates. As was the case with the FD estimates, women appear to have greater wage gains than men. When examining the results by workers' position in the wage distribution two things stand out: first, high-paid workers obtain a higher return than the medium-paid workers. These workers, in turn, have a higher wage return than the low-paid workers. Wage returns to job mobility increase with the position in the wage distribution. Overall, when position in the wage distribution is considered men have a higher return than women. This, however, was not always the case with voluntary mobility.

5. Conclusion

The WES data provides ample evidence in favour of a wage premium associated to job mobility. This is clearly the case with both overall mobility and voluntary job mobility, confirming prior research results. Voluntary mobility is rewarded with a higher return. Royalty (1998) as well as Gladden and Taber (2000) found that in the US voluntary changes of employer lead to wage gains; in Europe Perez and Sanz (2005) reached similar conclusions. There is also evidence of strong wage mobility without job mobility. However, the data does not provide sufficient evidence of a wage penalty for involuntary job mobility. This paper examines the wage-job mobility relationships using the survey data longitudinally and account only for employee's fixed effects. Future research will incorporate the linked nature of the survey into the analysis in order to account for both workplace and employee effects using multilevel models.

Acknowledgement

The author would like to thank Cynthia Bocci, Jean-Francois Beaumont, Harry Francois and Fritz Pierre for helpful comments they have provided during the development of this study.

References

Altonji, J.G. and Williams, N. (2005). Do Wages Rise with Job Seniority? A Reassessment, *Industrial & Labor Relations Review*, 58, pp. 370-397.

Angrist, J.A. and Krueger, A.B. (2001). Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments. *Journal of Economic Perspectives,* 15, pp. 69-85.

Antel, J.J. (1991). The Wage Effects of Voluntary Labor Mobility with and without Intervening Unemployment, *Industrial and Labor Relations Review*, 44, pp. 299-306.

Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages, *The Economic Journal*, 111, pp. F585-F606.

Becker, G. (1962). Investment in Human Capital: A Theoretical Analysis, *Journal of Political Economy*, 70, pp. S9-S49.

Blumen, I., Kogen, M. and McCarthy, P. (1955). The Industrial Mobility of Labor as a Probability Process, *Cornell Studies in Industrial and Labor Relations* 6, Ithaca: New York.

Burdett, K (1978). A Theory of Employee Job Search and Quit Rates, *American Economic Review*, 68, pp. 212-220.

Flinn, C.J. (1986). Wages and Job Mobility of Young Workers, *Journal of Political Economy*, 94, pp: S88-S110.

Gladden, T. and Taber, C. (2000). Wage Progression among Less-skilled Workers, in David E. Card and Rebecca M. Blank (eds.) *Finding Jobs, Work and Welfare Reform*, New York: The Russell Sage Foundation, pp. 160-192.

Gregory, M. and Roberts, J. (2001). Unemployment and Subsequent Earnings: Estimating Scarring among British Men 1984-94, *The Economic Journal*, 111, pp: F607-F625.

Jovanovic, B. (1979). Job Matching and the Theory of Turnover, *Journal of Political Economy*, 87, pp. 972-990.

Light, A. and McGarry, K. (1998). Job Change Patterns and the Wages of Young Men, *The Review of Economics and Statistics*, 80, pp. 276-286.

Mortensen, D. (1988). Wage Separations and Job Tenure: On-the-job Specific Training or Matching, *Journal of Labor Economics*, 6, pp. 445-470.

Munasinghe, L. and Sigman, K. (2004). A Hobo Syndrome? Mobility, Wages, and Job turnover, *Labour Economics*, vol. 11, nº. 2, pp. 191-218.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross-Section Data, *Econometrica*, 46, pp. 69-85.

Perez, J.H.G and Sanz, Y.R. (2005). Wage Changes through Job Mobility in Europe: A Multinomial Endogenous Switching Approach, *Labor Economics*, 12, pp. 531-555.

Pissarides, C. (1992). Loss of Skills during Unemployment and Persistence of Employment Shocks, *Quarterly Journal of Economics*, 107, pp. 1381-13911.

Stigler, G. (1962). Information in the Labour Market, *Journal of Political Economy,* 70, pp. 94-105.

Royalty, A. B. (1998). Job-to-job and Job-to-unemployment Turnover by Gender and Education Level, *Journal of Labor Economics*, 16, pp.441-479.

Semykina, A. and Wooldridge, J. M. (2006). Estimating Panel Data Models in Presence of Endogeneity and Selection: Theory and Application, discussion paper, Michigan State University, East Lansing, MI: Department of Economics.

Sicherman, N. and Galor, O. (1990). A Theory of Career Mobility, *Journal of Political Economy*, 98, pp. 169-192.

Topel, R. (1991). Specific Capital, Mobility and Wages: Wage Rise with Job Seniority, *Journal of Political Economy*, 99, pp. 145-176.

Vishwanath, T. (1989). Job Search, Stigma Effect and Escape Rate from Unemployment, *Journal of Labor Economics*, 7, pp. 487-502.

LONGITUDINAL DATA ANALYSIS TECHNIQUES

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

On the Use of Exploratory and Confirmatory Longitudinal Data Analysis Techniques

Marcel de Toledo Vieira, Ronaldo Rocha Bastos, Henrique Steinherz Hippert and Augusto Carvalho Souza[1]

Abstract

This paper discusses the use of various approaches for analysing longitudinal survey data, including alternative exploratory data analysis techniques and different regression modelling strategies to address longitudinal analyses of the British Household Panel Survey data on attitudes to gender roles, and their relation to demographic and economic variables. The general question in this article is: would one draw different conclusions and inferences, depending on the approach one chooses? Both exploratory and confirmatory longitudinal data analysis have been performed, by the adoption of correspondence analysis (CA) and regression modelling techniques for the analysis of adaptive relationships. Results from the CA have generally been confirmed by the regression models parameter estimates, which have often agreed in sign with the relationships displayed in the CA maps. Empirical evidence has shown that the selection of the analysis approach and modelling strategy is an important issue in the longitudinal data analysis context. We recommend that the choice should be therefore made taking into consideration the aims of the longitudinal analysis.

[1] Marcel de Toledo Vieira (marcel.vieira@ufjf.edu.br), Ronaldo Rocha Bastos (ronaldo.bastos@ufjf.edu.br) and Henrique Steinherz Hippert, Federal University of Juiz de Fora, Brazil; Augusto Carvalho Souza, Federal University of Minas Gerais, Brazil.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Goodness-of-Fit Measures for Models Based on Generalized Estimating Equations Approach

Punam Pahwa[1]

Abstract

An important part of any model selection process is the assessment of how well the model fits the data (goodness-of-fit). In the last two decades, many analytical methods have been developed for longitudinal data analysis, however, there is still a lack of standard reasonable goodness-of-fit measures for such models. For longitudinal data, we need goodness-of-fit statistics for selecting not only a correct response function but also for selecting an appropriate within-subject correlation/covariance structure. Goodness of fit statistics based on likelihood methods such as likelihood ratio test and Akaike's information Criteria (i) require repeated fittings of the data to a family of nested models, (ii) require complete specification of likelihood function, and (iii) can not be used to assess adequacy of models which are fitted by using generalized estimating equations (GEEs) approach. Vonesh et al developed three goodness-of-fit statistics: (i) rc – concordance coefficient to measure concordance between fitted and observed responses; (ii) $r(\omega\hat{})$ – a measure of concordance between assumed and true covariance structures; and (iii) - to test the equality between assumed and true covariance structure (indirectly by testing the equality between 'sandwich' and assumed covariance structure). These three measures are based exclusively on the model at hand. I propose to utilize these measures to assess the goodness-of-fit of models fitted utilizing GEEs approach to analyze longitudinal data collected (based on non-survey design) on respiratory health of Canadian grain elevator workers. An attempt will be made to modify these measures to assess the adequacy of models for longitudinal complex-survey data.

[1] Punam Pahwa, University of Saskatchewan, Canada (pup165@mail.usask.ca)

ADJUSTING FOR NON-RESPONSE AND ATTRITION

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Sample Loss from Cohort Studies: Patterns, Characteristics and Adjustments

Ian Plewis, Lisa Calderwood and Sosthenes Ketende[1]

Abstract

The omnipresence of non-response in longitudinal studies is addressed by assessing the accuracy of statistical models constructed to predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic curves and logit rank plots as ways of assessing accuracy. The ideas are applied to data from the first four waves of the UK Millennium Cohort Study and the results suggest that our ability to discriminate and predict non-response is not high. Conclusions are drawn in terms of methods of adjusting for non-response.

Key Words: Non-Response, ROC Curves, Gini Coefficient, Weights, Millennium Cohort Study.

1. Introduction

The overall theme of this paper is what can we learn from modelling the predictors of different kinds of non-response in longitudinal studies. It is common to model the logit (or probit) of the probability of not responding and then to use the predicted values from the model to generate non-response (i.e. inverse probability) weights. We can also use the information about the statistically important predictors from the estimated logistic regression as a description of the missingness mechanism and then incorporate this mechanism into a procedure for multiple imputation as described by, for example, Carpenter and Plewis (2010). Often, however, we lack any useful summary of the accuracy of our chosen model and hence we are uncertain about the value of our knowledge about missingness and its implications both for statistical adjustment and for the utility of possible interventions to prevent non-response.

