C3

Catalogue no. 11-522-XPE

SYMPOSIUM 98

Longitudinal Analysis for Complex Surveys

PROCEEDINGS

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche, microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-8615) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(780) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site:
http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialing area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**
Fax order line (Canada and United States)	**1 877 287-4369**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 11-522-XCB, is produced annually on CD-ROM for $60.00 in Canada. Outside Canada the cost is US $60.00. A paper version no. 11-522-XPE is also available at the same costs.

Please order by mail, at Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, K1A 0T6; by phone, at **(613) 951-7277** or **1 800 700-1033**; by fax, at **(613) 951-1584** or **1 800 889-9734**; or by Internet, at order@statcan.ca. For changes of address, please provide both old and new addresses. Statistics Canada products may also be purchased from authorized agents, bookstores and local Statistics Canada offices.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Methodology Branch

SYMPOSIUM 98

Longitudinal Analysis for Complex Surveys

PROCEEDINGS

Published by authority of the Minister responsible for Statistics Canada

October 1999

Catalogue no. 11-522-XPE
ISBN 0-660-17745-5

Frequency: Occasional

Catalogue no. 11-522-XCB
ISSN 1481-9678

Frequency: Annual

Ottawa

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Canadian Cataloguing in Publication Data

Symposium 98, Longitudinal Analysis for Complex Surveys
(1998 : Ottawa, Ontario)
Symposium 98, Longitudinal Analysis for Complex
Surveys : proceedings

CD-ROM bilingual version also issued under title:
Symposium 98, Longitudinal Analysis for Complex Surveys : proceedings.
ISSN 1481-9678
CS11-522-XCB

Paper version issued also in French under title:
Symposium 98, Analyse longitudinale pour les enquêtes complexes : recueil.
ISBN 0-660-96131-8
CS11-522-XPF

1. Longitudinal method – Congresses. 2. Statistics –
Methodology – Statistics. I. Statistics Canada. Methodology
Branch. II. Title.

HA12 S95 1999 300.7'27
C99-988034-9

PREFACE

Symposium 98 was the fifteenth in the series of international symposia on methodological issues sponsored by Statistics Canada. Each year the symposium focuses on a particular theme. In 1998, the theme was Longitudinal Analysis for Complex Surveys. The event was cosponsored with the Centre de recherche mathématiques from l'Université de Montréal.

The 1998 symposium attracted over 300 people from 14 countries who met over two days, in the Simon Goldberg Conference Centre in Ottawa to listen to experts from various statistical and government agencies as well as representatives from the private sector. A total of 25 papers was presented by the invited speakers. Aside from translation and formatting, the papers, as submitted by the authors, have been reproduced in these proceedings.

The organizers of the Symposium 98 would like to acknowledge the contribution of the many people, too many to mention individually, who helped make it a success. Over 40 people volunteered to help in the preparation and running of the symposium, and 25 more verified the translation of papers submitted for this volume. Naturally, the organizers would also like to thank the presenters and authors for their presentations and for putting their presentations in written form. Finally, they would like to thank Carole Jean-Marie and Lynn Savage for processing this manuscript.

The Proceedings of the 1999 annual symposium entitled Combining Data from Different Sources are currently being prepared. In 2000, the symposium will be replaced by the *International Conference on Establishment Surveys*, as was the case in 1993. This conference will be held in Buffalo, New York. The next Statistics Canada symposium will be held in Ottawa in 2001.

Symposium 98 Organizing Committee

Michael Hidiroglou

Tony Labillois

Sylvie Michaud

Denis Lemire

Georgia Roberts

Michel Latouche

STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data	**1993** - International Conference on Establishment Surveys
1985 - Small Area Statistics	**1994** - Re-engineering for Statistical Agencies
1986 - Missing Data in Surveys	**1995** - From Data to Information: Methods and Systems
1987 - Statistical Uses of Administrative Data	**1996** - Nonsampling Errors
1988 - The Impact of High Technology on Survey Taking	**1997** - New Directions in Surveys and Censuses
1989 - Analysis of Data in Time	**1998** - Longitudinal Analysis for Complex Surveys
1990 - Measurement and Improvement of Data Quality	**1999** - Combining Data from Different Sources
1991 - Spatial Issues in Statistics	**2000**- International Conference on Establishment Surveys II
1992 - Design and Analysis of Longitudinal Surveys	

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES PROCEEDINGS ORDERING INFORMATION

Use this two page order form to order additional copies of the proceedings of Symposium 98: Longitudinal Analysis for Complex Surveys. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 98 PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 6th FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada" - Indicate on cheque or money order: Symposium 98 - Proceedings Canada).

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	_____	@ $10
1987 -	Les utilisations statistiques des données administratives - FRENCH	_____	@ $10
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	_____	@ $12
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	_____	@ $10
1989 -	Analysis of Data in Time - BILINGUAL	_____	@ $15
1990 -	Measurement and Improvement of Data Quality - ENGLISH	_____	@ $18
1990 -	Mesure et amélioration de la qualité des données - FRENCH	_____	@ $18
1991 -	Spatial Issues in Statistics - ENGLISH	_____	@ $20
1991 -	Questions spatiales liées aux statistiques - FRENCH	_____	@ $20
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH	_____	@ $22
1992 -	Conception et analyse des enquêtes longitudinales - FRENCH	_____	@ $22
1993 -	International Conference on Establishment Surveys - ENGLISH (available in English only, published in U.S.A.)	_____	@ $58
1994 -	Re-engineering for Statistical Agencies - ENGLISH	_____	@ $53
1994 -	Restructuration pour les organismes de statistique - FRENCH	_____	@ $53
1995-	From Data to Information - Methods and Systems - ENGLISH	_____	@ $53
1995-	Des données à l'information - Méthodes et systèmes - FRENCH	_____	@ $53
1996-	Nonsampling Errors - ENGLISH	_____	@ $55
1996-	Erreurs non dues à l'échantillonnage - FRENCH	_____	@ $55
1997-	New Directions in Surveys and Censuses - ENGLISH	_____	@ $80
1997-	Nouvelles orientations pour les enquêtes et les recensements - FRENCH	_____	@ $80
1998-	Longitudinal Analysis for Complex Surveys - ENGLISH	_____	@ $60
1998-	L'analyse longitudinale pour les enquêtes complexes - FRENCH	_____	@ $60
1998	Longitudinal Analysis for Complex Surveys -BILINGUAL on CD-ROM	_____	@ $60
	PLEASE ADD THE GOODS AND SERVICES TAX (7%) (Residents of Canada only)	$_____	
	TOTAL AMOUNT OF ORDER	$_____	

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME ______________________________

ADDRESS ______________________________

CITY ______________________________

PROV/STATE ______________________________

COUNTRY ______________________________

POSTAL CODE ______________________________

TELEPHONE (______)______________________________

FAX (______)______________________________

For more information please contact John Kovar: Telephone (613) 951-86155, Facsimile (613) 951-5711.

LONGITUDINAL ANALYSIS FOR COMPLEX SURVEYS

TABLE OF CONTENTS

SESSION 3: CORRECTING FOR NONRESPONSE

Chairperson: John Kovar

SESSION 4: GROSS FLOWS I

President: J. Gambino

SESSION 5: GROSS FLOWS II

Chairperson: G. Hole

OPENING REMARKS

OPENING REMARKS

Michael Wolfson, Statistics Canada

Good morning and a very warm welcome to Symposium 98.

As a preface, let me say that I'd rather not be here! Originally, Dave Binder was slated to give these opening remarks. The great news is that David is on his way to a full recovery from his illness.

This is the 15th in this series of methodology symposia sponsored by Statistics Canada. The program for this year is impressive. I am particularly pleased to note a high level of international participation, including speakers from the United Kingdom, the U.S.A, France, Switzerland, and Israel.

This symposium originated with an invitation to participate in a workshop on the analysis of longitudinal data and complex survey data from the Centre de recherches mathématiques (CRM) of the Université de Montréal in March 1996. Statistics Canada subsequently suggested that we link the workshop to the symposium (bring together the symposium and workshop ideas). Here we are, 2 years later, with the results of this good idea.

I would like to extend a very warm welcome from Statistics Canada and the City of Ottawa to our visitors, and also to the personnel of Statistics Canada, who are very well represented in the program as well as in the audience.

Since its start in 1984, this series of symposia has covered many survey-related topics. To name just a few, we have dealt with statistics from small regions, administrative data and the impact of technology on surveys.

This year's topic is closely related to the topics of two of our previous symposia: Analysis of Data in Time, in 1989, and the Design and Analysis of Longitudinal Surveys, in 1992. Among the new elements in this symposium are questions surrounding methods for taking complex survey designs into account in the analysis, and the software environments required for managing the complex data sets resulting from longitudinal surveys.

The general topic of Symposium 98 is very timely for Statistics Canada, since we have recently launched several major new longitudinal surveys, and we have created important new longitudinal data sets drawing on administrative data. As a result, we are now in the throes of analysing these new longitudinal data -- or rather, to put it in the positive light it deserves -- beginning to harvest the first fruits of these important data development investments.

This recent shift in emphasis within Statistics Canada's data collection programs -- from almost a complete reliance on cross-sectional designs, to an increasingly important number of longitudinal surveys and administrative data sets -- represents a fundamental maturation in the ways we think about our role as a statistical agency.

Cross-sectional data are fine for monitoring socio-economic patterns and trends in Canadian society. And they can provide suggestive information about possible explanations for key patterns – such as the positive association between education and income, or between income and health.

However, it is only with longitudinal data that we can begin to have some hope of teasing out the underlying causal stories. For example, is it low income that leads to poor health, or is it failing health that leads to declines in income? And even leaving the difficult questions of causal inference aside, these longitudinal data offer a much richer opportunity for description and monitoring of social processes – in particular, descriptions of dynamics and transitions.

Historically, one major example of the new descriptive insights afforded by longitudinal data is the U.S. Panel Study on Income Dynamics, and its early finding that, while the overall proportion of families living in poverty did not change that much from year to year, there was considerably more movement of individual families into and out of poverty over time.

The need to begin uncovering these kinds of dynamic patterns and underlying causal pathways has been appreciated by our colleagues in government policy ministries. As a result, they co-operated in securing for Statistics Canada the new funding for our three premier longitudinal surveys –

- the National Longitudinal Survey of Children and Youth (NLSCY),
- the National Population Health Survey (NPHS), and
- the Survey of Labour and Income dynamics (SLID),

all of which began collecting data from households in 1993 or 1994.

The NLSCY or "kiddies" survey selected cohorts of children between birth and age 15 years, and is following them with household interviews of the parents, and the kids themselves as appropriate, every 2 years. In addition, data are gathered from the childrens' teachers and principals, which we have coupled with "ecological" data on neighbourhood socio-economic characteristics from the population census.

The second major new survey, the NPHS, has a core nationally representative sample of about 17,000 individuals, including those living in institutions. They too are being followed longitudinally every two years. In addition, all members of their household at interview time are included in the survey, and to varying degrees, provinces have purchased substantial extra sample. Moreover, about 95% of the sample respondents have given us permission to link the survey to their health care records, along with their health insurance numbers to facilitate the linkage. As a result, we are in the process of creating an extra-ordinarily rich longitudinal data set combining both self-report and clinical descriptions of health state trajectories. Thus, we are finally able to join, at the microdata level, the two solitudes of what I would call the vernacular and bio-medical views of health.

The third survey, SLID, follows cohorts of about 15,000 families annually over 6 years. And every 3 years, a new cohort is drawn. Thus, we have in each year a total sample of about 30,000 families, half in each of two longitudinal cohorts. After the end of each year, individual respondents are asked in detail about their incomes and their jobs, and are asked to reconstruct the start and end times of each job they had during the previous year. So SLID gives income and labour market transition data with sub-annual detail over a multi-year period.

In addition to these new surveys, the federal ministry of Human Resources Development has supported Statistics Canada in developing administrative data sources – specifically a 10% longitudinally linked sample of personal income tax returns starting in 1980 (the LAD). This is also valuable as a complementary source of data on income dynamics. While the data are of excellent quality, and their marginal cost to us is low, there is a problem of a lack of co-variates.

I would characterise this as a "wide but thin" data set – i.e. many observations but few potential "explanatory" variables. In contrast, the three new longitudinal surveys I just mentioned are "narrow but deep". For example, they have insufficient sample size to provide much in the way of sub-provincial estimates, but they do include a wealth of covariates with which we can begin to explore the rich structure of relationships associated with the patterns in respondents' unfolding life course trajectories.

Also, in a precedent-setting initiative with the Manitoba Ministry of Health and the University of Manitoba, we have created a linked file drawing together person-level longitudinal Manitoba health care administrative records, socio-economic data from the 1986 population census, and disability data from the Health and Activity Limitations Survey.

Moreover, with the major improvements we are making in our business register, and new surveys such as the Workplace Employee Survey (WES), we will also have a growing base of longitudinal data on businesses. WES is the first panel survey at Statistics Canada where data about employees and employers will be gathered both jointly and on a longitudinal basis. This will allow us to study the "pit face" of the labour market – bringing together at long last the supply and demand curves that everyone sees on the first day of Econ 101, but which until now could only be studied in isolation, at least at the micro level. In terms of contemporary labour market issues, WES has been designed to shed light on such questions as the interaction of profitability and the technologies adopted by the employer, and the wages and skills of the employees.

This new wealth of data brings with it, of course, a wealth of statistical challenges – many of which are the focus of this Symposium. For example, the data we are collecting can have very complex structures – as in the NLSCY where we have data not only for the child (with content varying by age), and his or her parents, but also the teacher, the school principal, and contextual data such as neighborhood characteristics and attributes of classmates.

The software issues related to handling, processing and analyzing these data are non-trivial. They also appear sufficiently unique to a relatively small group of statistical agency and researcher users that I would not hold my breath for an off-the-shelf solution to emerge quickly from the private sector – notwithstanding all the recent hype about "data warehousing" and "data mining".

Another challenge is inference when we have data at several levels of observation, such as children in the context of their schools and neighborhoods. Analogous questions are arising in the health area, where recent evidence clearly suggests that social context, over and above individual characteristics, has an influence on longevity. But we have yet to tease out of the data better insights as to the main elements of the causal pathways.

Yet a further challenge, one of particular interest to me, is one that I don't see reflected in the program. This is how to cluster individuals by the "character" of their longitudinal pathways. I am thinking here, for example, about patients who have had heart attacks – where we already have masses of detailed longitudinal data on all their subsequent doctor and hospital visits, as well as "upstream" data on their risk factor and socio-economic backgrounds. I have this image of a river delta, with many different rivulets wending their way through – and correspondingly for the cohort of cardiac patients, some wending their way through the health care system via coronary bypass surgery, others via angioplasty, etc. The question is how best might we create a map of these rivulets, i.e. infer the extent of structure in the subsequent treatment profiles of this group of patients?

Well, that can perhaps be a challenge for a future symposium. For now, you already have a rich tableau of stimulating statistical issues. I wish you all well in these discussions. I am sure that we in Statistics Canada will benefit from the intellectual discourse, and I trust that everyone will be able to conclude the symposium having gained useful new insights and ideas.

SESSION 1

PREPARING/STORING/SOFTWARE

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

CYCLE 2 AND BEYOND: PREPARING AND STORING LONGITUDINAL DATA OF THE NATIONAL POPULATION HEALTH SURVEY

Peter Fobes and Leslie Geran[1]

ABSTRACT

The National Population Health Survey (NPHS) is a family of surveys with multiple objectives, one of which is to provide information on a panel of people who will be followed over time to reflect the dynamic process of health and illness. Data for the first cycle of the NPHS - Households Survey were collected from June 1994 to June 1995, and were released in September 1995. Data for the second cycle were collected from June 1996 to August 1997. One of the primary outputs for the second cycle is a longitudinal master file. This paper will describe six major strategies that were developed to process the longitudinal master file.

KEY WORDS: Longitudinal surveys; Data processing; Health surveys; Relational databases; Metadata storage.

1. INTRODUCTION

Statistics Canada's National Population Health Survey (NPHS) is a family of surveys with multiple objectives, one of which is to provide information on a panel of people who will be questioned every two years for up to twenty years to reflect the dynamic process of health and illness. Data for two cycles of the NPHS have been collected and processed. This paper describes how the longitudinal data were prepared and stored in order to facilitate longitudinal analysis, and discusses the processing problems unique to the processing of longitudinal survey data.

1.1 Objectives of the Survey

The NPHS was created to fill a data gap identified by the National Task Force on Health Information in 1991. The NPHS is designed "to be flexible, and to produce valid, reliable and timely data … to be responsive to changing requirements, interests and policies". The survey has several broad objectives. The NPHS provides measures of the level, trend and distribution of the health status of the population. It also provides information on a panel of people who will be followed over time to reflect the dynamic process of health and illness. Provinces, territories and other clients are permitted to supplement the survey with content or sample. NPHS survey data can be enriched by linking to administrative data such as vital statistics and health services utilisation.

Turning these objectives into sets of data to fill the information gap require that both cross-sectional and longitudinal data be collected, processed and stored. The need to provide both cross-sectional and longitudinal data in a timely manner was a major force behind how the processing of the longitudinal data evolved.

[1] Peter Fobes (fobepet@statcan.ca) System Development Division and Leslie Geran (gerales@statcan.ca), Health Statistics Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

1.2 Survey Design

The original household sample consisted of a core sample of approximately 22,000 households in 10 provinces to ensure reliable estimates by sex and broad age groups. [1] In all provinces but Quebec, the Labour Force Survey (LFS) design of a multi-stage stratified sample of dwellings selected within clusters was used to draw the sample. In Quebec, the sample was drawn from the sample design of the 1992 - 1993 health survey organised by Santé Quebec. The design of the Enquête sociale et de santé (ESS) is a two-stage design similar to that of the LFS.

During collection, when a dwelling was identified as being in-scope, the basic demographic data (name, date of birth, age, sex, and marital status) for each household member was entered into a roster. Every member of the household was asked to complete a short general questionnaire, and one member of the household was selected to answer an in-depth health questionnaire. The longitudinal panel is defined as every selected household member who had completed at least the general questionnaire in Cycle 1: 17,276 respondents.

Sample collection is carried out in four collection periods per cycle, in June, August, November and February. A small fifth collection period in the following June is used to convert non-response from the previous collection periods. In Cycle 1, the collection lasting from June 1994 to June 1995, the collection period was two weeks long. In Cycle 2 (June 1996 to June 1997) and beyond, collection periods are approximately six weeks long. LFS interviewers trained in NPHS concepts use laptop computers to collect the data with a computer-assisted interviewing (CAI) questionnaire.

The longitudinal panel was augmented in both Cycle 1 and Cycle 2 by sample buy-ins by provinces wishing to obtain more reliable sub-provincial estimates. In Cycle 3, the core sample will be augmented to ensure cross-sectional representativity.

1.3 Survey Content

There are four major parts to the survey content each cycle. First, core questions are asked every cycle. In the General questionnaire, there are sections on socio-demographic questions such as country of birth, level of education, labour force activity, and income. The General questionnaire also has a few health-related questions such as two-week disability, activity restrictions, and chronic conditions. In the Health questionnaire, core content includes height and weight, smoking, alcohol consumption, injuries, physical activities, drug consumption, mental health, and social support.

Second, every cycle has focus content. In Cycle 1, baseline data on work and life stressors were collected. In Cycle 2, access to selected health care services such as physical check-ups, dental care, blood pressure checks, flu shots, and women's health concerns were featured. In Cycle 3, family health history and self-care questions will be featured. Focus content may be repeated in future cycles.

Third, clients may buy content. Health Canada bought large sets of questions on smoking behaviour in both Cycle 1 and Cycle 2. Several provinces also bought questions on mental health and coping behaviours.

Fourth, in Cycle 2 and beyond, sets of probing questions are asked. Selected data from previous cycles are part of the sample file, and these data are available when the computer generates the CAI questionnaire. Respondents who give answers inconsistent with responses from the previous cycle are asked probing questions to resolve the inconsistency. The answers to the probing questions are used to edit the longitudinal data. In this respect, longitudinal processing starts with questionnaire design.

2. PROCESSING STEPS

Preparing the longitudinal data followed the same steps as the cross-sectional data, because the Cycle 2 data would also appear on a cross-sectional data file. Additional steps were needed to create a longitudinal file. The Cycle 1 master file was combined with the Cycle 2 file, but we found that some of the processes had to be repeated in order to have valid, consistent longitudinal data (see Graph 1: Processing Steps).

[1] Separate designs were used for surveys of the institutionalised population and for the population of the two territories. These surveys will not be further discussed in this paper, but the broad approaches outlined below were used to create longitudinal files for these populations.

For the cross-sectional processing of every NPHS cycle, data are collected using computer-assisting interviewing. Sample files are sent electronically to the regional offices, and the cases are assigned to interviewers. Interviewers contact and interview respondents, and send completed questionnaire files back to head office via the regional office. At head office, the data files are examined for completeness, and a preliminary response code is assigned. Long answers that require coding are stripped off the files and sent for coding. Interviewer notes are examined for any important data that could not be entered into the CAI questionnaire.

Next, the reformat stage standardises answer codes sets from the different software packages that make up the CAI application. Reformat sets the flow of the questionnaire, which may have changed as a result of critical updates. Reformat also generates a final response code, updates any non-critical updates such as the re-coding of other specify answers, and adds the coding for the long answers, such as the standard occupational classification (SOC).

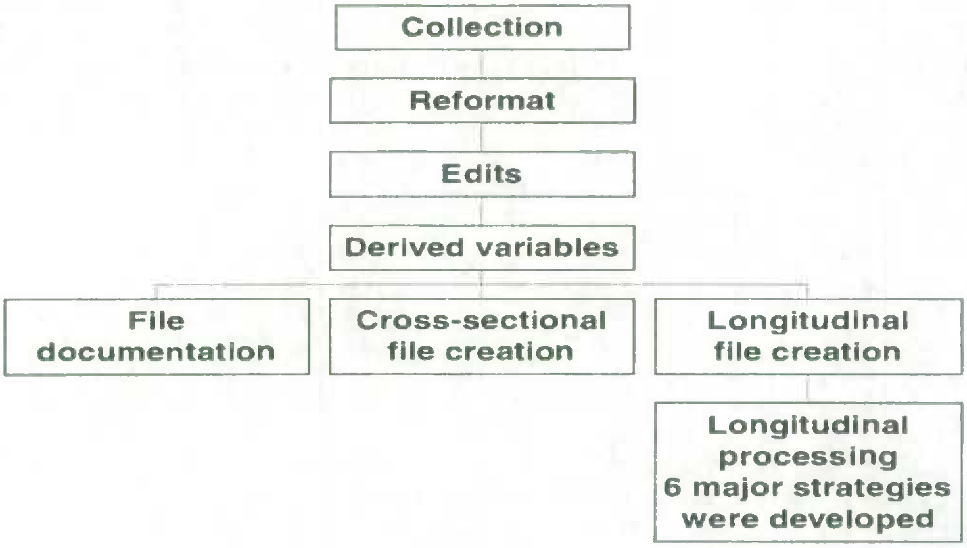
The editing stage resolves any inconsistencies between answers. Very few edits are needed because the CAI program includes range checks and logical references to previous answers.

Next, derived variables are created which use the answers of one or many collected variables to create a new variable. For example, from the responses to questions on a respondent's height and weight, a new variable called body mass index is calculated. This is a ratio of a person's height to weight, and is used to estimate whether or not a person weighs an appropriate amount.

The final step to create a master data file is to drop any intermediate processing variables, and add sample weights.

After the Cycle 2 derived variable program had been run, all of the variables were ready to create cross-sectional master files. Additional steps were needed to create a longitudinal file. The Cycle 1 master file was combined with the Cycle 2 file, but we found that some of the processes had to be repeated in order to have valid, consistent longitudinal data. We developed and implemented strategies to resolve six major longitudinal processing issues.

Graph 1: Processing Steps


Collection
Reformat
Edits
Derived variables
File documentation
Cross-sectional file creation
Longitudinal file creation
Longitudinal processing 6 major strategies were developed

2.1 Naming Convention

The first issue was how to refer to questions, variables and concepts over time. Some sort of naming convention was needed to distinguish the variables created from the same questions from different cycles. We created a naming convention where each position within the 8 character variable name gives information about the variable (current statistical software packages limit the length of a variable name to 8 characters). Characters one and two are an acronym of the section heading on the questionnaire. The third character denotes whether the variable is core content or a buy-in question. The fourth character is for the cycle. The fifth character describes how the data are obtained – collected, derived or grouped. Characters six, seven and eight are reserved for the question number or variable name. For example, HWC6DBMI refers to a variable in the height weight section, core content, from Cycle 2 (1996). The variable is derived, and the name is BMI (body mass index).

Each Cycle 2 variable was named using this convention, and the variables from Cycle 1 were re-named. The processing programs use the re-named variables.

2.2 Coding

The second issue was how to code long answers to questions when the classification systems change over time. We do not want to artificially introduce change in a coded variable because of a change in the classification system. Five major sets of classification systems are used to code long answers collected in the NPHS. Causes of activity limitations are coded with the International Classification of Diseases (9th revision). Drugs are coded with the Canadian Anatomical Therapeutic Chemical (ATC) Classification System. Labour force variables are coded using the Standard Occupational Classification (1980) and the Standard Industrial Classification (1980). Geographic variables use the Postal Code Conversion File (PCCF) to link to the 1991 Census geography. All of these classification systems are currently under revision.