The ideas in this paper are illustrated by data from the Millennium Cohort Study (MCS), the fourth in the series of internationally renowned birth cohort studies in the UK. A brief description of the MCS is given in Section 2 along with the patterns of non-response over its first four waves. Section 3 addresses the issue of how the accuracy of predictive models for non-response might be measured and introduces two summary measures: the Gini coefficient derived from the area under the receiver operating characteristic curve (ROC), and the slope of the logit rank plot as introduced by Copas (1999). Section 4 explores alternative specifications for models that predict non-response and Section 5 concludes with a discussion of the implications of the findings for statistical adjustment.

2. The Millennium Cohort Study

The wave one sample of MCS includes 18,818 babies in 18,552 families born in the UK over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. As practically all mothers of new-born babies in the UK are eligible to receive Child Benefit, the Child Benefit register was used as the sampling frame and the initial response rate was 72%. Areas with high proportions of Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered as described in Plewis (2007a). The first four waves took place when the cohort members were (approximately) nine months, 3, 5 and 7 years old. Partners were interviewed whenever possible and data were also collected from the cohort members themselves and from their older siblings.

2.1 Sample loss from the Millennium Cohort Study

Table 2.1-1 shows how the MCS sample has diminished over time after wave one. The sample loss consists of wave non-respondents – cases that are missing at wave t but not at one or more subsequent waves – and attrition cases that, once

[1] Ian Plewis, Social Statistics, University of Manchester, UK, M13 9PL (ian.plewis@manchester.ac.uk); Lisa Calderwood, Centre for Longitudinal Studies, Institute of Education, London, UK, WC1H 0AL; Sosthenes Ketende, Centre for Longitudinal Studies, Institute of Education, London, UK, WC1H 0AL.

missing, do not reappear. It is not possible definitively to allocate cases to one of these two non-response categories until the end of the study but Table 2.1-1 does indicate that sample loss from the MCS, in common with most longitudinal studies in the social sciences, consists of a mixture of wave non-response and attrition and is therefore non-monotonic. Table 2.1-1 also shows that non-respondents divide roughly 50:50 into refusals and other non-productives (not located, not contacted etc.) at wave two but that refusals become more dominant thereafter. Note that the eligible sample size – which excludes child deaths and emigrants - increases between waves two and three; some cases omitted at wave one in England were recruited for the first time at wave two.

Table 2.1-1
Sample loss from MCS by non-response type

	Wave 2, age 3 yrs	Wave 3, age 5 yrs	Wave 4, age 7 yrs
Wave non-response	8.3%	3.3%	n.a.
Attrition	9.9%	16%	n.a.
Total	18%	20%	26%
Refusal	9.1%	12%	19%
Other non-productive	9.2%	7.3%	7.4%
Eligible sample size	18,385	18,944	18,756

n.a. not applicable as wave non-response is undefined at the most recent wave.

2.2 Predictors of MCS non-response at wave two

Research reported in Plewis (2007b) and Plewis et al. (2008) on the predictors of different types of non-response in the MCS is summarised in Table 2.2-1.

Table 2.2-1
Predictors of non-response, MCS wave two

Wave one predictor	Wave non-response	Attrition	Refusal	Other non-productive
Moved residence after wave one	✓	×	×	✓
UK country	✓	✓	✓	✓
Family income	×	✓	✓	×
Refused to answer income qn.	×	×	✓	×
Ethnic group	✓	✓	×	✓
Tenure	✓	✓	×	✓
Accommodation type	✓	✓	✓	✓
Mother's age	✓	✓	✓	✓
Education	✓	✓	✓	✓
Provided stable address	×	✓	✓	✓
Cohort member breast fed	✓	✓	✓	✓
Longstanding illness	✓	✓	✓	✓
Partner present	✓	✓	✓	✓
Partner but no interview	✓	✓	✓	✓

We see from Table 2.2-1, based on a multinomial logistic regression, that variables measured at wave one that predict attrition do not necessarily predict wave non-response (and vice-versa). The same is true for predictors of refusal and other non-productives. Note that some of the predictors – refused to answer the income question, providing a stable address and no partner interview – are not variables of substantive interest and arguably fall under the heading of 'paradata' (Couper and Lyberg, 2005). We comment briefly on the value of variables like these in the process of adjusting for missing data in our concluding section.

3. Summarizing the accuracy of predictions

It is clear from Table 2.2-1 that the different types of non-response at wave two are systematically related to variables measured at wave one. What is less clear is how well the models discriminate between, or predict, respondents and non-respondents. We can think of the functions estimated from the logistic regressions as risk scores and we can then ask about the accuracy of these risk scores. A widely used method of assessing the goodness-of-fit of models for binary or categorical outcomes is to use one of several possible pseudo-R^2 statistics. Apart from their rather arbitrary nature, which thus makes comparisons across datasets difficult, pseudo-R^2 are not useful in this context because they assess the overall fit of the model and do not distinguish between the accuracy of the model for the respondents and non-respondents separately.

There are two related components of accuracy: classification (or discrimination) and prediction (Pepe, 2003). Classification refers to the conditional probabilities of having a risk score (s) above a chosen threshold (t) given that a person either is or is not a non-respondent. Prediction, on the other hand, refers to the conditional probabilities of being a non-respondent given a risk score above or below the threshold.

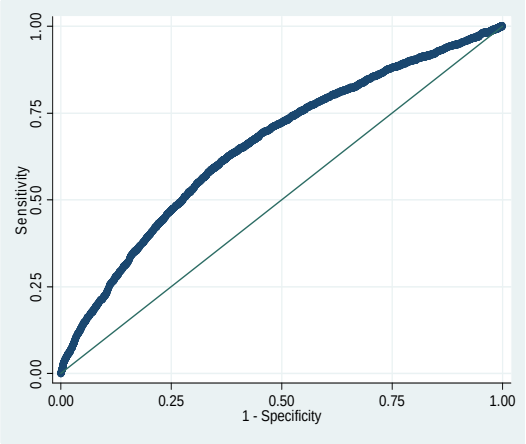
More formally, let D and $\bar{D}$ refer to the presence and absence of the poor outcome and + ($s > t$) and – ($s \leq t$) refer to positive and negative tests derived from the risk score. Then, for classification, we are interested in P(+|D), the true positive fraction (TPF) or sensitivity of the test, and P(-|$\bar{D}$), its specificity, equal to one minus the false positive fraction (1 – FPF). For prediction, however, we are interested in P(D|+), the positive predictive value (PPV) and P($\bar{D}$|-), the negative predictive value (NPV). If the probability of a positive test (P(+) = τ) is the same as the prevalence of the poor outcome (P(D) = ρ) then inferences about classification and prediction are essentially the same. Then sensitivity equals PPV and specificity equals NPV. Generally, however, (TPF, FPF, ρ) and (PPV, NPV, τ) convey different pieces of information.

The extent to which risk scores discriminate between respondents and non-respondents can be used as an indication of how influential, and possibly how effective our statistical adjustments are going to be. A lack of discrimination suggests either that there are important predictors missing from the risk score or that a substantial part of the missingness mechanism is essentially random. The extent to which risk scores predict whether a case will be a non-respondent in the next or subsequent waves is an indication of whether any intervention to reduce non-response, however well-designed and targeted, will be successful.

3.1 Receiver Operating Characteristic curve

We can plot the true positive fraction (i.e. sensitivity) against the false positive fraction (i.e. 1 - specificity) for any threshold t. This is known as a Receiver Operating Characteristic (ROC) curve (Figure 3.1-1). The ROC curve is always anchored at coordinates (0,0) and (1,1) and for large samples and at least some continuously measured predictors it is smooth with a monotonically declining but always positive slope. Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The diagonal line joining the point [0, 0] (sensitivity = 0, specificity = 1: everyone is predicted not to have a poor outcome so the threshold on the probability scale is one) to [1, 1] (sensitivity = 1, specificity = 0: everyone is predicted to have a poor outcome so the threshold is zero) is the ROC that would be obtained if the chosen variables used to construct the risk score do not explain any of the variation in the outcome. Consequently, it is the AUC – the area enclosed by the ROC curve and the diagonal – that is of interest and this can vary from 1 down to 0.5. The Gini coefficient (= 2*AUC – 1) is, however, used here as it is a more natural measure which varies from 0 to 1.

Figure 3.1-1
ROC curve



3.2 Logit rank plots

Copas (1999) proposes the logit rank plot as an alternative to the ROC as a means of assessing the predictiveness of a risk score. If the risk score is derived from a logistic regression then a logit rank plot is a plot of the linear predictor from the logistic regression model against the logistic transform of the proportional rank of the risk scores. More generally, it is a plot of logit(p_i) against the logits of the proportional ranks (r/n) where p_i is the estimated probability of a poor outcome for case i and r is the rank position of case i ($i = 1..n$) on the risk score. The slope of this relation – which can vary from zero to one - is a measure of the predictive power of the risk score and Copas argues that it is more sensitive to changes in the specification of the model underpinning the risk score than the Gini coefficient is. The slope is scale-independent and can therefore be used to compare risk scores for the outcome of interest. A good estimate of the slope can be obtained by calculating quantiles of the variables on the y and x axes and then fitting a simple regression model.

3.3 Results

Table 3.3-1 gives the Gini coefficients and slopes of the logit rank plots for both overall non-response at wave two and for the different types of non-response. The estimate of 0.39 for the Gini coefficient for overall non-response is relatively low, indicating that discrimination between non-respondents and respondents from the risk score is not good but it is slightly better for wave non-respondents and other non-productive than it is for attrition and refusal. A similar picture emerges from the slopes of the logit rank plots although these bring out more clearly the differences in predictiveness for the different types of non-response.

Table 3.3-1
Accuracy measures, MCS wave 2

	ROC: Gini	Logit rank plot: slope
Overall non-response	0.39	0.45
Wave non-response	0.43	0.52
Attrition	0.39	0.41
Refusal	0.37	0.37
Other non-productive	0.52	0.58

95% confidence limits generally ± 0.02

We find that a 'consent' variable – the respondent expresses willingness at wave one for the survey records to be linked to health data from administrative sources – is also associated with non-response at wave two although this variable was not used in the earlier work. Its inclusion in the model, however, has very little effect on the Gini coefficient although the slopes of the logit rank plots are higher. This provides some reassurance that risk scores might be relatively robust to some misspecification of the logistic regressions.

It is possible to use variables measured at wave t+1 to predict wave non-response at wave t. We find that change in accommodation type and in partnership status between waves t-1 and t+1, and family income at wave t+1 all predict wave

non-response with the Gini coefficient rising from 0.43 to 0.46. This has implications for adjustment methods as discussed below.

4. Alternative strategies for predicting non-response

One of the difficulties faced by analysts wishing to use weights to adjust for non-response is that, ideally, the weights need to be re-estimated at each wave. The search for new predictors of non-response then recalculating and depositing the weights for secondary analysts can, however, be time-consuming and possibly wasteful of scarce technical resources. Consequently, it is worth considering whether the changes in the weights after wave two are sufficiently large to justify recalculation at each wave. Here we focus on wave four of MCS and compare Gini coefficients for three models for overall non-response based on:

1. Wave one variables, wave one values of these variables and wave one coefficients.
2. Wave one variables, wave one values, wave three coefficients.
3. Wave one values, wave three values and wave three coefficients.

We find that only four of the 15 wave one variables that were predictive of wave two non-response are not predictive of wave four non-response and the estimated Gini coefficients for the three models are 0.36 (n = 17862), 0.37 (n = 17862) and 0.36 (n = 12729), only a little smaller than the estimate of 0.39 at wave two. Note also that item non-response leads to a substantial decrease in sample size for the third strategy and this is a general difficulty when constructing non-response weights.

Further work will look at the discrimination and prediction at wave four based on the inclusion in the model of other wave three variables.

5. Conclusions

Three main points emerge from this paper. The first is that using a framework that is constructed around different kinds of conditional probabilities and risk scores generates summary measures of accuracy like Gini coefficients and slopes of logit rank plots that enable us to make comparisons across models that predict non-response. These comparisons provide a means of assessing the usefulness of introducing extra predictors into the models and of comparing predictiveness and discrimination for different kinds of non-response. We find, at least for the Millennium Cohort Study considered here, that our best models for missingness are not especially accurate and this lack of accuracy suggests that we might find it difficult effectively to target interventions that might prevent non-response. The issue of targeting brings in questions about the optimum cut point on the risk score that, in turn, requires a consideration of the costs and benefits of intervention that goes beyond the scope of this paper.

The second point is that models developed to generate non-response weights at wave two might be satisfactory to use at later waves. If this point were supported by further investigations, in particular of models that incorporate variables measured for the first time after wave one – for example, children's performance on educational tests taken at home – then this would suggest that efforts to estimate fresh non-response models at each wave might be misplaced and it would also get round some of the problems created by item non-response in the models used to generate the inverse probability weights.

The final point relates to the implications of our results for statistical adjustment other than by using inverse probability weights. We find that our models of missingness are, for wave non-respondents, improved by including variables from a later wave and therefore these variables could be included in a selected imputation process for a particular model of interest. Moreover, the fact that some of the important predictors of non-response are variables that are unlikely ever to feature in a model of interest – providing a stable address, for example - means that they might be used as instruments in a joint Heckman-type model that considers the models of interest and missingness simultaneously and thus allows for non-ignorable missingness. Carpenter and Plewis (2010) provide an example.

References

Carpenter, J. and Plewis, I. (2010). Analysing Longitudinal Studies with Non-Response: Issues and Statistical Methods, in M. Williams and P. Vogt (eds.), *Handbook of Methodological Innovations*. Newbury Park, Ca.: Sage.

Copas, J. (1999). The Effectiveness of Risk Scores: The Logit Rank Plo", *Applied Statistics*, 48, pp. 165-183.

Couper, M. P. and Lyberg, L. E. (2005). The Use of Paradata in Survey Research, paper presented at the 54th Session of the International Statistical Institute, Sydney, Australia.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Boca Raton, Fl.: Chapman and Hall/CRC.

1.1 Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford: OUP.

Plewis, I. (ed.) (2007a). *The Millennium Cohort Study: Technical Report on Sampling* (4th. ed.), London: Institute of Education, University of London.

Plewis, I. (2007b). Non-response in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology*, 10, pp. 325-334.

Plewis, I., Ketende, S. C., Joshi, H. and Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the Millennium Cohort Study, *Journal of Official Statistics*, 24, pp. 365-385.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Analysis of attrition in the Longitudinal Study of Child Development in Quebec (ÉLDEQ) from 1998 to 2008

Catherine Fontaine and Robert Courtemanche[1]

Abstract

The Quebec Longitudinal Study of Child Development (QLSCD) began in 1998. After 10 years of data collection, there are 1,402 respondents compared to 2,120 in the first cycle. In light of a possible extension of the QLSCD starting in 2011, it is necessary to assess the limitations imposed by sample attrition, including the longitudinal bias of the estimates. The analysis examines the adjustment of non-response through weightings. Finally, some solutions are put forward to alleviate the effects of attrition.

Key Words: Attrition, Nonresponse, Bias.

1. Introduction

1.1 General description of the survey and its objectives

The Quebec Longitudinal Study of Child Development began in 1998. It is intended to be representative of children born in Quebec in 1997-1998 (single births only). A sample of 2,817 children was selected from the master file of live births in Quebec, using a three-stage stratified sample design. No additions have been made to the sample since 1998. The main objective of the survey is to identify the factors which, having come into play in early childhood, contribute to the social adaptation and academic success of young Quebecers. Collection was carried out annually from 1998 to 2006, then biennially since 2006. Various collection instruments were used in each collection, including the computerized questionnaire completed by the interviewer, which is administered to the person most knowledgeable about the child (PMK). Measurements are also performed on the child, through cognitive tests, an assessment of physical condition or the paper questionnaire administered to the child. A "cross-sectional" weighting is calculated for each cycle to represent all children born in Quebec in 1997-1998 who completed at least one collection instrument in that cycle.[2] Other weightings are added in different cycles, including a longitudinal weighting as well as weightings for secondary collection instruments. Up to the 2006 cycle, weighting classes were created by means of segmentation (using the CHAID algorithm). When the classes were created, the weighted proportion of respondents was calculated using a reference cycle (prior to the cycle to be weighted). The 1998, 2000 and 2002 cycles were used as reference cycles for the different weightings.

1.2 Description of study

The attrition study is part of a re-evaluation process for the continuation of the survey. Initially, the survey was planned for a single phase (1998 to 2002). Then, a second phase was added, from 2003 to 2010, covering the child's elementary education years (kindergarten through grade 6). Now, the intention is to continue the survey for a third phase, from 2011 to 2015, thus covering the child's secondary school years (secondary 1 to 5). However, since the sample underwent attrition between 1998 and 2008, there is reason to evaluate the limitations that this attrition imposes on analyses performed, and to consider solutions that could be implemented in the third phase.

2. Attrition in the QLSCD

Attrition is generally defined as the premature loss of units selected for longitudinal follow-up, in relation to the subset of units that responded in time 1. In the QLSCD, the attrition that can be described as "definitive" includes children whose

[1]Catherine Fontaine et Robert Courtemanche, Institut de la statistique du Québec, 200, chemin Sainte-Foy, Québec (Québec), G1R 5T4, Canada.
(catherine.fontaine@stat.gouv.qc.ca et robert.courtemanche@stat.gouv.qc.ca).
[2] Excluded are children who died and those who left Quebec definitively.

family has left Quebec definitively, deaths and parents' persistent refusals to participate in the survey. However, it may not be possible to identify some cases of definitive attrition before the end of the survey. In particular, this applies in the case of families who do not respond in a survey cycle but are re-contacted in each cycle and respond in a later cycle. These cases may be considered "temporary" attrition. And indeed, this is the approach favoured by the QLSCD: most non-respondents in an earlier cycle are contacted in a later cycle.

The following table shows attrition starting with the sample of respondents in 1998.