For Cycle 2 and beyond, we decided to adopt a strategy that the same classification systems be used throughout all cycles. When a new classification system is adopted, historical data will be re-coded. For example, the drug coding system changed between Cycle 1 and Cycle 2. All of the Cycle 1 long answers for drugs were re-coded, and the answers placed on the longitudinal file. If time, space, and budgets allow, long answer variables may be coded using more than one classification system. If re-coding is not done, the differences between classification systems must be noted in the documentation. For example, the urban / rural geographic variable in Cycle 1 was based on an LFS sample design indicator, not census geography. The documentation warns that changes in estimates calculated at the urban / rural level may be as a result of the change in definition and not due to respondents who moved.

2.3 Longitudinal Edits

The third issue we needed to address was how to handle inconsistent answers from different cycles. We needed a longitudinal editing strategy. Upon examining the data, we found two types of inconsistencies. Type 1 inconsistencies were variables that could be regarded as constants, for example, date of birth and sex. Cycle 2 birth dates were different from Cycle 1 birth dates for some respondents. In many cases, the month and the day of the birth was changed from a non-response ("Don't know" or "Refusal") to a valid response. In 55 cases, the change in birth date was large enough to create a large change in the age at interview. In many of these cases, an interviewer note confirmed that the Cycle 2 response was a correction. We decided to accept the Cycle 2 response, and we changed the Cycle 1 data to be consistent with the Cycle 2 date of birth. The age at Cycle 1 interview was re-calculated using the Cycle 2 birth date, and the skip patterns of the Cycle 1 variables were changed. Some derived variables were re-calculated, for example, the age at immigration and the age when the respondent started smoking. We expect the number of these inconsistencies to be reduced over time, since they appear to be Cycle 1 input errors or recall errors by respondents.

Type 2 inconsistencies did not have a major impact on questionnaire flow. We examined pairs of variables such as the Cycle 1 and Cycle 2 responses to the questions "ever smoked", "ever had a mammogram", and "ever drank alcohol". A surprising amount of inconsistency was found in some of the variables. For example, 194 women who said they had had a mammogram in Cycle 1 said they had never had one in Cycle 2. About half of these women said they had had the mammogram two or more years before the Cycle 1 interview. It was difficult design an edit for these inconsistencies, since the inclination was to believe the most recent answer or edit variables based upon probing questions, which were absent for these questions. In addition, having only two cycles of data to compare gave little to base an edit. After consultation with several analysts, it was determined that these inconsistencies are of analytic interest to researchers, who should decide for themselves how to interpret the inconsistency. In Cycle 3, some of these inconsistencies will be probed.

2.4 Longitudinal Derived Variables

The fourth issue was what kind of variables should be derived from the longitudinal data and placed on the data file. Many change variables could be calculated, but the analysts we consulted wanted to calculate their own variables. We created only four longitudinal derived variables for the Cycle 1 / Cycle 2 longitudinal file, and these were for two reasons. Not all variables can be put on the master file for confidentiality reasons. For example, a change in residence variable was created from the addresses that do not appear on the file. We also created a longitudinal variable if the calculation was difficult for the analyst. For example, we created a variable for the number of days between Cycle 1 and Cycle 2 interviews. This saves the researcher from creating dates from six month, day and year variables.

2.5 Linking to Administrative Data

The fifth issue was what type of administrative data should be put on the file. Records for respondents who died before the Cycle 2 interview are included in the longitudinal file. The Cycle 2 data have been filled with "not stated" codes, except for the variables concerning the death of the respondent. Preliminary status codes for death are assigned to records when the NPHS respondent relations team receives post office returns marked "deceased" or when, during data collection, interviewers talk to relatives of the deceased and try to obtain a date of death. These records, along with the records of respondents who were not traced in Cycle 2, are linked to administrative records for confirmation.

Statistics Canada collects provincial mortality records and combines them to produce a yearly national file, but the national file lags behind the survey data, and deaths occurring in the later part of the data collection period cannot be confirmed in time to put the results on the longitudinal file. We decided that for Cycle 2 and beyond, the preliminary death status codes will be put on the longitudinal file, and confirmed data will be added in future cycles, as the administrative records are made available. When a match is made, the date of death and a code for the cause of death are copied to the longitudinal file.

The longitudinal data may be linked to other administrative data (for example, doctor visits) in future cycles, although for confidentiality reasons these data might be placed on a secure file other than the master file.

2.6 Data storage and retrieval

The sixth issue we faced was how to design a data storage and retrieval system that could handle a large amount of data and documentation, and could be easily used by a variety of analysts. Three major factors influenced how the NPHS longitudinal data could be stored and retrieved. First, the survey content changes greatly from cycle to cycle. Not every variable appears in every cycle.

Second, there are many variables. For the Cycle 1 / Cycle 2 longitudinal file there are approximately 2300 variables. Most analysts are interested in only a subset of variables such as mental health or chronic conditions. It is important for the users to be able to easily choose the variables of interest (including the appropriate sample weight), select a population and create a subset of the longitudinal file for their particular analysis needs.

Third, the technical capabilities of our analysts and the computer resources available to our analysts vary greatly. For example, some analysts have access to large and powerful hardware and sophisticated analysis software such as SAS, while others may only have access to spreadsheets such as Excel. In addition, not all analysts are familiar with relational database concepts. We do use a relational database to store our data dictionary. The data dictionary is used to create frequency and record layout reports for information products such as public-use files and sharing files. The analysts, however, do not need to manipulate the data dictionary database in order to read the reports, although they would need to manipulate the data database to do analysis.

For storing the longitudinal data file, we are considering several options. Option 1 is storing the data as one big flat file, with each cycle of data concatenated to the previous cycles. This how we are storing our file of Cycle 1 and Cycle 2 longitudinal data. It was relatively easy for us to construct, and analysts are very familiar with this type of data storage. The users manually browse the record layout report and content report to locate the variables they are interested in and then use software such as SAS or SPSS to load the data file and select the variables and population of interest. With Cycle 3 and beyond, however, the size of the file would become unwieldy. Without some type of retrieval software, even the simple task of finding a variable on a record layout would be onerous.

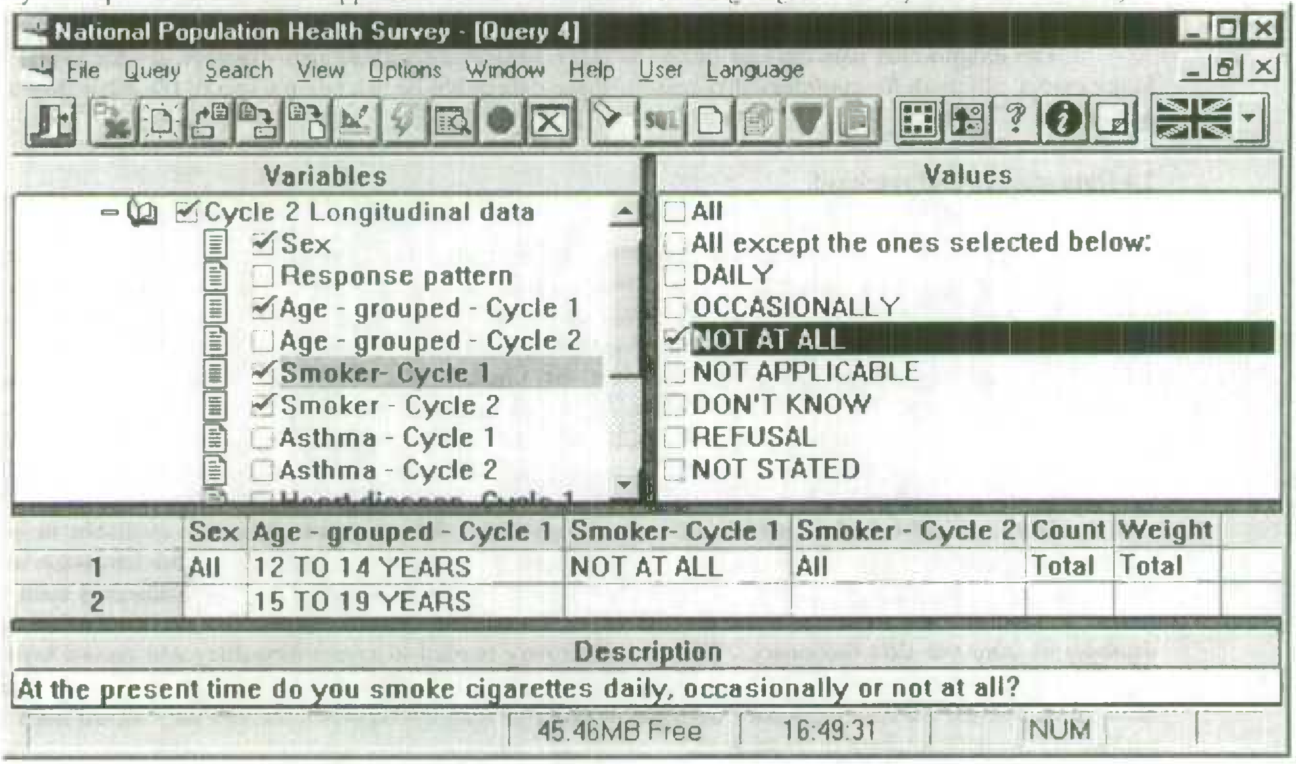
Option 2 is storing the data as a set of time series, but not all variables appear in every cycle. Many blanks would appear in the time series, and this is not an efficient use of space. This method would be more appropriate for a longitudinal survey whose content does not change throughout the life of the survey.

Option 3 is storing the longitudinal file as a series of relation database tables linked by a common key. We have developed a prototype with the Cycle 1 and Cycle 2 data using Microsoft ACCESS, but other database run time systems could be used such as FoxPro. Users would need to understand the database schema and use structured query language (SQL) to retrieve records from the file. In addition, the data dictionary itself could be released to the users to facilitate locating variables and file contents. Again, not all users have the time or resources to understand the schema, so we would like to present the relational database appear as a big flat file.

A data retrieval system developed at Statistics Canada does this. We are prototyping software called the Information Retrieval and Meta-information Administrator system (IRMA). IRMA is a generic software program that handles large amounts of data. For the NPHS application, IRMA binds the data dictionary to the longitudinal file with a point-and-click interface that allows the user to do two major operations: to quickly view a frequency with selection criteria, and to create a subset of records with these criteria.

IRMA requires a Windows operating system, and both the data dictionary and the data must be stored in relational tables. IRMA is written in C++. It creates SQL requests through the point-and-click interface and these requests are passed to the database using the open database connectivity (ODBC) protocol. "Query 4" is an example of a data request using IRMA.

System options include data suppression rules and random rounding to guard survey data confidentiality.



Variables:

- ☑ Cycle 2 Longitudinal data
  - ☑ Sex
  - □ Response pattern
  - ☑ Age - grouped - Cycle 1
  - □ Age - grouped - Cycle 2
  - ☑ Smoker- Cycle 1
  - ☑ Smoker - Cycle 2
  - □ Asthma - Cycle 1
  - □ Asthma - Cycle 2

Values:

- □ All
- □ All except the ones selected below:
- □ DAILY
- □ OCCASIONALLY
- ☑ NOT AT ALL
- □ NOT APPLICABLE
- □ DON'T KNOW
- □ REFUSAL
- □ NOT STATED

	Sex	Age - grouped - Cycle 1	Smoker- Cycle 1	Smoker - Cycle 2	Count	Weight
1	All	12 TO 14 YEARS	NOT AT ALL	All	Total	Total
2		15 TO 19 YEARS				

Description: At the present time do you smoke cigarettes daily, occasionally or not at all?

45.46MB Free | 16:49:31 | NUM

3. CONCLUSION

For the first NPHS longitudinal file, the motto was "keep it simple". The simplest possible data model was used. All data for all respondents are on one square file. Few longitudinal edits were done, and few longitudinal derived variables were created. In database terms, the file structure is highly de-normalised. Processing efforts were directed into creating a file that would be easy for the analysts to understand and use, instead of trying to determine what the analysts want to do. A naming convention was created, and classification systems were changed if necessary. For the few longitudinal edits that were done, all Cycle 1 and Cycle 2 variables were reviewed to make sure that the changes did not introduce any errors in the flow of the questionnaire. The data file structure selected, a flat file, is familiar to most analysts.

For Cycle 3, processing the longitudinal file may become more complicated if imputation for non-response is done. The processing team is much better able to meet that challenge now that the groundwork for "keeping it simple" has been set.

REFERENCES

Fobes P, and Lim P.K. (1998). The Information Retrieval and Meta-Information Administrator System. Internal Document, Statistics Canada.

Geran L. (1997). Longitudinal Edit Checks. Internal document, Statistics Canada.

National Health Information Council (1991). *Health Information for Canada: Report of the National Task Force on Health Information.*

Statistics Canada (1995). *National Population Health Survey Overview 1994-95.* (Statistics Canada, Cat. No. 82-567) .