Table 2.1-1
Attrition in the QLSCD from 1998 to 2008

Cycle	Children's age	Number of respondents	Attrition (%)
1998	5 months	2,120	0.0
1999	17 months	2,045	3.5
2000	2.5 years	1,997	5.8
2001	3.5 years	1,950	8.0
2002	4 years	1,944	8.3
2003	5 years	1,759	17.0
2004	6 years (kindergarten)	1,492	29.6
2005	7 years	1,528	27.9
2006	8 years	1,526	28.0
2008	10 years	1,402	33.9

Note: This table shows the number of respondents covered by cross-sectional weighting.

The attrition analysed is defined as non-response to a survey cycle, for cycles 1999 to 2008, in relation to the children who participated in the survey in 1998. This definition therefore includes both "definitive" and "temporary" cases of attrition. It also corresponds to the concept of non-response for which processing is carried out in each cycle by means of weightings. This processing increases the weights of survey respondents in a given cycle to compensate for the loss of non-respondents, compared to a previous cycle. It should be noted that the weights represent only those children eligible for the survey in a given cycle (cases where children have definitively moved outside Quebec or died are excluded).

The estimate of longitudinal bias may be defined as the difference between a statistic measured for a variable of interest in the population and the weighted estimate of that statistic in the sample at time *t*. However, this definition implies that the statistic can be calculated in the population. At the time of this analysis, such a measure was unavailable. Therefore, the concept of longitudinal coherence was needed to replace the concept of longitudinal bias.[3] Longitudinal coherence can be estimated by calculating the difference between the weighted estimate of a characteristic at time 1 (when the sample size is large) and the same estimate at a later time (time *t*).

3. Methods of analysis

3.1 Choice of characteristics to be studied

Three methods were used to perform this attrition analysis. The first two methods focus on the longitudinal coherence aspect, while the third concerns the precision aspect. They will be described in detail in the sections below. For these methods, the same subset of nine variables, measured in the 1998 cycle, was used. These variables were chosen primarily because they are related to measures of academic success. The nine variables are listed below:

1) household income in 1998 (under $15 000);
2) baby's health in 1998 as perceived by PMK (good, poor or fair);
3) adequacy of household income in 1998 (insufficient or very insufficient);
4) language spoken at home by the mother in 1998 (neither French nor English);
5) family type (lone parent);
6) mother's immigrant status (European or non-European immigrant);
7) mother's highest level of schooling in 1998 (high school diploma or less);

[3] For a definition of the concept of longitudinal coherence, see Franklin *et al.* (2007).

8) health region of resident in 1998 (Montréal or Laval);
9) mother's age in 1998 (under 25).

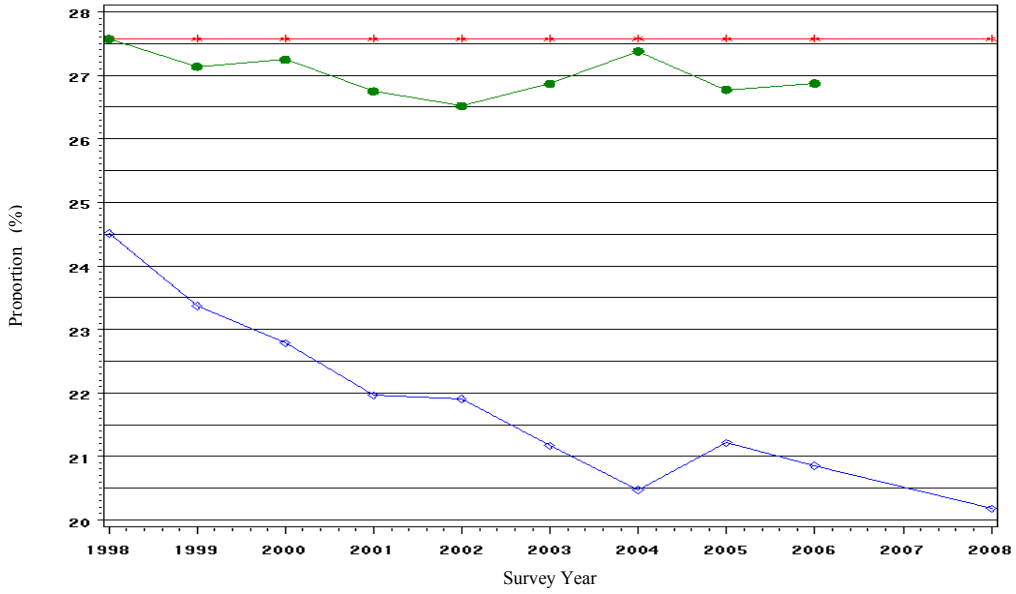
3.2 Analysis of longitudinal coherence

3.2.1 Charts

To provide a visual illustration of longitudinal coherence, charts were produced for each of the nine characteristics studied. Essentially, these charts show how the estimate of a given characteristic evolved from 1998 to 2006. Two curves are shown, based on the different sampling weights used in the estimate (with or without adjustment for non-response). These curves can be compared to a reference line, which is the estimate of the characteristic in time 1. The charts showed that estimates with an adjustment for non-response reduce the effect of losses in the sample. By contrast, unadjusted estimates generally take the form of a curve moving downward over the years. An analysis of the charts also shows that two characteristics seem more problematic as to the level of longitudinal coherence (see Section 4 for further details).

A single chart is provided as an example, namely the one illustrating adequacy of household income. The curve of the estimates adjusted for non-response[4] does not follow the same trend as the curve for estimates unadjusted for non-response.[5] It remains fairly close to the reference estimate[6] (1998 cycle), whereas the curve for unadjusted estimates becomes increasingly divergent. This indicates that the weighting has done its work.

Chart 3.2.1-1
Chart illustrating longitudinal coherence



3.2.2 Analysis of relative differences

Although they are useful, the charts do not shed light on the scope of longitudinal coherence, because they differ in scale. For this reason, a calculation was performed measuring the relative difference between each of the weighted estimates for a given cycle and the corresponding weighted estimate for the 1998 cycle. The calculation was performed for cycles 1999 to 2006 (data from cycle 2008 had not been weighted at the time of the analysis). The weighted estimates include the adjustment for non-response.

[4] Identified by the "● " sign.
[5] Identified by the " ◊ " sign.
[6] Identified by the "+" sign.

Formula 1: Measurement of relative difference

$$Relative\ difference_t = \frac{\left(\hat{p}_{cross-sectional\ weighted,\ cyclet\ t} - \hat{p}_{cross-sectional\ weighted,1998\ cycle}\right)}{\hat{p}_{cross-sectional\ weighted,1998\ cycle}}$$

for times t =1999 to 2006.

Table 3.2.2-1 shows, for each of the nine characteristics, the maximum, in absolute value, among the relative differences calculated for years 1999 to 2006.

Table 3.2.2-1
Maximum, in absolute value, of the relative difference per characteristic

1998 characteristic	Maximum (%)
Language spoken at home by mother	18.3
Mother's immigrant status	14.8
Family type	6.7
Health region	6.2
Household income	5.9
Mother's highest level of schooling	5.8
Adequacy of income	3.8
Mother's age	3.1
Baby's health	2.5

3.3 Analysis of precision

The third aspect of the attrition analysis concerns precision and power. It will be dealt with simply by presenting the number of respondents who exhibited a given characteristic in 1998, and who were still respondents in 2006 or 2008. The following table shows the losses (unweighted) for the characteristics studied.

Table 3.3-1
Change in the number of respondents between 1998 and 2008

1998 characteristic	1998 cycle	2006 cycle	2008 cycle	Loss between 1998 and 2008 (unweighted)
Language spoken at home by mother	128	55	53	59 %
Mother's immigrant status	253	128	120	53 %
Household income	248	140	121	51 %
Family type	171	100	86	50 %
Adequacy of income	511	316	284	44 %
Mother's highest level of schooling	564	370	327	42 %
Health region	571	371	338	41 %
Mother's age	159	109	96	40 %
Baby's health	473	330	292	38 %
Number of respondents in cycle	2,120	1,526	1,402	34 %

Note: This table shows the number of respondents covered by cross-sectional weighting.

4. Findings and proposed solutions

The two variables with the poorest level of longitudinal coherence, among the variables studied, are usually characteristics that serve to define the non-response model. These variables are the mother's immigrant status and the language spoken by the mother at home. In other words, the probability of responding varies according to these characteristics, and the adjustment that is done serves to reduce the non-response bias for these characteristics (and those related to them). However, it appears that even these adjustments do not serve to re-establish a good level of longitudinal coherence. Table 3.3-1 shows that these two characteristics have undergone sizable losses since 1998, proportionally speaking, in relation to the other variables. It is therefore more difficult to minimize the biases in this case. Also, the method of segmenting the

sample of respondents (which creates weighting groups) may have used these two variables to define the non-response model for only a subgroup of children.

The two variables that show the lowest losses of respondents (approximately 40%) also exhibited a very good level of longitudinal coherence. These are the variables representing children whose mother was under 25 years of age and children whose health as a baby was perceived to be good, fair or poor in 1998. One of these variables (mother's age) is considered in the modeling of non-response, while the other is not.