Tambay J-L. and Catlin G. (1995). Sample design of the National Population Health Survey. *Health Reports* (Statistics Canada, Cat. No. 82-003), 7(1): 29-38.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH: CONFIDENTIALITY AND REMOTE ACCESS

Jean Pignal[1]

ABSTRACT

In 1994, Statistics Canada introduced a new longitudinal social survey that collects information from about 23,000 children spread over 13,500 households. The objective of the National Longitudinal Survey of Children and Youth is to measure the development and well being of children until they reach adulthood. To this end, the survey gathers together information about the child, parents, neighbourhood as well as family and school environment. As a consequence, the data collected for each child, is provided by several respondents, from parents to teachers, a situation which contributes to an increased disclosure risk. In order to reach a balance between confidentiality and the analytical value of released data, the survey produces three different microdata files with more or less information. The master file that contains all the information is only available by means of remote access. Hence, researchers do not have direct access to the data, but send their request in the form of software programs that are submitted by Statistics Canada staff. The results are then vetted for confidentiality and sent back to the researchers. The presentation will be devoted to the various disclosure risks of such a survey and to the tools used to reduce those risks.

[1] Jean Pignal, Special Surveys Division, Statistics Canada, Jean Talon Building, 5th floor, section D5, Tunney's Pasture, Ottawa (Ontario) Canada, K1A 0T6

Proceedings of Statistics Canada Symposium 98
Longitudinal analysis for Complex Surveys
May 1998

RECENT DEVELOPMENT OF SOFTWARE TO ANALYSE LONGITUDINAL STUDIES: EXAMPLES FROM THE OFFICE FOR NATIONAL STATISTICS LONGITUDINAL STUDY

Michael Rosato[1]

ABSTRACT

In epidemiology analysis of longitudinal data is commonly accepted as providing the most robust measures of association between putative risk and selected outcomes such as death or cancer. SMARTIE is a SAS application for efficient analysis of longitudinal data. Based on person days at risk, it can handle multiple exits from and re-entries to risk, and derives outcome measures such as survival rates, Standardised Mortality Ratios (SMRs) and Cancer Incidence Ratios (SIRs). Summary data can be produced in a format easily ported to any modelling package such as Stata 5.0. We discuss the background to its development, the overall program structure, its command language, and finally we say something about the organisation of outputs. Findings from survival studies using the Longitudinal Study of the Office for National Statistics (ONS) are used to demonstrate features of SMARTIE. This study is based on one per cent of the population of England and Wales. It is continually updated with the addition of new members and with information from birth, death and cancer records, and from the census.

1. BACKGROUND

Longitudinal studies share common key features. A selected population is tracked over a de-limited period of time and information on relevant exposures, and events occurring to members, are collected. This accumulating information eventually encapsulates relevant life histories that can then be subjected to statistical analysis. One such study is the Longitudinal Study (LS) based at the Office for National Statistics of Great Britain. While it is described in detail elsewhere (Goldblatt, 1990 and Hattersley, 1995), it can briefly be defined as a record linkage study integrating information collected at successive national censuses (1971, 1981 and 1991) for a representative one per cent sample of the population of England and Wales. This linkage also incorporates registrations of routine vital events such as birth, death and cancer. The study originally included approximately 500,000 individuals drawn from the 1971Census. With the addition of new members the sample is now about 850,000 individuals, and the follow-up period for inclusion of events is currently twenty-five years. The LS is confidential, a strict 'safe environment' is maintained around it, and access is very strictly controlled.

Figure one outlines the life history of a typical (if hypothetical) LS member. Born in 1955, this member contributed to all three censuses included in the study. She had two children by her mid-twenties, migrated shortly after the 1981 Census, and returned about five years later. She suffered the death of a spouse around 1990, registered with cancer in 1993, and finally died two years later.

The drawbacks of this approach are well known. It takes time to build up sufficient events, and the computer analyses of large datasets can be both time-consuming and frustrating. With SMARTIE we directly address this problem. It is fast, easy to use, flexible and ideally suited to standard repeated analyses of very large studies. The name itself is an acronym and, like many acronyms, its literal meaning is now

[1] Michael Rosato. Longitudinal Study Unit, Room B7/10, Office for National Statistics, 1 Drummond Gate, London SW1V2QQ.

lost. It was commissioned by the team which manages the LS, and while written to be generally applicable software, the specific requirements of handling LS data was a fundamental design concern.

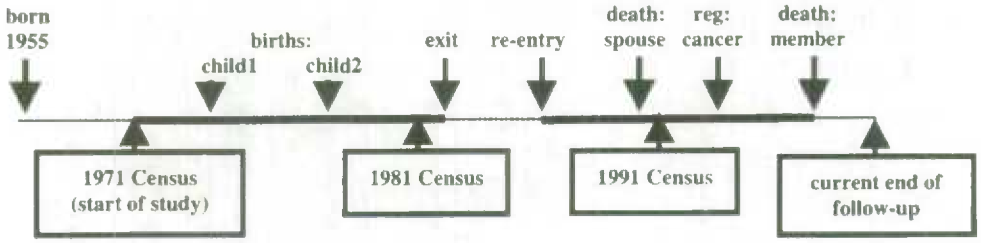


figure one: example of life history of hypothetical longitudinal study member

2. ORGANISATION OF THE PROGRAM

SMARTIE has two distinct parts, as shown in figure two. The first is a screen driven preparation phase, which cleans and re-organises a dataset for analysis. Here a formal period of follow-up is allocated, and all event occurrences are linked with corresponding dates and tested to ensure they fall within this. These are then tested for internal consistency. Exits from and re-entries to surveillance are linked with the events of interest, temporally ordered and tested for feasibility. Relevant ICD revisions are applied to terminal events (for example death, cancer registration) based on date of occurrence. Output from this is a SAS dataset that therefore comprises ordered 'life histories' of study members (as in figure one). A single study dataset can spawn a number of analyses.

In this paper we concentrate on the analysis process, outlined in the lower part of figure two. The user prepares a script (A) of SMARTIE commands. This is input to a code generator (B), which outputs a SAS program encapsulating the required analysis (C). This is then submitted as a SAS job. It calls on a library of pre-written macros (D), links the 'valid' data (E) and a set of standard rates (F), generates the required statistics and, finally, outputs results (G). This generated SAS program can remain hidden (it is possible to use SMARTIE without ever seeing this code).

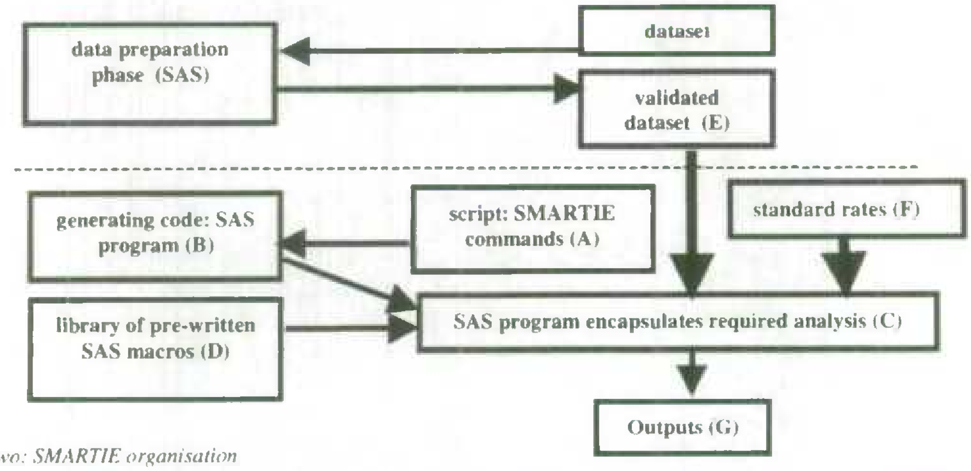

data preparation phase (SAS)
dataset
validated dataset (E)
generating code: SAS program (B)
script: SMARTIE commands (A)
standard rates (F)
library of pre-written SAS macros (D)
SAS program encapsulates required analysis (C)
Outputs (G)

figure two: SMARTIE organisation

3. SMARTIE COMMANDS

This comprises twenty commands, the detail of which is presented in the SMARTIE documentation. They were designed to be easy to use and provide facilities to customise outputs, link datasets for analysis, include sets of standard rates, restrict data before analysis, and re-code variables within an analysis. In these proceedings we concentrate on the overall effect of the more important of these.

The first enables the user to generate summary variables, specifically to flexibly define series of both *terminal events* and *population sub-groups* of interest to a particular analysis. One example of each is given below. They comprise a command header, followed by a list of one or more SAS conditional statements, each of which is preceded by a title naming the defined group. The list terminates with the statement 'finish'. There are constraints to the form of this list: it is currently limited to a maximum of ninety-nine groups; and the individual SAS conditions must be syntactically correct. However, in practice, these are not a barrier to use. There is no requirement for logical exclusivity in these conditions so individuals can be classified to more than one group in any analysis.

```
cause group section
  group 1: lung cancer
  if caused=162 then csgp=1;

  group two: IHD
  if (caused in (410,411, 412, 413, 414))
     then csgp=2;
finish
```

```
tabulation group section
   group one: social class=professional
   if (socl=0) then tbgp=1;

   group 2:  social class=non-manual
   if ((socl ge 0) and (socl le 2)) then tbgp=2;
finish
```

In using the person years at risk method to derive estimates of a population at risk the most basic structure is the *age group-time period* matrix (Lexis) through which this population ages. In SMARTIE all risk is accumulated in days, and the program allows two possible methods of traversing this matrix. Figure three shows analysis in terms of *age at event*: study members are aged as they pass through time. In this case passing time is measured in terms of calendar periods.

Figure four, on the other hand, shows traversal in terms of *age at entry*. In this the age of the individual is held constant as at entry, and they are aged through the time dimension only. Here passing time is measured in terms of the length of survival from first entry to risk. In analysis based on *age at entry* an additional dimension, isolating the calendar period of *entry to risk*, is required. This is necessary because the definition of the risk period dimension used takes no account of calendar period. Its inclusion therefore allows adjustment for calendar period of entry.

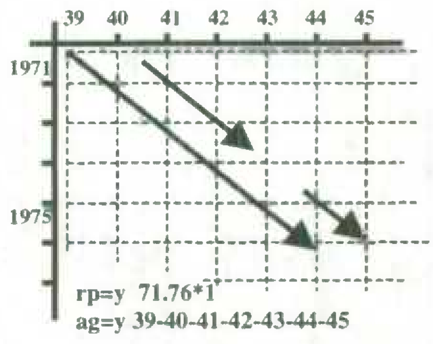


figure three: lexis, showing analysis by age at death

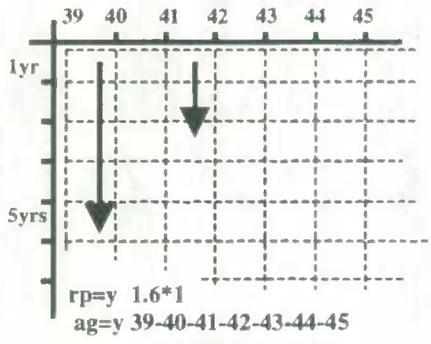


figure four: lexis, showing analysis by age at entry

The unbroken line in Figure three shows a person entering the study at the beginning of 1971 aged exactly 39 years, adding 365 days to each cell crossed (366 days for leap years). This individual finally exits at the end of 1975 aged 43 years, after contributing 364 days to the final cell. The second broken line shows a more complicated scenario: the individual enters the study aged 39 in 1970, contributes risk as before, but exits the study after three years, returning sometime in 1973. SMARTIE can handle multiple exits from and re-entries to risk. The terminal events are similarly handled, accumulated into the cells in which they occur. These two sets of information (estimates of the population at risk and counts of the events) form the base numbers for computing the rates. All outputs are therefore automatically adjusted for age and time period.

The two sets of commands defining these matrices are specified as in the examples below:

```
age groups=y   0.99*5                                age groups=y   0-5-10-15-2-25-30-35-40-45-50-55
risk periods=y  1971-1972-1973-1974-1975-1976        risk periods=y  0-1-2-3-4-5-10-15-20
                                                     entry periods=y 1971-1977*2
```

They take one of two forms. The first of these is a list of contiguous points specifying the boundaries of the age groups or time periods required (for example, between the exact year 1971, 1972 and so on). If ranges are regular however, a simpler statement of the form 'from x to y in units of z' (or, from exact age zero years to exact age 99 years of age in groups of five years) can be used.

These five elements (populations sub-groups, terminal events, and the components of the age-time matrix) form the major dimensions of any analysis. Results are cause-specific, and output as wafers defined by the age-time matrix. Complete sets of cause-specific results are produced for each of the defined population sub-groups. There are two forms of *output*. The first of these is a tabular report comprising selected groupings drawn from the data elements computed in the analysis. These items include observed events, person years of risk, rates, expected events, indirectly standardised ratios (e.g. SMR, SIR), and associated confidence intervals. These tabular outputs can become quite unwieldy, and while cause groups and tabulation groups can be aggregated within the commands themselves, this is not possible for the age group and time period dimensions. However, two sub-totalling commands are available to allow the user to collapse the time-period matrix and re-aggregate the results: expected values are computed using the original age group and time period specifications and then aggregated. The observed events are aggregated, and standard rates likewise re-generated. The re-computed standardised ratios over these sub-totals are therefore output as adjusted for the selected age group and time period dimensions.