Other findings are also contained in the study of Fontaine and Courtemanche (2009). Thus, all these analyses reveal that the 1998 characteristics that are most affected by attrition in the QLSCD are the mother's immigrant status, household income, the language spoken by the mother at home and family type. If they are analysed and also examined in relation to academic success, there is an increased risk of longitudinal bias, despite the fact that these are characteristics that are usually taken into account in when adjusting for non-response for the different instruments/cycles. Lastly, they are generally associated with small numbers of respondents in 2008, which can impose limitations on analysis. What, then, are the solutions needed to correct this situation in the future?

The solution put forward to improve the longitudinal coherency aspect consists in methodological studies to examine a method of adjusting for non-response as an alternative to the method currently used. Both these methods function by correcting the reference weights for a previous cycle using the inverse of the weighted response rate for the current cycle, within a subset of children: these are the weighting groups. The current method uses a segmentation of the data set, based on tests of association (chi-square). The alternative method considered, called the score method, creates groups according to the values of a score, which is the estimated probability of response resulting from a logistic regression model. Recent studies suggest the benefits of the score method (Haziza and Beaumont, 2007). The result of these methodological studies indicates that the four characteristics that had been targeted as being the most associated with attrition exhibit an improvement in their level of longitudinal coherence when the score method is used. The latter method was therefore implemented in weighting the data from the 2008 cycle.

As to the aspect concerned with the precision of the estimates, the solutions proposed seek to limit the decrease in the number of respondents, which has been especially steep in the more recent cycles. Developing respondent loyalty, reducing the response burden (by limiting the number of collection instruments) and using administrative data (for the sample as a whole) are all measures likely to improve the power of the analyses performed.

5. Conclusion

In conclusion, the study on attrition in the QLSCD up to 2008 confirmed that the sub-set of children affected by this loss come from families with one (or more) of the following characteristics: low household income; lone parent situation; immigrant mothers; mothers who speak neither French nor English at home. This subset is a domain of interest when studying academic success.

On the other hand, this analysis is compromised by the risk of potential bias and the number of children available for analysis purposes. Therefore, a change in the method of creating weighting classes was implemented when the data for 2008 were weighted, in an effort to reduce the effects of attrition on longitudinal bias. It will also be possible to target this sub-group of children during data collection to minimize future non-response. This has already been done in previous collections.

This study of attrition thus showed the important role that weighting plays in minimizing biases due to attrition, since weighting improves the level of longitudinal coherence of characteristics related to academic success. This finding has been confirmed by a supplementary analysis based on population data, available from the ministère de l'Éducation, des Sports et des Loisirs du Québec.

In closing, users of QLSCD data have been informed of the findings of the attrition study. This will enable them to assess the extent to which those findings may limit the conclusions that they can draw from their future analyses.

References

Fontaine, C. and R. Courtemanche, (2009). Étude de l'érosion pour l'Étude longitudinale sur le développement des enfants du Québec de 1998 à 2008, rapport non publié, Québec, Canada : Institut de la statistique du Québec.

Franklin, S. et al. (2007). The National Longitudinal Survey of Children: 10-Year Review of the Methodology, rapport non publié, Ottawa, Canada: Statistique Canada.

Haziza, D. and J.-F. Beaumont, (2007). On the construction of imputation classes in surveys, *International Statistical Review*, 75, p. 25-43.

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Using Paradata in the Modelling of Nonresponse for a Longitudinal Survey: Application to the Survey of Labour and Income Dynamics

Beatrice Baribeau, Wisner Jocelyn, and Junior Chuang[1]

Abstract

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey that collects data about work, income and family circumstances to produce longitudinal and cross-sectional estimates. The survey currently uses frame information, from the Labour Force Survey (LFS), to produce a nonresponse adjustment using the Chi-Square Automatic Interaction Detection (CHAID) algorithm. In this paper we investigate nonreponse adjustments that utilise paradata in comparison with the current nonresponse adjustment. The results of this study show that while the current nonresponse adjustment yields some lower standard errors associated with some key estimates, the model incorporating both paradata and frame information provides advantages in of higher correlation with response and non-income variables as well as a reduction in pseudo-relative bias.

Key Words: Paradata, Nonresponse Adjustment, Pseudo-Relative Bias.

1. Introduction

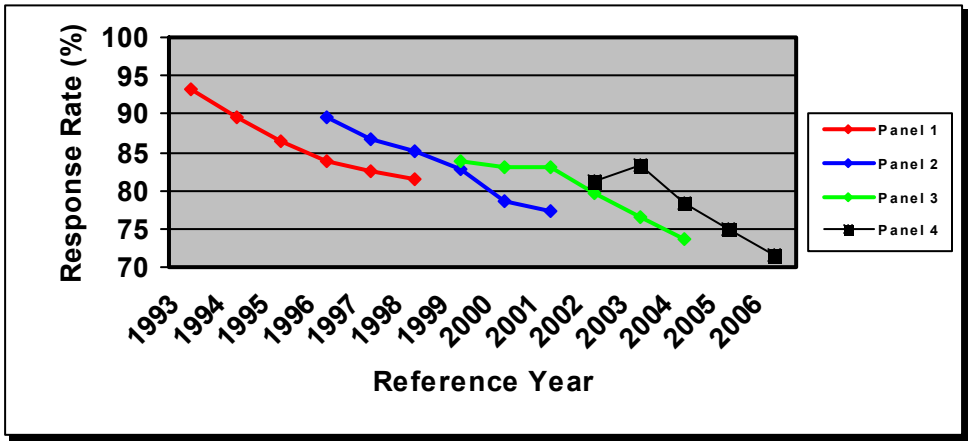
1.1 Introduction to the Survey of Labour and Income Dynamics

The Survey of Labour and Income Dynamics (SLID) is an annual, longitudinal, telephone survey that collects data about work, income and family circumstances to produce longitudinal and cross-sectional estimates. It longitudinally follows a sub-sample of individuals from approximately 17,000 responding households from the Labour Force Survey (LFS). Each panel in SLID is followed for six years, with panels being introduced every three years. Since the sample is taken from responding LFS households, basic demographic information on persons in SLID is readily available. Households with a hard refusal or two consecutive years of nonresponse are removed from collection. This aspect of collection affects the use of paradata in nonresponse modelling given that paradata is only available on those households present in collection.

Figure 1.1.-1 shows the longitudinal response rates by wave and panel. The year listed on the bottom of the figure represents the year of reference for the survey. The year of collection is one greater than the year of reference. This figure clearly demonstrates a decline in response rates over panel life as well as from one panel to the next. This decline is a motivating factor in determining an optimal nonresponse adjustment. Other factors to consider in the analysis of nonresponse adjustments is the introduction of responsive design for the reference year 2009 and a redesign of the survey with an indefinite panel life. Responsive design uses the paradata to focus collection efforts in order to increase response rates and ensure sample representativity. Using a nonresponse adjustment with the same variables nicely compliments this method. The redesign which could include a panel of indefinite length means that as the panel ages, the frame information will become more and more outdated. The panel used in this research was panel four, the last panel demonstrated. It was chosen due to its multiple waves available for analysis and its relevance, being a recent panel.

[1] Beatrice Baribeau, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Beatrice.Baribeau@statcan.gc.ca; Wisner Jocelyn, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Wisner.Jocelyn@statcan.gc.ca; Junior Chuang, Simon Fraser University, Canada

Figure 1.1-1
Longitudinal response rate by wave and panel



1.2 Current nonresponse methodology

The survey currently uses frame information, from the Labour Force Survey (LFS), to produce dichotomous variables to model nonresponse. The nonresponse adjustment is based on the Chi-Square Automatic Interaction Detection (CHAID) algorithm which retains the most significant variables in explaining response. It builds a tree based on these variables where the leaves of the tree are the response homogenous groups (RHGs). Restrictions are placed on the RHGs to contain a minimum of 30 units with a minimum weighted response rate of 50%. The inverse of the weighted response rate is used to adjust the weights of respondents, while nonrespondents are assigned a weight of zero. A benefit of using the CHAID algorithm is that in building the tree, the algorithm allows dissymmetry and thus is less restrictive than clustering methods, see Beaument (2005) or propensity methods as in Watson and Wooden (2006).

1.3 Available paradata

SLID is a computer-assisted telephone survey. The available paradata (see Scheuren (2005) for an insightful discussion on the concepts of paradata) comes from the Blaise Transaction History (BTH) files. The BTH files contain information such as a record of each call to a household, the time and day of the call and the outcome of the call (i.e., busy signal, interview obtained, etc.). The information is at the household level, that is, the calls represent calls to the household, not particular individuals within the household. Households may have multiple records, each one representing an individual phone call. Since calls can vary in characteristics such as the day of the call, only the most recent call was kept per household. The last call is the most probable to correspond to the outcome of a response or a nonresponse. Similarly to the frame information, the paradata information can be converted to dichotomous variables such as “number of call to the household” to be used by the CHAID algorithm.

2. Study assumptions

2.1 Panel Four data

Five waves of panel four were used for this study. Approximately 1 860 (from 34 000) units were excluded from the study. The bulk of exclusions were persons who were children in wave one that became adults by wave five (1 700 units) and nonrespondents for whom no BTH information as a nonrespondent existed (1 150 units). The two groups contain considerable overlap. The first group was excluded due to a lack of variables of interest in wave one, which would have contributed an unpredictable component to the pseudo-relative bias.