A secondary form is the output of a standard rectangular dataset of frequencies, comprising a series of records, each indicating a cell of the tabular report, and its associated observed events and person days of risk. These can then be ported to other packages for further analysis. In this situation, SMARTIE also produces the necessary control files to facilitate this with minimal effort.

The general form of this command is a list of data items required for output, as shown in the following;

```
outputs=T (ox, oxsul, oxsul+a+t, p)
      ou=D (st)
outputs=T (ox, oxsul, oxsul+a+t, p)    d(st)
```

The first example, T(..), specifies that four tabular reports are required. The first comprises observed and expected results; the second reports observed, expected, standardised ratios and associated upper and lower confidence intervals; the third repeats this but also includes row and column sub-totals over each of the wafers; and the last returns the set of person years of risk. The user can therefore develop standard outputs to suit individual needs. The second example, D(..), requires that a standard dataset be produced (in this example, for further analyses in Stata). The third simply extends this idea and concatenates both into a single command.

Formally, the outputs can be specified as follows: if the accumulated observed events (**o**) in population group **g** at age group **a** in time period **t** are represented by **d(o, g, a, t)** and the person years of risk by **p(g, a, t)** then the cause specific rates **r** can be given as

$$r(o, g, a, t)=d(o, g, a, t)/p(g, a, t).$$

The computation of expected events (e) can be determined as

$$e(o, g, a, t)=p(g, a, t) * [d(o, s, a, t)/p(s, a, t)]$$

where **s** is a standard group with which the comparison of rates is to be made.
Finally, the standardised ratio (**sr**) can be specified as

$$sr\ (o, g, a, t)=[d(o, g, a, t)/e(o, g, a, t)] * 100.$$

Ninety five per cent confidence intervals for the standardised ratios based on **d** deaths are derived using the following formulae

$L=100 * EL/e$, and
$U=100 * EU/e$

where **e** is the expected events for that cell,
$EU=1.94 + d - 1.96 \sqrt{(d + 0.96)}$, and
$EL=0.96 + d - 1.9602 \sqrt{(d\ )}$, if the number of deaths is less than 900, or
$0.962 + d - 1.96 \sqrt{(d + 0.11)}$ if the number of deaths is greater than 900.

4. EXAMPLES

The use of SMARTIE is comprehensively detailed elsewhere (Rosato et al). The examples below draw on recent analyses using LS data, and examine mortality patterns by the regions of England and Wales. Using standardused mortality ratios for the period 1988-1994 were produced in SMARTIE and the results input to a GIS system. Two cohorts were traced. Those present at the 1971 census were classified to their region of residence in 1971, and those present at the 1981 census to their region of residence at the 1981 census. Figure five shows the results: the darker the shading the higher the mortality. Death rates were higher in the north of the country than in the south, regardless of which census the region of usual residence was measured. These findings confirm the existence of a persistent north/south health divide in British society.

figure five: mortality of all men aged 40-64 by standard region of England and Wales, all causes, 1988-1994

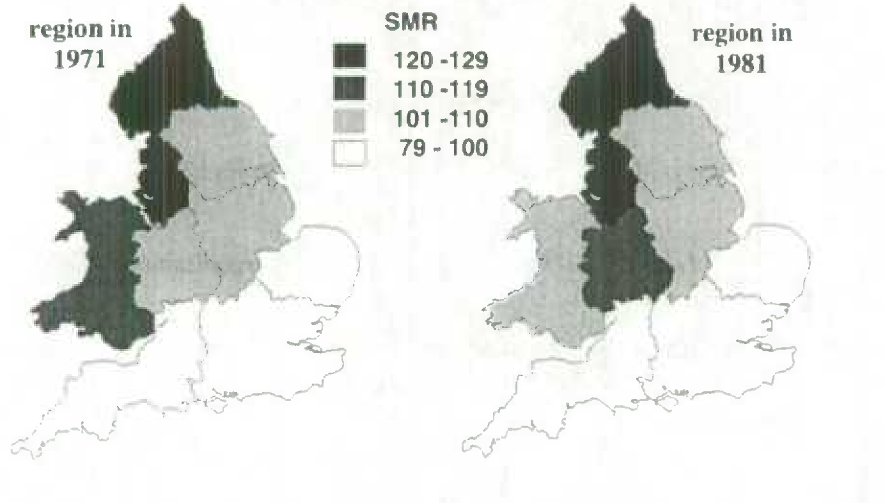
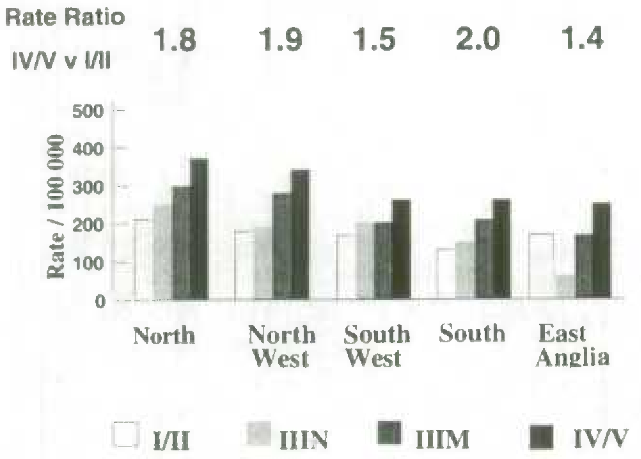


Figure six extends this analysis to examine the effects social class within selected regions. In this case SMARTIE data was ported to Stata and directly standardised rates for social class by standard region The same regional pattern is apparent, with higher mortality in the north of the country than in the south. Concentrating on the social class variation within each region, we can see that the death rates rise incrementally, uniformly lowest in the professional classes (I/II), and highest in the partly skilled and unskilled. The death rates in every class in the two southern regions are lower than the same class in the northern regions. The interesting figures are at the top of the slide. These measure the excess in the death rates of men in the unskilled manual classes when compared with those in the professional classes. In the northern regions death rates for men in classes IV/V were 80% to 90% higher than those of men in classes I/II. In the southern regions, however, with overall lower mortality, this disparity between classes is at least as great as in the north. Lower mortality levels therefore do not necessarily lessen inequality.

figure six: selected standard regions of England&Wales, by social class, death rates 1988-1994



SMARTIE is now a core element in the analysis of LS data. We welcome enquiries from researchers on both the ONS Longitudinal Study and the SMARTIE program. The unit can be contacted using the address given for the author.

REFERENCES

Goldblatt,P (1990). *Longitudinal study: Mortality and social organisation.* Series LS no. 6 London HMSO.

Hattersley,L and Creeser,R. (1995). *Longitudinal study 1971-1991. History, organisation and quality of data.* Series LS no. 7 London HMSO

Rosato,M, Harding,S, McVey.E and Brown.J (1998). *Research implications of improvements in access to the ONS Longitudinal study.* Population Trends 91 Spring 1998.

SESSION 2

MULTI-LEVEL MODELLING

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

MULTILEVEL MODELS FOR REPEATED BINARY OUTCOMES: ATTITUDES AND VOTE OVER THE ELECTORAL CYCLE

Yang, M[1]., Heath, A[2]. and Goldstein, H[1].

ABSTRACT

Models for fitting longitudinal binary responses are explored using a panel study of voting intentions. A standard repeated measures multilevel logistic model is shown inadequate due to the presence of a substantial proportion of respondents who maintain a constant response over time. A multivariate binary response model is shown a better fit to the data.

KEYWORDS: Longitudinal binary data; multivariate multilevel model; political attitudes; voting.

1. INTRODUCTION

The electoral cycle has become an established feature of voting behaviour, both in Britain and in other European countries. After an initial 'honeymoon' between the new government and the electorate, disillusion often sets in and government popularity, whether measured by opinion polls, by-elections or midterm elections such as the European and local elections - tends to decline. In most cases, there is then some recovery in the government's standing in the run-up to the next general election (Miller, Tagg and Britto, 1986; Miller and Mackie, 1973; Reif, 1984; Stray and Silver, 1983)

There are various possible explanations for this pattern. One possibility is that voters make their mid-term decisions on rather different criteria from those they use at a general election. Thus in the mid-term, votes at a by-election or at the European election are unlikely to lead to a change in government. These occasions may thus be used by some voters to communicate their dissatisfaction rather than to change the government. This point may hold with even more force for mid-term opinion polls. (Miller and Mackie, 1973, pp.265-6).

Gelman and King (1993) have provided a more detailed theory about the way in which the opinion poll series changes in meaning as time passes. They suggest that, at the start of the campaign voters' responses are thus based on unenlightened preferences, using on whatever information they happen to have to hand about the candidates, and by polling day are able to base their decisions on 'fundamental variables'. That is, the voters learn how the candidates' policies relate to their own ideologies. Fundamental variables such as the voters' ideologies thus come to acquire greater weight as the campaign progresses.

While there are important institutional differences between the American Presidential campaigns studied by Gelman and King and the British party campaigns, similar processes may be at work here. Our hypothesis, then, is that variables such as the voters' ideologies will have relatively greater weight on their actual voting decisions in general elections than they do on decisions in mid-term elections or on vote intentions conveyed to opinion pollsters. The latter, we suspect, will be more influenced by the free information which the voters have at hand from the mass media about current political stories and events.
Gelman and King use a series of independent random samples conducted at different stages of the campaign in order to test their hypotheses. A more efficient method, however, for understanding change in voters' behaviour is to use a panel study, with repeated observations on the same respondents. In the present paper

[1] Institute of Education, University of London
[2] Nuffield College, Oxford

we use a three-wave panel study covering a complete electoral cycle from 1983 to 1987 to illustrate the modelling procedures. There are three important features of the structure of the two sets of data: (i) a hierarchical structure with voters nested within constituency and years nested within a voter; (ii) repeated dependent binary outcomes (vote or vote intention); (iii) time-dependent covariates representing voters' ideologies and perceptions of the parties and their leaders.

To tackle the dependency problem, the *arbitrary multinomial model* may be considered (Cox, 1972). It fits k binary outcomes by considering the 2^k possible categories, of which $2^k - 1$ of them can then be fitted in the model with a multinomial error distribution. In our case $k = 3$ (years) and the bottom level of year will no longer exist after this reformulation. This model cannot easily accommodate time-dependent covariates which are one level lower than the new multinomial response, and as Cox points out this model gives little insight into the structure of the data (Cox, 1972: p115).

Along the same lines Zeger, Liang and Self (1985) proposed a model with a first-order auto-correlation structure over time to take up the dependence and included time-independent covariates. Consistent estimates based on the full EM procedure were derived. Although this approach can be extended to accommodate time-dependent covariates with some modification to their first-order Markov model, it does not deal with effects above the subject level, in our case the constituency effects.

To take into account the clustering and to model the contextual effects found in the data, Goldstein (1986) proposed the multilevel model using iterative generalised least squares (IGLS) estimation. Under normality this leads to ML estimates. For repeated responses over occasions the model can be extended naturally by adding an additional level at the bottom of the data structure, giving three levels in all. To extend this method to the case of a repeated binary response variable, we may use a generalised linear model formulation (Goldstein & Rasbash, 1996). At the voter level, we can consider modelling the probability of a positive response as a smooth, for example polynomial, function of time. Another approach is to use a multilevel multivariate logistic model. Like any multivariate model, the dependence between the responses can be modelled by the covariance structure at the individual level, in this case the *biserial covariance* (Goldstein, 1995).

In this paper we examine two models: a standard three-level repeated measures logistic model and a multivariate multilevel logistic model. We compare the results from these with those obtained by applying separate two-level models to each round of the panel. For binary responses, we use the procedures known as PQL quasi-likelihood estimation with a second-order Taylor series approximation (Breslow & Clayton, 1993; Goldstein, 1991, Goldstein & Rasbash, 1996) which have been incorporated into the program *MLn/MLwiN* (Rasbash & Woodhouse, 1995).

2. THE DATA FROM THE 1983-6-7 PANEL

In this paper we use the 1983 - 86 - 87 British Election Panel Study. Respondents were interviewed on three occasions: first in 1983 immediately after the general election, second in the autumn of 1986, and third in 1987 immediately after the general election of that year. The panel thus covers a complete electoral cycle, with one round of interviews taking place between the two general elections, The 1983 British Election Survey was a clustered random sample with 3955 respondents interviewed in 250 constituencies (for full details see Heath *et al.*, 1985, Appendix I). For cost reasons, the panel was based on a subset of these respondents. Respondents in 112 of the original constituencies were selected to provide the panel. Of the 1698 respondents selected in this way a total of 869 (52%) completed all three waves. The two main sources of non-response were the difficulty of locating respondents who had moved between 1983 and 1987 and the refusal of located respondents to participate in an interview. (Heath *et al* 1991, Appendix II). The numbers of respondents used are 1526, 1008 and 823 respectively in the three years. There is considerable dependence between responses in the three rounds of the panel. Individuals voting for Conservative on all three occasions made up 27.7%. Those voting for or against Conservative Party both in 1983 and 1986 made up 74.3%. Similarly the percentages for the same votes are 63.3% in 1983 and 1987, 66.2% in 1986 and 1987. Our response variable is vote or vote intention. To simplify the treatment, we shall in our analysis dichotomise the response, contrasting Conservative votes with votes for all other parties.