2.2 Longitudinal imputation of the paradata

In cases where past nonresponse led to removal from the collection sample, no paradata existed. In these circumstances, a longitudinal imputation from the most recent wave of paradata was performed provided the data for the unit was also a nonrespondent. The information from a past wave of nonresponse may not have reflected the situation in the more recent wave of nonresponse but it was more current than the frame information. In some cases in the last wave of available

paradata, the unit was a respondent; these records were removed from the study. This study was limited in that excluded nonresponding units could differ from the included nonresponding units.

3. Analyses of the nonresponse adjustments

3.1 Correlation between auxiliary variables and response

The formation of RHGs depends on auxiliary variables being related to response; hence it is desirable to have strong correlation between response and auxiliary variables. Figure 3.1-1 shows the absolute value of correlations between auxiliary variables taken from the LFS and response for reference years 2002-2006. The absolute value of the correlation was used to demonstrate the magnitude of the correlation, which for the purpose of comparison between auxiliary variables was of more interest than the direction.

As demonstrated in Figure 3.1-1, correlation increased over the panel life for most variables. The variables most correlated to response were renter, single and married; consequently, they were the first branches of the tree that created the RHGs.

Figure 3.1-1
Absolute value of correlation between LFS variables and response over the panel life

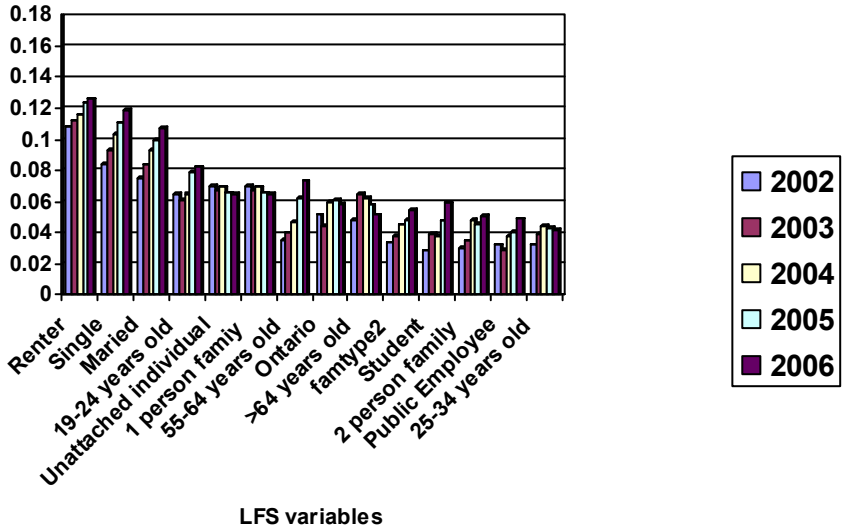
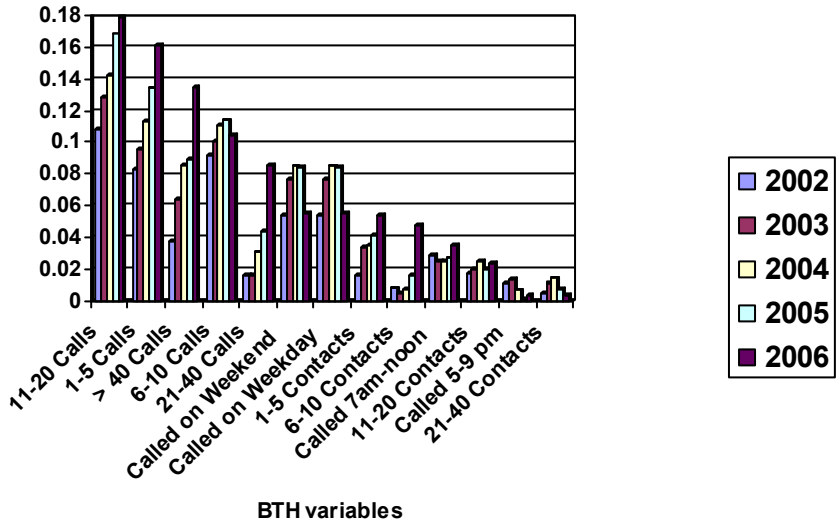


Figure 3.1-2 shows the absolute value of correlation for the auxiliary variables from the paradata. Like the LFS variables, the BTH variables became more correlated to response over time. However, three BTH variables exceeded the highest LFS correlation. The most correlated BTH variable, "11-20 calls", had an absolute correlated value of 0.18. When both LFS variables and BTH variables were included in the CHAID algorithm to perform a nonresponse adjustment, BTH variables were always selected to form the first branches of the tree.

Figure 3.1-2
Absolute value of correlation between BTH variables and response over the panel life



3.2 Correlation between key auxiliary variables and variables of interest

The second measure for the auxiliary variables was correlation between key auxiliary variables and variables of interest. It is desirable to have a strong correlation between the key auxiliary variables and the variables of interest since this leads to a lower standard error associated with estimates.

Overall neither source of auxiliary variables was strongly correlated to the variables of interest. From table 3.2-1 one can see that the frame variables were more correlated to income than the paradata variables. This was also true for individual sources of income. However, the paradata variables tended to be more correlated to non-income variables as demonstrated in table 3.2-2.

Table 3.2-1
Correlation between auxiliary variables and income

LFS variables	Correlation	BTH Variables	Correlation
Single	0.10	1-5 calls	0.05
Renter	0.08	6-10 calls	0.01
> 64 years old	0.06	11-20 calls	0.02
Ontario	0.04		

Table 3.2-2
Correlation between auxiliary variables and hours worked

LFS variables	Correlation	BTH Variables	Correlation
Renter	0.06	1-5 calls	0.14
Single	0.02	6-10 calls	0.06
Ontario	0.02	11-20 calls	0.05
> 64 years old	0.01		

3.3 The pseudo-relative bias

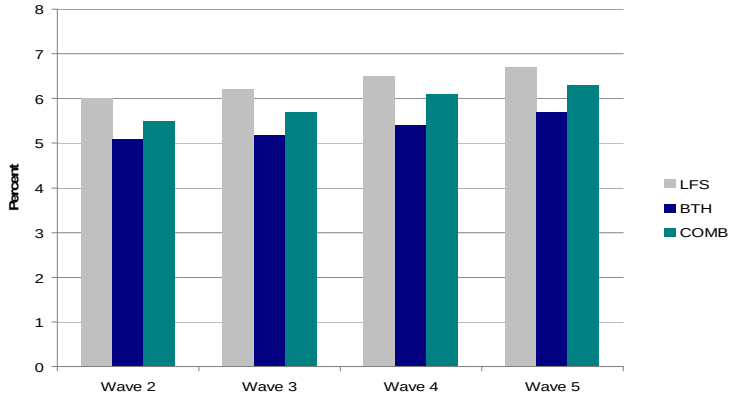
Without the true population estimates, bias cannot be measured. Following Singh et al (1995), we used a measure of longitudinal consistency dubbed the pseudo-relative bias. In it, the wave one estimate can be seen as the best guess of the true population estimate with subsequent waves experiencing attrition. With the attrition due to nonresponse, there is the potential for nonresponse bias if the nonrespondents differ from the respondents, unless it is accounted for by the nonresponse adjustment. If the nonresponse adjustment adequately adjusts for the nonrespondents then using the adjusted weights of a later wave combined with the variable of interest values from wave one should produce an estimate close to that of the wave one estimate. Measurement of the percentage change of the new estimate from the wave one estimate is the pseudo-relative bias. This is expressed as:

$$pseudo\_rel\_bias = \left(\frac{\hat{Y}_{1,S}^{(wi)} - \hat{Y}_{1,LFS}^{(w1)}}{\hat{Y}_{1,LFS}^{(w1)}} \right) * 100$$

$\hat{Y}_1^{(wi)}$ is an estimate for a particular variable of interest (Income, Salary, Rent) and a given source S (LFS, BTH, Combined LFS+BTH) using weight from wave *i (*wi*)* and *y-values* from wave one.

Using income to measure the pseudo-relative bias, Figure 3.3-1 shows the pseudo-relative bias for three nonresponse adjustments. The first adjustment is based solely on LFS variables, the second uses only BTH variables while the third uses a combination from both sources (labelled as COMB). As longitudinal attrition increases with each wave, so does the pseudo-relative bias. Figure 3.3-1 shows that the more dependent the adjustment is on the paradata, the lower the observed pseudo-relative bias. This trend is also observed in several other income and non-income variables.

Figure 3.3-1
Pseudo-relative bias of income by adjustment



3.4 The standardised interquartile range

Examining the interquartile range was done to gauge the homogeneity of the weights after the nonresponse adjustment. Homogeneity of the weights within an RHG can lead to better stability in the estimation of the standard errors. The standardised interquartile range (SIQR) is defined as (Q3-Q1)/median. Table 3.4-1 shows the averaged SIQR over all RHGs for the three adjustments. A lower SIQR is desirable. The LFS only adjustment produces the lowest SIQR. This result is as expected given the LFS variables include province and the correlation between province and the magnitude of weights.

Table 3.4-1
Means of the Standardised Interquartile range

Adjustment	2002	2006
LFS	0.596	0.929
BTH	1.201	1.283
Combined	1.083	0.736

3.5 Standard errors for total income

The analyses performed in 3.3 and 3.4 used the weights from after the nonresponse adjustment. Standard errors on the estimates for income were examined after performing an adjustment for influential values and calibration. Laroche (2007), provides a detailed discussion of different types of weights used for SLID. The higher correlation of LFS variables to variables of interest as well as the lower SIQR indicated that the variability of estimates may be more stable than for weights based on the BTH files alone. As indicated in table 3.5-1 this was indeed true. Table 3.5-1 shows the coefficient of variation (CV) and percentage change of the standard errors (S.E.) from the LFS estimates to the BTH only adjustment and the combined paradata and LFS adjustment, by age group. The only estimate that was statistically significantly different is for 55-64 year-olds in 2001. When estimates were examined by province similar results were seen. While overall, there was an increase in the standard errors of estimates, CVs remained low.

Table 3.5-1
Coefficient of variation and percentage change in standard errors for income by age

	BTH		Combined	
Age in 2001	CV	Percentage change of S.E.	CV	Percentage change of S.E.
<20	3.1%	-1.6%	3.0%	-4.4%
20-24	2.9%	-7.3%	2.9%	-7.5%
25-34	2.0%	1.8%	2.1%	2.8%
35-44	1.8%	0.7%	1.8%	0.8%
45-54	2.5%	4.9%	2.5%	3.7%
55-64	2.5%	2.3%	2.5%	-0.2%
65+	2.0%	-2.4%	2.3%	12.3%
All ages	3.1%	2.8%	3.0%	3.1%

4. Conclusions and future work

The paradata auxiliary variables were more correlated to response and some non-income variables than the frame variables. The frame variables were more correlated with income variables and produced a lower SIQR than the paradata variables. The higher correlation with income and lower SIQR indicated a greater stability in variability in estimates for income, as was confirmed in a lower standard error for income. However, given that CVs of estimates for income remained low when the nonresponse adjustment relied only on the paradata and that it reduced the pseudo-relative bias, the trade off between variability and bias was acceptable. Bearing in mind other factors such as the introduction of responsive design and a planned redesign which will incorporate an indefinite length of time per panel, an ideal nonresponse adjustment could incorporate both the frame and paradata variables.

Future work includes producing estimates and CVs for each year of panel four. Before a new nonresponse adjustment that uses paradata can be incorporated, a treatment for currently excluded nonresponding units must be proposed and tested. Estimates and correlations for more non-income variables should be produced and pseudo-relative biases after calibration should be calculated.

References

Beaumont, J-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31, 227-231.

Laroche, S. (2007). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, Catalogue no. 75F0002MIE — No. 007, Statistics Canada.

Scheuren, F. (2005). Paradata from concept to completion, *Proceedings of Statistics Canada International Symposium 2005*.

Singh, A.C., Wu, S. and Boyer, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 396-401.

Watson, N. and Wooden, M. (2006), Modelling Longitudinal Survey Response: The Experience of the HILDA Survey, HILDA PROJECT DISCUSSION PAPER SERIES No.2/06, Australia.

STATISTICS CANADA SYMPOSIUM - 25TH ANNIVERSARY BANQUET

Proceedings of Statistics Canada Symposium 2009
Longitudinal Surveys: from Design to Analysis

Statistics Canada Symposium - 25th Anniversary Banquet

Gordon Brackstone

1. Introduction

When I was first asked to say a few words on this occasion, I pictured us in a windowless nondescript hotel ballroom somewhere, not in this west-coast rain forest backed by totem poles and looking out on the majestic Ottawa River. I hope I can do justice to the occasion and to the surroundings.

I have been asked to talk about the Symposium as someone who was there at the start, and for 20 more after that. I want to emphasize that I was there at the start but I was not the starter. David Binder's short note in the Program Book tells you a little of how the Symposium started through the Methodology Research Committee. Also you will have seen the exhibit and video that the Organizers have put together at the Conference centre, with many mementoes and snippets of earlier Symposia, and a lot of familiar faces looking a lot younger. My role in getting these Symposia started could best be described as one of benign acquiescence. Having put in place a Methodology R & D Committee, my contribution was not to suppress a good initiative and not to interfere too much. Thanks to the Committee (under the initial leadership of MP Singh and later David Binder) my interventions were rarely needed and were usually ignored anyway - although annually I did get involved in deciding the topic for the following year – just to make sure that we didn't focus on the analysis of data from complex surveys every year, as some might have had us do, and also to ensure that the Symposia remained on a firm budgetary footing.

Let me start by examining what exactly we are celebrating this year. 2009 minus 25 equals 1984, the year the first Methodology Symposium was held. The more astute among you may be wondering how we can be holding the 25th Symposium in 2009 if we have held one every year since 1984. Don't worry too much about it – 25 or 26, it is close enough for Government statistical work. It seems that after I retired they lost count. Or the counting rules were developed by a statistician of the third kind. What is certain is that this year is the 25th Anniversary of the first Symposium.

I will say first a few words about that year 1984, and more specifically about the circumstances and environment in which this Symposium Series was launched - the circumstances both in Statistics Canada, and in the field of survey methodology more broadly. Then I will have a few comments and observations to make about the Symposium series over the years, finishing with some thoughts about the future and all of this before you get dessert.

2. 1984

The book on 1984 has already been written. For those of us who grew up in the 1950s and 1960s, 1984 was always a special year, long-awaited thanks to Orwell's work. It represented a distant, but not too distant, future in which Big Brother watched over citizens, and a Ministry of Truth defined reality and rewrote history accordingly. Concepts like NewSpeak, ThoughtCrime, and DoubleThink were part of the scenario. Orwell wrote in 1948, in the Post-War period during which many new Government social programs were being put into place, and Government indeed was destined to play a much larger role in the life of citizens than had been the case earlier. He had also seen the roles Governments had taken on in the Soviet Union and in Germany before the War.

Anyway, 1984 eventually arrived and some of Orwell's predictions had indeed come to pass, others had not. For example, Newspeak was well established. StatCan was an example of Newspeak, (as were the clever and irreverent names ending in Can that were thought up for other Ministries). *FedProv* provided a good example of Newspeak and Doublethink in one concept. Its beauty is that it can, as required, stand for harmonized common approaches to important national issues, or it can mean constant jurisdictional bickering with no hope of resolution. We didn't have a Ministry of Truth, Minitrue, but we did have Statistics Canada who defined truth, plus or minus two standard errors, 19 times out of 20. And within STC the System of National Accounts was not shy about rewriting history, they openly published historical revisions. It was only later that Unemployment Insurance became Employment Insurance.

Orwell had much to say about the international scene – and the role of war as a necessary condition for economic stimulus (he hadn't heard of Infrastructure spending), but he fell short in one respect: he did not predict a Methodology Symposium.

3. Situation at STC

Meanwhile back in the Pasture, 1984 was a watershed year at Statistics Canada for quite different reasons. It was the year of a major reorganization which defined an organisational structure with six Fields that has stayed virtually unchanged in STC to this day. This change reflected the transformative thinking of Chief Statistician, Martin Wilk, who had resolved to make the organization more outward-looking in its relations with clients, and more internally collaborative and efficient in its operations. And that is what he did in four short years at the helm. Martin Wilk was a strong believer in Research and Development, in the importance of analysis, and in reaching out to other organizations to help achieve objectives.

One of the organizational changes in that 1984 restructuring was the reassembly of a Methodology Branch. From 1979 to 1984 the Methodology Divisions had been scattered within Subject Matter Branches in different Fields. During this period of separation (if I may use that term here in Quebec), the Directors of the Methodology Divisions saw the need to maintain the close contacts and joint activities between their Divisions that they had enjoyed prior to 1979. In fact, the need for these contacts and exchanges was, in a sense, even more important during this period of separation. One of the initiatives in this context was the creation of a Methodology Research Committee across the Divisions to coordinate the R & D work that was not directly associated with particular subject-matter programs. As David Binder has mentioned in his note, this Committee was the source of the proposal for the first Methodology Symposium. That Committee continued in operation when the Methodology Branch was formed in 1984 and continues to this day, with some changes in name and function.

It is significant to note several other Institutions associated with Methodology that date from this period. The ACSM, STC's external advisory committee on Methodology issues, which meets twice annually, will hold its 50th meeting next year. The internal Methods and Standards Committee (the management committee responsible for corporate issues in this area) held its first meeting in 1983. The journal, *Survey Methodology* celebrated its 25th Anniversary a few years ago in 2000, but its adoption as an official STC publication occurred in 1984 (interestingly the last issue of Survey Methodology in its old format was the one devoted to papers from the 1st Methodology Symposium in 1984). Regular methodology Interchanges with the U.S. Bureau of the Census were also instituted during this period, and subsequently extended to BLS as well.
So this was a period when the Agency had recognized the importance of Methodology to its programs, was recognizing the need to strengthen its external connections and consultations, and was concerned with strengthening internal collaboration and efficiency. The Symposium Series emerged from this environment.