As measures of voters' fundamental values related to the candidate's policies on nuclear defence, unemployment (versus inflation), tax cuts (versus government spending) and privatisation (versus nationalisation) respectively. We use four scales, left/right wing variables $x_1 - x_4$ on a tweenty-one point sclae. We have two groups of variables to reflect the more topical 'headline' themes to which the voters will have been exposed over the course of the electoral cycle, namely evaluations of the political leaders (variables $x_5 - x_6$) and party images (variables $x_7 - x_{10}$). Evaluations of the party leaders (Mrs Thatcher and Mr Foot/Kninock) were asked on four-point scales running from very effective to ineffective. Voters' images of the two major parties (Conservative and Labour) were on two sets of dicotomise questions: "is the party an extreme or otherwise (coded 0 or 1)?" "is the party united or otherwise (coded 0 or 1)?"

Our central hypothesis, then, is that the parameter estimates for variables $x_1 - x_4$ would be relatively larger in the general election years of 1983 and 1987 while the estimates for variables $x_5 - x_{10}$ would have relatively greater impact on vote intention in the mid-term election year of 1986.

3. MODELS

Model 1: three-level repeated measures model. We treat year as the repetition at level 1 (indicated by t) nested within individuals (indicated by i), while individuals are nested within constituency j. Let z_t be the vector of indicator variables for $t = 1,2,3$ or 1983, 1986 and 1987 respectively, namely, $z_{1ij} = 1$ if t=1983, zero otherwise; $z_{2ij} = 1$ for t=1986 and zero otherwise; $z_{3ij} = 1$ for t=1987 and zero otherwise. Since year is now level 1 our notation reflects this with t being the index for the first subscript. We shall use s_{ij} to denote the measurement of time (1,2,3) as a continuous score. We can write a model as follows for the probability of a positive response π_{tij}

$$\log it(\pi_{tij}) = \sum_{t=1}^{3} \beta_{0,t} z_{tij} + \sum_{t=1}^{3}\sum_{h=1}^{10} \beta_{h,t} z_{tij} x_{h,tij} + \sum_{t=1}^{3} v_{tj} z_{tij} + u_{ij} \qquad (1)$$

$$u_{ij} = u_{0ij} + u_{1ij} s_{ij}, \qquad v_{tj} \sim N(0,\Omega_v), \qquad u_{ij} \sim N(0,\Omega_u)$$

$$\Omega_v = \begin{pmatrix} \sigma^2_{v1} & & \\ \sigma_{v12} & \sigma^2_{v2} & \\ \sigma_{v13} & \sigma_{v23} & \sigma^2_{v3} \end{pmatrix}, \qquad \Omega_u = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$$

where v_{tj} are the residual terms at constituency level associated with the intercepts of years. Thus, for the j^{th} constituency the marginal population mean for Conservative voting for 1983, 1986 and 1987 respectively on the logit scale is given by $\beta_{0,1} + v_{1j}$, $\beta_{0,2} + v_{2j}$ and $\beta_{0,3} + v_{3j}$ respectively. The term u_{ij} estimates the departure on log odds of the i^{th} individual voter in the j^{th} constituency from the overall log odds, here assumed to be a linear function of time. Thus, the variance-covariance Ω_u is modeled as a quadratic function of the continuous time S. The response is $y_{tij} \sim Bin(1, \pi_{tij})$ with independent distribution error at level 1 as $\pi_{tij}(1-\pi_{tij})$ across all the three years, and as same as before we can model extra-binomial variation by allowing a multiplier to the level error distribution.

Model 2: a multilevel multivariate logistic model. Using the same notation as in the case of the repeated measures model (1), a general multivariate logistic model for our data may be written

$$\log it(\pi_{tij}) = \sum_{t=1}^{3} \beta_{0,t} z_{tij} + \sum_{t=1}^{3}\sum_{h=1}^{10} \beta_{h,t} z_{tij} x_{h,tij} + \sum_{t=1}^{3} v_{tj} z_{tij} \qquad (2)$$

We make the same assumptions as for model (1), except that there is no level 1 variation, but at level 2 we allow the binomial variates to covary. Namely the following variance-covariance structure among individual voters is assumed for fitting this model:

$$\Omega_2 = \begin{pmatrix} \pi_{1ij}(1-\pi_{1ij}) & & \\ \sqrt{\pi_{1ij}\,\pi_{2ij}(1-\pi_{1ij})(1-\pi_{2ij})} & \pi_{2ij}(1-\pi_{2ij}) & \\ \sqrt{\pi_{1ij}\,\pi_{3ij}(1-\pi_{1ij})(1-\pi_{3ij})} & \sqrt{\pi_{2ij}\,\pi_{3ij}(1-\pi_{2ij})(1-\pi_{3ij})} & \pi_{3ij}(1-\pi_{3ij}) \end{pmatrix}$$

At this level, therefore, we estimate a covariance structure in which the diagonal terms are the binomial variances and the off -diagonal terms are biserial covariances. We may also, as before, allow three extra-binomial variation parameters one for each of the diagonal terms only, which leaves the off-diagonal terms being estimated freely from the data. When these binomial variances are relaxed, the off-diagonal terms estimate just the biserial correlation coefficients between any two years pairwisely. We can elaborate the model by forming the interaction terms between the explanatory variables and the year indicators to fit the main effects of them in the fixed part according to equation (2). To compare the effects of the same explanatory over the three years, the joint hypothesis tests on $\beta_{h.83} = \beta_{h.86}$ and $\beta_{h.86} = \beta_{h.87}$ using the approximate Wald statistics are carried out.

A two-level logistic model is also fitted to the data by the election year to examine the marginal distribution of each year, ignoring the dependency of outcomes between years.

4. RESULTS

In Table 1 we list the estimates of those models without fitting the effects of the covariates. Comparing to the marginal model, both models (1) and (2) give close estimates to the fixed effects in the models, and the random effects in terms of variances by year at constituency level. However model (1) holds an entirely different assumption at voter level compared to that of model (2), fitting a quadratic function of time span to the voter level residuals. This results a severe under-dispersion at the bottom level of year, and non-Normal distribution to the voter level residuals, indicating a failure in accommodating the dependency of outcomes within voter. Therefore model (1) is inadequate in this form to fit our data.

Model (2) has biserial terms at voter level to take up the dependency of outcomes between years. The estimates of covariances may be termed as biserial correlation coefficients in our case estimated as 0.54, 0.62 and 0.61 respectively, compared to the raw proportions of the same votes 0.74, 0.63 and 0.66. As we expected, this treatment also gives us a binomial variance at the same level for each year between voters with the extra-binomial parameters are estimated all close to one, indicating that these biserial terms have almost fully taken up the dependency of outcomes.

Model (2) then is fully elaborated for fitting the main effects of all covariates in order to test out our assumption about the electoral cycle. Parameter estimates are displayed in Table 2. The joint test is based on two linear contrasts of $\beta_{h.83} = \beta_{h.86}$ and $\beta_{h.86} = \beta_{h.87}$ for testing the equality of impacts between occasions for each variable crossing the three waves, which performs the χ^2 test in the MLn program. Results are in the last column of Table 2.

The random effects across constituencies are rather small with almost zero estimates to 1983 and 1986, from fitting the full model (2). It is not necessary to explore any random slope in this case. The significant results are only found from two variables: score on nationasition vs privatisation x_4 with the expected pattern and image on Labor Party, x_{10} (divided versus united) with a reversed pattern. Overall, there is not strong evidence giving unequivocal support to the substantive theory described in the introduction. Possibly this is because the non-election round of interviews was conducted rather too late in the electoral cycle, being held in the autumn of 1986 less than twelve months before the June 1987 election. By the autumn, the Conservatives had already recovered their popularity in the opinion polls and the panel study did not therefore really capture the phase of mid-term disillusion with the government.

Table 1 Parameter estimates, SE in brackets

Parameter	Model (1)	Model (2)	Marginal model
Fixed effect: Intercept 1983, $\beta_{0,1}$	-0.39 (0.08)	-0.42 (0.08)	-0.40 (0.08)
Intercept 1986, $\beta_{0,2}$	-0.77 (0.07)	-0.81 (0.08)	-0.75 (0.08)
Intercept 1987, $\beta_{0,3}$	-0.30 (0.07)	-0.32 (0.08)	-0.22 (0.08)
Random effect:			
Constituency level: Var(83)	0.35 (0.08)	0.41 (0.10)	0.38 (0.09)
Var(86)	0.20 (0.08)	0.18 (0.09)	0.14 (0.09)
Var(87)	0.19 (0.08)	0.22 (0.09)	0.13 (0.09)
Cov(83,86)	0.29 (0.07)	0.31 (0.08)	N/a
Cov(83,87)	0.28 (0.07)	0.31 (0.08)	N/a
Cov(86,87)	0.20 (0.07)	0.21 (0.08)	N/a
Extra-binomial variance			
Voter level: 1983	σ^2_{u0} : 2.10 (0.21)	0.96 (0.04)	0.96 (0.04)
1986	σ_{u01} : 0.08 (0.05)	1.00 (0.05)	0.97 (0.05)
1987	σ^2_{u1} : 0.00 (0.00)	0.98 (0.05)	0.98 (0.05)
Biserial cov(83,86)	N/a	0.54 (0.03)	N/a
Biserial cov(83,87)	N/a	0.62 (0.04)	N/a
Biserial cov(86,87)	N/a	0.61 (0.04)	N/a
Variance within voter	0.38 (0.01)	--	--

Table 2. Estimates of the main effects from fitting model (2) and tests for equality over occasions

Parameter	Estimate (SE) 1983	Estimate (SE) 1986	Estimate (SE) 1987	Equality test χ^2_2
β_1	0.08 (0.01)	0.12 (0.02)	0.07 (0.02)	4.77
β_2	0.05 (0.01)	0.03 (0.01)	0.03 (0.02)	2.01
β_3	0.05 (0.01)	0.07 (0.02)	0.08 (0.02)	2.30
β_4	0.09 (0.01)	0.04 (0.02)	0.05 (0.02)	6.55 ($p<0.05$)
β_5	-0.93 (0.12)	-0.92 (0.13)	-1.03 (0.21)	0.25
β_6	0.49 (0.08)	0.46 (0.09)	0.60 (0.10)	1.24
β_7	0.91 (0.12)	0.94 (0.14)	0.95 (0.17)	0.07
β_8	-0.39 (0.12)	-0.40 (0.15)	-0.34 (0.18)	0.08
β_9	-0.55 (0.15)	-0.24 (0.15)	-0.60 (0.24)	2.98
β_{10}	0.16 (0.27)	-0.10 (0.18)	0.57 (0.22)	6.03 ($p<0.05$)

5. CONCLUSION

For analysing the repeated binary data with time dependent covariates as in our paper, the two-level logistic model is shown to be a useful tool to provide basic information on the contextual variation and the possible association of the covariates with the outcome, based on the marginal distribution. However it has drawbacks of lacking efficiency where responses at some occasions are missing and of its failure to model the dependency among the repeated responses. The three-level repeated measures logistic model provides a straightforward and efficient way to model the full data, and enable us to perform the approximate χ^2 test for our centre assumption. However this model fails to accommodate the dependency of outcomes, and violates the basic assumption of independent error distribution in the binomial case.

The multilevel multivariate logistic model assumes binomial error at each occasion with a biserial covariance structure at voter level to take care of the dependence between the repeated outcomes. It has the same advantages as the repeated measures model in terms of the efficiency from pooling all the data in one model. The model's predictions for the overall probability of voting show a reasonable agreement with the raw probabilities. The estimated variance among constituencies for each year is similar to that from the marginal models fitted to each year separately, and the same Binomial assumption holds for the lowest level error distribution by year. The Normality assumption for higher level residuals is adequate. It is also possible to generalise the multivariate model for the general repeated measures case with any number of occasions, but this will involve setting up an explicit model for the autocorrelation structure, and work on this is currently under way. This model can also readily to be extended for study the tactical voting at constituency level.

REFERENCES

Breslow, N.E. & Clayton, D. G. (1993) Approximate inference in generalised linear models. *J. American Statist. Assoc.*, **88**: 9-25.

Cox, D. R. (1972) The analysis of multivariate binary data. *Applied Statistics*, **21**, 113-20.

Gelman, A. and King, G. (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, **23**, 409-51.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* **73**: 43-56.

Goldstein, H. (1995) *Multilevel Statistical Models*. Edward Arnold: London, Halsted Press: New York.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, **159**, 505-513.

Heath, A. F., Jowell, R.M. and Curtice, J.K. (1985) *How Britain Votes*. Oxford: Pergamon.

Heath, A. F., Jowell, R.M., Curtice, J.K., Evans. G., Field, J. and Witherspoon, S. (1991) *Understanding Political Change: The British Voter 1964-1987*. Oxford: Pergamon.

Miller, W.L., Tagg, S. and Britto, K. (1986) Partisanship and party preferences in government and opposition: the mid-term perspective. *Electoral Studies*, **5**, 31-46.

Miller, W.L. and Mackie, T. (1973) The electoral cycle and the asymmetry of government and opposition popularity. *Political Studies*, **21**, 263-79.

Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*, V1.0, Institute of Education, University of London.

Reif, K. (1984) National electoral cycles and European elections, 1979 and 1984. *Electoral Studies*, **3**, 244-55.

Stray, S. and Silver, M. (1983) Government popularity, by-elections and cycles. *Parliamentary Affairs*, **36**, 49-55.

Zeger, S.T., Liang, K.Y. and Self, S.G. (1985) The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, **72**:1, 31-8.

(This work was carried out as part of the Multilevel Models Project funded by the Economic and Social Research Council (UK) under the programme for the Analysis of Large and Complex Datasets.)

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

RANDOM EFFECTS MODELS FOR LONGITUDINAL DATA FROM COMPLEX SAMPLES

Julia L. Bienias,[1] Laurel A. Beckett,[1] and Charles L. Owen[1]

ABSTRACT

Longitudinal studies with repeated observations on individuals permit better characterizations of change and assessment of possible risk factors, but there has been little experience applying sophisticated models for longitudinal data to the complex survey setting. We present results from a comparison of different variance estimation methods for random effects models of change in cognitive function among older adults. The sample design is a stratified sample of people 65 and older, drawn as part of a community-based study designed to examine risk factors for dementia. The model summarizes the population heterogeneity in overall level and rate of change in cognitive function using random effects for intercept and slope. We discuss an unweighted regression including covariates for the stratification variables, a weighted regression, and bootstrapping; we also did preliminary work into using balanced repeated replication and jackknife repeated replication.

KEY WORDS: pseudo maximum likelihood, bootstrapping, resampling, random effects

1. INTRODUCTION

Longitudinal studies are very useful, because they permit better characterizations of change and the assessment of possible risk factors, and complex sample designs are often used for large-scale longitudinal studies for several reasons. First, it is often important to obtain accurate descriptions of change for subgroups such as children or minorities, and thus they may be oversampled. Second, oversampling groups at high risk may give more power for a given cost. Finally, longitudinal studies are typically expensive to carry out; limitations on resources may make it necessary to accomplish multiple goals with a fixed budget, and thus to design an efficient sampling strategy. There has been little experience, however, in adapting the increasingly sophisticated models proposed for longitudinal data to the complex sample setting.

We focus on random effects models of change (Laird & Ware, 1982) in a study of cognitive function in the elderly. Such models can be used to characterize change directly and also summarize the heterogeneity in overall levels and patterns of change. When the sample design has been constructed to overrepresent people with unusually rapid or slow changes, then the analysis must adjust for the sampling design in order to describe the source population. We compare different approaches to parameter estimation and variance estimation. The classical sampling approach is to use pseudo-maximum likelihood estimates obtained by weighting each person's data in the analysis to reflect the sampling, with a sandwich estimator based on a first-order Taylor series expansion to obtain variance estimates (Särndal, Swensson, & Wretman, 1992; Skinner, Holt, & Smith, 1989). Alternatively, we examine a bootstrap approach (Efron, 1982) and using the stratum characteristics as covariates in the model (Korn & Graubard, 1991; also DuMouchel & Duncan, 1983). Previous studies have examined the effects of different approaches to allowing for the effects of complex sampling designs on estimation in linear regression (DuMouchel & Duncan, 1983; Kovar, Rao, & Wu, 1988; Krewski & Rao, 1981) and logistic regression (Korn & Graubard, 1991; Roberts, Rao & Kumar, 1987).

[1] Julia L. Bienias, Rush Institute for Healthy Aging, 1645 W. Jackson Blvd., Suite 675, Chicago, IL 60612, USA; Laurel A. Beckett, Charles L. Owen: same address

2. THE SETTING

East Boston, Massachusetts, a geographically defined, largely working-class urban community, was one of four sites of the Established Populations for Epidemiologic Studies in the Elderly (EPESE). A complete census was carried out in 1982-84. Based on information obtained during an initial in-home interview, a stratified random sample (age group (5) × sex × memory performance (4)) was chosen for yearly follow-up. The sample we model here contains 388 people and 1743 observations (Evans et al., 1989; also Beckett et al., 1992.)

The outcome of interest was a summary measure of cognitive function based on averaging z scores (using population estimates of mean and standard deviation) from eight individual cognitive performance tests. This composite measure had a more Gaussian distribution than the individual tests and minimized the floor and ceiling problems. In addition, it should reduce error variation and effectively extend the range within which individual differences are discriminable. Figure 1 is a plot of each person's set of measurements of this cognitive function summary score. We considered a model using age as the only substantive fixed effect:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon, \tag{1}$$

where $\mathbf{Y}$ is a vector of observations, $\mathbf{X}$ and $\mathbf{Z}$ are known covariate matrices, β is an unobservable vector of fixed effects, and γ is an unobservable vector of random effects. We assume that the parameters of γ are normally distributed person-specific random effects with variance-covariance $\mathbf{G}$, ϵ are normally distributed within-person errors with covariance $\sigma_\epsilon^2\mathbf{I}$, and that γ and ϵ are independent. The entries in $\mathbf{G}$ and the error variance σ_ϵ^2 are additional parameters to be estimated. Thus $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{ZGZ}'+\sigma_\epsilon^2\mathbf{I})$, and we denote the density function by $f(y)$.

3. PARAMETER ESTIMATION

Suppose that the community consists of N individuals for each of whom longitudinal data $\mathbf{Y}$ are generated from a multivariate normal distribution via a random effects model (1). Let $g(\mathbf{y},\theta) = \delta/\delta\theta \log f(\mathbf{y})$, where θ denotes the parameters (β and variance parameters) to be estimated. The function g is an unbiased estimating function, and the maximum likelihood estimates (MLEs) of the parameters in θ are given by the roots of the equations

$$\textstyle\sum_{i=1}^{N} g(\mathbf{y}_i,\theta) = 0 \ . \tag{2}$$

The MLEs of the coefficients β can be obtained as functions of the MLEs of the variance parameters, and both can be calculated by the EM algorithm (Laird & Ware, 1982).

If data were available for the entire community, then the solution of the estimating equations (2) would be an estimate of the superpopulation parameters of the model (1). When the survey community data are not completely known, the solution of (2) defines a finite population parameter for the community (Godambe & Thompson, 1986). Under stratified sampling, taking a weighted mean of the estimating functions for the sampled individuals will give both an estimate of the finite population parameters and a sample-based estimate of the superpopulation parameters. If we denote the inverse of the sampling probabilities under the stratified design by w_{hi} for person i in stratum h, then the pseudo MLEs are given by the roots of:

$$\textstyle\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} g(\mathbf{y}_i,\theta) = 0 \ , \tag{3}$$

These are consistent in three different senses (Särndal, Swensson, & Wretman, 1992). First, if all the people in the community were known, the pseudo MLEs would be the roots of the estimating equations (2), and thus would equal the finite population parameters, or finite population consistency. Second, if the sample size n and the population size N were both allowed to go to infinity, while holding the design fixed, the pseudo MLEs would converge in probability to the finite population parameters. Third, again allowing both n and N to go to infinity, the pseudo MLEs would converge in probability to the superpopulation parameters.

4. MODEL-BASED VARIANCE ESTIMATION

Some authors have suggested that regression coefficients can be estimated without using the sampling weights, provided that the coefficients β, given the predictors $\mathbf{X}$, the random effects γ, and the residuals ϵ do not depend on any variables used in the stratification (DuMouchel & Duncan, 1983; Korn & Graubard, 1991). If, however, the relationship is thought to depend on some stratification variables, one can add these variables as

in important ways, without incurring the penalty of increased standard errors because of adjustment for unequal sampling weights. Interpretation is complicated by the fact that no single, population-wide estimate is provided for the effect of the primary predictors (the so-called census coefficients). Additionally, uncertainty about the population due to clustering or weights is not accommodated.

5. VARIANCE ESTIMATION: THE BOOTSTRAP

An alternative approach to variance estimation is to use bootstrap replication techniques (Efron & Tibshirani, 1995). This and similar approaches have been applied for estimates of means and regression parameters for univariate outcomes (Rao, Wu, & Yue, 1992; Sitter, 1992). Using the bootstrap requires careful specification of the probability model that gives rise to the data. The model has three components: the sampling design, the fixed parameter estimates, and the error structure. Each component needs to be replicated by the bootstrap resampling procedure in such a way that the distribution of the bootstrap parameter estimates mimics the sampling distribution in the original probability model. Our setting consists of a random effects model with unknown coefficients but a normal error structure, a community with people divided up into strata, each person with a vector of outcomes arising from the normal model, and a random sample without replacement in each stratum. Our goal is to estimate the unknown superpopulation parameters, rather than just the parameters that are the roots of equation (3).

We begin by sampling separately from each stratum h, sampling entire cases rather than fixing the predictors $\mathbf{X}$ and resampling residuals from the regression model because the observed predictors $\mathbf{X}$ are typical of the stratum but may vary in the larger community. We resample with replacement a sample of size n_{h-1} rather than n_h, to compensate for the bias in standard error estimates from the association within the original community strata (Davison & Hinkley, 1997); we do not reduce the stratum size further by the sampling fraction, because we want to preserve the uncertainty associated with having the total community generated from a superpopulation model. The weights are adjusted to reflect this smaller sample size, so that the resulting pseudo MLEs will be consistent. The bootstrap replication is then repeated M times, giving estimates θ_m^* at replication m for each parameter. The sample variance v^* can be calculated for each parameter, and a studentized $100(1-\alpha)\%$ interval calculated centered at the pseudo MLE $\hat{\boldsymbol{\theta}}$:

$$\hat{\theta} \pm (v^*)^{1/2} z_{\alpha/2} \ . \tag{4}$$

The coverage probability of this interval (4) depends on: the approximate normality of the sampling distribution of the pseudo MLEs, the centering of the bootstrap distribution at the pseudo MLE, and the accuracy of the sample variance of the bootstrap variance as an estimate of the sampling variance of the pseudo MLEs. The first two assumptions can be checked empirically. If Q-Q plots of the bootstrap pseudo MLE parameter estimates suggest nonnormality, appropriate normalizing transformations of the bootstrap estimates may be suggested by the empirical distribution of variances estimated from each bootstrap replication, plotted against the bootstrap point estimate (Davison & Hinkley, 1997, Ch 5.2). Alternatively, bootstrap percentiles could be used, but would require substantially larger numbers of bootstrap replications (Efron & Tibshirani, 1993). Bias correction approaches can be used if the empirical distribution of the replicate estimates is not centered at the full-sample pseudo MLE (Efron & Tibshirani, 1993).

The assumption that the bootstrap standard deviation is a consistent estimate of the true sampling standard error is more difficult to establish. For estimation of a mean, the bootstrap resampling procedure gives consistent estimates of the standard error (Davison & Hinkley, 1997). Thus, the bootstrap standard deviation of the mean of the estimating equation (3), evaluated at a fixed parameter value, is a consistent estimate of the standard error under sampling from the model, provided the model is correct. Bootstrap estimates of the standard error of coefficients from linear regression have also been shown to be consistent, if the number of parameters is small relative to the number of data points and the resampling is not subject to near-collinearity in the design matrix. These considerations suggest that the bootstrap standard deviations should be consistent estimates of the standard errors of the fixed-effect coefficients in the random effects models, with at most modest adjustment for bias or nonnormality. The estimates of the variance components, however, are more likely to have skewness and bias that would violate the assumptions of and require transformation.

6. RESULTS

The estimated model using the full sample weights is **cog = -0.40 - 0.05 age - 0.04 lag - 0.003 age lag +** γ_0 + γ_1, where "cog" is the summary measure of cognitive function defined in Section 2, "age" is taken as the difference in age from 80, "lag" is the time between the follow-up interview and baseline, γ_0 is a random variable for the individual person's intercept for the trajectory over time, and γ_1 is a random variable for the individual person's slope. We also have

$$\mathbf{G} = \begin{bmatrix} 0.4564 & 0.0168 \\ 0.0168 & 0.0046 \end{bmatrix} ; \; \hat{\sigma}^2 = 0.6081 \; .$$

Figure 2 shows the predicted trajectories for each person, based on the model. Figure 3 shows the same predicted values, but with the thickness of the line proportional to the sampling weight, to emphasize the unequal probabilities of selection and their impact on the estimation; more of the community have "flat" lines than appear in Figure 2.

Fitting a model without weights but including terms for sex, "poor vs. good" and "either intermediate vs. good" memory groups yields the following fit conditional on those variables as fixed effects alone and interacted with "lag:" **cog = -0.07 - 0.05 age - 0.03 lag - 0.004 age lag +** γ_0 + γ_1; **G**(1,1) = 0.3364; **G**(1,2) = 0.0193; **G**(2,2) = 0.0100; $\hat{\sigma}^2 = 0.1085$.