Given his own priorities and impact, it was fitting that the first Symposium should be opened by Martin Wilk, that it should focus on the analysis of survey data, and that it should be co-sponsored by the Laboratory for Research in Probability and Statistics at Carleton and Ottawa universities, an organization that continued to co-sponsor many of the subsequent Symposia, and was instrumental in helping the early Symposia to emerge on a firm financial footing.

4. Professional development

One of the ongoing challenges of managing a large methodology staff is to provide adequate professional development opportunities. With a staff of 200 or more (now closer to 300 I think), it isn't possible that all will have regular Conference attendance possibilities. Conference attendance for Government employees, particularly if outside Canada, has for a long time required special paperwork and approvals. In addition to weighing the inherent value of attendance to the employee and the organization, it was always necessary to consider also the public perception of extensive foreign travel in the prevailing budgetary climate.

Furthermore, in the early 80s there weren't a great number of Conference opportunities of direct relevance to our Methodology work. Many of the issues we faced were only faced by other national statistical agencies and so similar experience had to be found outside Canada. The ASA meetings always provided a range of sessions of relevance to our work, even if these were only a small proportion of the total sessions there. But we could seldom send more than 20 or so people to the ASA. The SSC did not have much of direct relevance at that time – its Survey Methods Section was still in the future. The ISI, especially IASS, provided some good relevant sessions biennially, but few employees got to go to those meetings which were usually held outside North America. So laying on an annual Symposium on a topic directly relevant to our work, and which a large number of our own employees could potentially attend, was an attractive idea. If we couldn't

send our people abroad as much as we would like, then we could at least arrange for some notable speakers to come here. That was, to my mind, always one of the main justifications for the Symposium.

The Symposium also provided the additional benefits, incentives and recognition to those who presented papers, as with any Conference. The imposed deadlines are often a significant incentive in ensuring that real work progress is made. The visibility of the output can enhance professional reputations, and hopefully the corporate reputation too. The Symposium could also provide younger employees an opportunity to gain experience in presenting papers in a familiar environment before venturing further afield.

Of course, since the 1980s, many other Methodology meetings have developed, including ad hoc Conferences on particular topics, often under the aegis of ASA, as well as inter-Governmental meetings under the auspices of UN agencies or the EU. But the Symposium has remained an important professional development opportunity for our methodologists, and hopefully for many from other countries too.

5. Topics and themes

If you look at the list of 25 (or 26) topics covered by the Symposium series over the years it is like time travel through the Methodology challenges and priorities of the last quarter century. The titles reflect the priorities of the day.

As I have already said, it was fitting that the first Symposium, opened by Martin Wilk, should focus on the analysis of survey data. The traditional focus of Methodology within Statistics Canada had been on survey design, and this was an important signal that the analysis of data was the ultimate user concern, and should be getting more attention from methodologists, and consideration at the design stage. Analysis recurs in the Symposium themes periodically to this year. As François mentioned in his opening remarks yesterday, this year's theme of longitudinal surveys can itself be traced back through earlier Symposia since 1989 and 1992. The 1989 Symposium on the Analysis of Data in Time was itself identified as a follow up to an earlier meeting on Panel Surveys in Washington. In 1989 the analysis of data in time was still limited by the lack of longitudinal surveys.

The second Symposium on Small Area Statistics occurred at a time when Statistics Canada was proposing a more organized and coherent approach to this topic across the Agency. This was a full-fledged conference held in the Ottawa Congress Centre – recently demolished. This Symposium combined some of those broad organizational considerations with the methodological interest in small area estimation. It resulted in a Wiley book on Small Area Statistics.

In 1986 we were back in the basement of the Coats Building discussing Missing Data in Surveys. In looking back I at first thought that the records of this Symposium had ironically gone missing, but they have shown up in Volume 12 of *Survey Methodology*. The following year, 1987, the topic was Administrative Data Use and the venue was the Old Railway Station, then the Government Conference Centre. The subject was very topical with renewed efforts to broaden tax data use and the mounting awareness of privacy issues being prominent. I think it is correct to say that this was the first Symposium to produce a self-standing set of Proceedings. And by this time we were clearly referring to the Symposia as a series, and this was the 4th.

While the large Conferences like 1985 and 1987 were successful, they involved a lot of work, and moved us away from STC premises. Some of us felt a certain attachment to the cosy surroundings of the STC basement, but particularly the ease with which STC employees could attend sessions there. But capacity was restricted and access control to the building for visitors was a problem. From 1988 to 1996 we stayed in the Simon Goldberg centre at STC but we were bulging at the walls and the fire marshals were on our case. Also we really only had a single Plenary room. From the 1994 event I clearly remember the closing Panel discussion, a format we used in several Symposia to try to keep the audience till the end. The Chief Statisticians of Australia, Canada, France and New Zealand duked out a four-rounder on who was the best! We had the idea of alternating topics so that one year we would have a major off-site Symposium on a broad topic, and the next year a more narrowly focused topic in the basement. So in 1997 we moved to the Palais de Congrès (this year's site) for the first time, and in 1998 and 1999 we were back at STC. It seemed that even the narrower topics were attracting more registrants than could safely fit into the basement and we had to abandon the alternation idea. It seemed that we were incapable of organizing a small Symposium, only a large one. After 1999 the Symposium has not yet returned to the Simon Goldberg centre in the basement at STC.

A perceived relative neglect of business surveys from the methodology perspective led to the planning of the first ICES conference as our 1993 Symposium. This was the first Symposium not to be held in the Ottawa-Gatineau area, but it was

still close enough, in Buffalo, New York, that we could send as significant number of STC participants. We rented a bus and headed off down the turnpike. That topic was repeated in 2000 in Buffalo, and again in Montreal in 2007.

The issue of Data Quality permeates all our Symposia, but was the focus of attention in 1990, 1996 (Non-sampling Error), and 2001. We have usually avoided restrictions on the subject-matter area of application (except those implicit in ICES), but in 2006 interest in the subject of Population Health was deemed broad and challenging enough to warrant a Symposium addressed to measurement issues in that area.

6. Influence and Emulation

One sign of success is emulation. I believe the Methodology Symposium Series has been emulated in two senses. Firstly, within Statistics Canada, other professional communities have recognized the value of a regular opportunity for staff to present and discuss their work with peers inside and outside the organization. Today, both the IT community and the Economists/Sociologists hold annual professional conferences.

Secondly, several other statistical offices have instituted regular Methodology conferences. It may be a touch presumptuous to claim that these are emulations of the STC model, but we can at least say that these other offices have reached a similar conclusion to STC about the value of such meetings.

7. Bilingualism

I can't end without mentioning one of the main distinguishing features of the Symposia series, its bilingualism. This reflects the working official languages of Canada and allows STC employees to present and participate in their official language of choice. The vast majority of external participants are participating in English, but I have always been struck by the appreciation, and sometimes surprise, of those who do come from other francophone countries that they really can participate in French if they wish. I have also observed that a hidden benefit of bilingualism is the improvement that is made to the English version of a paper as a translator struggles to make sense of it in French – and presumably vice versa too. Translators force us to express ourselves more clearly in our first language.

8. The Future

Now what about the future? I know you are thinking the future is dessert and it better come soon. Bear with me. Reaching a 25th Anniversary is very nice, but it doesn't necessarily imply that reaching a 50th Symposium should be the next target. The real question is whether the Symposium is continuing to meet the needs of the Methodology community both within STC and internationally, and whether it is still the best way to meet those requirements. Much has changed in the world since 1984 – particularly the means of communication available for discourse between statisticians. As we have seen, the Symposium has adapted and evolved in many ways since 1984, but it remains primarily a physical gathering of humans for a brief period each year. And it does require a significant amount of preparation to bring about – as generations of organizing committees can attest. It is a valid question to ask whether, in this day and age, there may not be other mechanisms that can supplement or even replace an annual Symposium. Over the years, I do not recall a great deal of criticism of the Symposia. People are very polite, especially in Canada, and generally people seemed well-pleased with the opportunity to attend. But, if I had to identify one negative comment that recurred from time to time, it would be a regret that there was not more time for discussion on many of the issues presented. This is not a criticism unique to this series of Symposia but is a function of many large meetings and conferences where the format, formality, and the need to keep on schedule, tend to inhibit discussion. If the future shape of the Symposium was under consideration, this would be the aspect where improvement or alternatives might focus. Don't forget to fill in your evaluation forms - no more banquet speeches perhaps.

9. Conclusion

When we held the first Symposium in 1984, I did not expect to be here 25 years later celebrating this anniversary. That we have reached this far is the result of the dedication, perseverance and support of organizers, contributors and participants over this whole period. I would particularly like to mention the program committees and local arrangement committees who have ensured that the Symposia actually happened. These are the people who work largely behind the scenes to ensure that

what we want is in place. Over 25 years that is probably 400 people who have helped to make this series possible. Many of them are here tonight. I would like to end by thanking those Program and Local arrangements organizers who have allowed us to get this far. Your work has been no less important to progress in Methodology than the work of those authors whose papers have taken the spotlight over the past quarter century. Thank you.