The following table summarizes the variance results. The column "Weighted SRS" contains estimates from treating the data as a simple random sample with weights (from SAS PROC MIXED®). "Design Effects" contain the ratio of the variance estimates to those from the Weighted SRS fit. For comparison, if the sampling weights are ignored, the estimates are: Intercept: -0.40; Age: -0.05; Lag: -0.05; Age by Lag: -0.0003; **G**(1,1) = 0.4628; **G**(1,2) = 0.0315; **G**(2,2) = 0.0111; $\hat{\sigma}^2 = 0.1083$. That is, the estimated coefficients for the fixed effects are nearly identical. For the random components, the estimate of the variance for the intercept effect is very similar to the estimates that take into account the sample design. However, the estimate of the variance for the slope effect is more than twice the size when the design is ignored and the estimated covariance between the intercept and slope components is also about twice the size. Finally, the estimated residual variance is about 82% smaller when the design is ignored (0.1083 vs. 0.6081).

7. DISCUSSION

but may overestimate the residual variance and the variances associated with the random effects, particularly the variance for the slope parameter. This would be expected, in turn, to lead to an increased chance of concluding there is heterogeneity in the data even if there were none.

We additionally wish to consider other replicate methods, namely jackknife repeated replication and balanced repeated replication (BRR; also called balanced half sample) (Wolter, 1985). In particular, we next wish to address the issue of design consistency of these approaches, and consider under what circumstances the various approaches to variance estimation may differ. Other authors (e.g., Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998) have discussed hierarchical models in the context of a hierarchical sample design. Our setting differs from these in that we have one mechanism assumed to be generating the data which is separate from the mechanism used in sampling, as opposed to the sample design being hierarchical itself. In some preliminary implementations of the jackknife and BRR, we saw a wide range of design effects, particularly for the variance estimates of the variance parameters. This deserves further attention to both the theory and to understanding circumstances under which these methods may fail, as well as discussion of the relation to inference (e.g., see Pfeffermann, 1993).

		Standard Error Estimates	Design Effects	
			Variance Estimation Method	
		Weighted SRS	Model-Based	Bootstrap
Fixed Effects	Intercept	0.0380	2.66	0.81
	Age at Start	0.0052	0.78	0.59
	Lag	0.0076	4.11	1.73
	Age * Lag	0.0010	1.21	1.00
Random Effects	Intercept	0.0409	0.52	1.49
	Slope	0.0011	3.31	0.83
	Cov (Intercept, Slope)	0.0057	0.93	1.51
Residual	σ^2	0.0258	0.04	4.88

REFERENCES

Beckett, L. A., Scherr, P. A., and Evans, D. A. (1992). Population prevalence estimates from complex samples. *Journal of Clinical Epidemiology*, **45**, 393-402.

Davison, A.C., and Hinkley, D. V. *Bootstrap Methods and Their Application.* Cambridge, UK: Cambridge Press, 1997.

DuMouchel, W. H., and Duncan, G. J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, **78**, 535-543.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* NY: Chapman Hall.

Evans, D. A., Funkenstein, H. H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., Hebert, L. E., Hennekens, C. H., and Taylor, J. O. (1989). Prevalence of Alzheimer's Disease in a community population of older persons: Higher than previously reported. *Journal of the American Medical Association*, **262**, 2551-2556.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, **54**, 127-138.

Korn, E. L., and Graubard, B. I. (1991). Epidemiologic studies utilizing surveys: Accouting for the sampling design. *American Journal of Public Health*, **81**, 1166-1173.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in surveyestimates. *Canadian Journal of Statistics*, **16**, 25-45.

Krewski, D., and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, **9**, 1010-1019.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 317-337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, Part 1, 23-40.

Rao, J. N. K., Wu, C. F.J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209-217.

Roberts, G., Rao, J. N. K., and Kumar, S. (1987). Logistic regession analysis of sample survey data. *Biometrika*, **74**, 1-12.

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

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, **83**, 231-241.

Skinner, C. J., Holt, D., and Smith, T. M. F. (Eds.) (1989). *Analysis of Complex Samples*. NY: Wiley.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. NY: Springer-Verlag.

Acknowledgment: This research was supported by grants (AG11101, AG10161) from the National Institute on Aging, National Institutes of Health, USA.

Figure 1. Raw Data

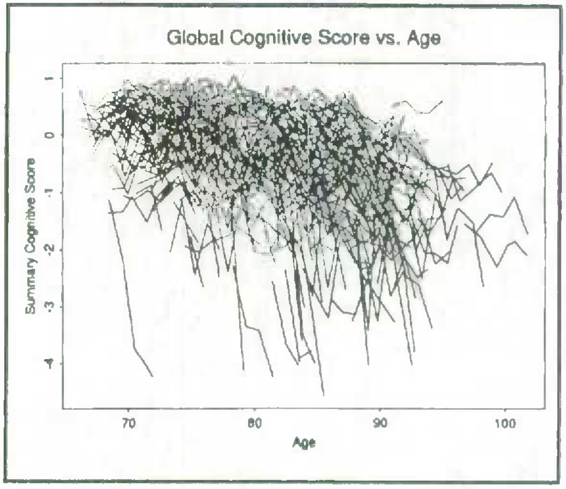


Figure 2. Fitted Model

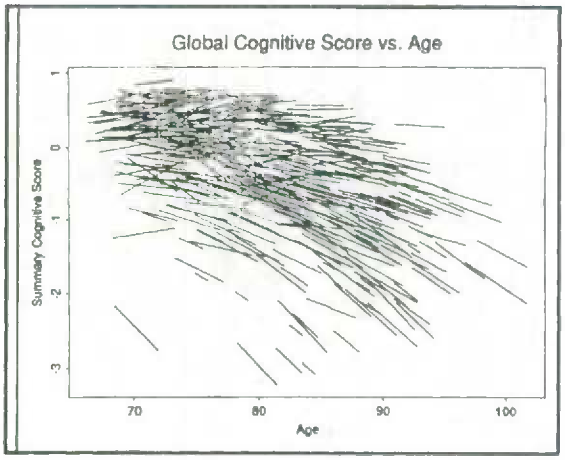
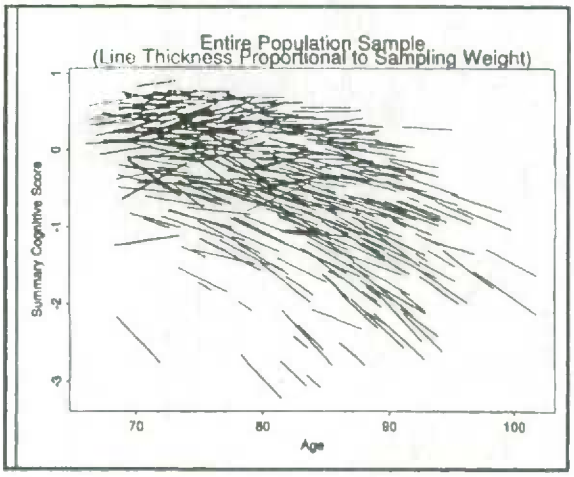


Figure 3. Model Showing Weights



Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

MULTILEVEL MODELING OF COMPLEX DATA STRUCTURES WITH MULTIPLE UNIT MEMBERSHIP AND MISSING UNIT IDENTIFICATIONS

Harvey Goldstein[1], Peter W. Hill[2], and Jon Rasbash[3]

ABSTRACT

This paper presents a method for handling longitudinal data in which individuals belong to more than one unit at a higher level, and also where there is missing information on the identification of the units to which they belong. In education, for example, a student might be classified as belonging sequentially to a particular combination of primary and secondary school, but for some students, the identity of either the primary or secondary school may be unknown. Likewise, in a longitudinal study, students may change school or class from one period to the next, so 'belonging' to more than one higher level unit. The procedures used to model these structures are extensions of a random effects cross-classified multilevel model.

1. INTRODUCTION

Across a wide range of disciplines, it is commonly the case that data have a complex hierarchical structure. Subjects may be clustered not only into hierarchically ordered units (e.g., students nested within classes, within schools), but may also belong to more than one unit at a given level of a hierarchy. In this paper we use educational data to illustrate our ideas. For example, a student might be classified as belonging sequentially to a particular combination of primary school and secondary school, in which case the student will be identified by a cross classification of primary schools and secondary schools. Alternatively, a particular student may spend a proportion of time in one school and the remaining proportion in another school. In this case, the student has multiple membership of units at a given level of clustering.

Goldstein (1987) and Raudenbush (1993) present the general structure of a model for handling complex hierarchical structuring with random cross classifications. For example, assuming that we wish to model the achievement of students taking into account both the primary and the secondary school attended by each student, then we have a cross classified structure, which can be modeled as follows:

$$y_{i(j_1j_2)} = (X\beta)_{i(j_1j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1j_2)}, \qquad [1]$$
$$j_1 = 1,\dots J_1, \quad j_2 = 1,\dots J_2\,, \quad i = 1,\dots N$$

in which the score of student i, belonging to the combination of primary school j_1 and secondary school j_2, is predicted by a set of fixed coefficients $(X\beta)_{i(j_1,j_2)}$. The random part of the model is given by two level 2 residual terms, one for the primary school attended by the student (u_{j_1}) and one for the secondary school attended (u_{j_2}); and the usual level 1 residual term for each student. We note that the

[1] Harvey Goldstein, Institute of Education, University of London
[2] Peter W. Hill, Centre for Applied Educational Research The University of Melbourne
[3] Jon Rasbash, Institute of Education, University of London

latter may be further modeled to produce complex level 1 variation (Goldstein, 1995, Chapter 3).

Rasbash and Goldstein (1994) give details of a method for estimating cross-classified models using a simple hierarchical formulation and a set of (0,1) dummy variables for each unit of one of the cross-classified random variables. The dummy variables are introduced as explanatory variables into the random part of the model and the variances of the random coefficients of these dummy variables are constrained to be equal, thus providing an estimate of the between unit variance. The method can be used to analyze a wide variety of models with the only serious limitation being the computational demands generated by models with a large number of cells in the cross classification. Examples of several kinds of frequently occurring situations in which it may be appropriate to use a multilevel random cross classification model are given by Goldstein (1995).

This paper sets out extensions of the random cross classification to the case of multiple unit membership and also to situations in which there is incomplete information on the units to which students belong. These procedures are all implemented in the software package *MLwiN* (Rasbash et al., 1998). We first consider the case where students belong to more than one secondary school.

2. THE MULTIPLE MEMBERSHIP MODEL

Suppose that we know, for each individual, the weight π_{ij_2}, associated with the j_2-th secondary school for student i (for example the proportion of time spent in that school) with $\sum_{j_2=1}^{J_2} \pi_{ij_2} = 1$. These weights, for example, may be proportional to the length of time a student is in a particular school during the course of a longitudinal study. Note that we allow the possibility that for some (perhaps most) students only one school is involved so that one of these probabilities is one and the remainder are zero. We can now rewrite [1] as follows:

$$y_{i(j_1,j_2)} = (X\beta)_{i(j_1,j_2)} + u_{j_1}^{(1)} + \sum_{j_2} u_{j_2}^{(2)} \pi_{ij_2} + e_{i(j_1,j_2)}$$

$$\sum_{j_2} u_{j_2}^{(2)} \pi_{ij_2} = \pi_i u^{(2)}$$

$$u^{(2)^T} = \{u_1^{(2)},.....u_{J_2}^{(2)}\} \qquad [2]$$

$$\pi = \{\pi_1,...\pi_{J_2}\}$$

$$\pi_{j_2}^T = \{\pi_{1j_2},...\pi_{Nj_2}\}$$

where N is the total number of students, $\mathrm{u}^{(2)}$ is the $J_2 \times 1$ vector of secondary school effects and j_1 indexes the primary school. Thus [2] is a 2-level model where the level 2 variation among secondary schools is modeled using the J_2 sets of weights for subject i ($\pi_1,...\pi_{J_2}$) as explanatory variables, with $\pi_{\mathrm{j}2}$ the N x 1 vector of student weights for the j_2th secondary school. We have

$$\mathrm{var}(\mathrm{u}_{\mathrm{j}_2}^{(2)}) = \sigma_{\mathrm{u}2}^2, \quad \mathrm{cov}(\mathrm{u}_{\mathrm{j}_1}^{(1)} \mathrm{u}_{\mathrm{j}_2}^{(2)}) = 0$$

$$\mathrm{var}(\sum_{\mathrm{j}_2} \mathrm{u}_{\mathrm{j}_2}^{(2)} \pi_{\mathrm{ij}_2}) = \sigma_{\mathrm{u}2}^2 \sum_{\mathrm{j}_2} \pi_{\mathrm{ij}_2}^2 \qquad [3]$$

We note also that for any fixed or random part explanatory variables defined at the school level the value of this variable will be that of the school to which a student belongs at the occasion of measurement.

3. THE MISSING IDENTIFICATION MODEL

Just as proportions or probabilities other than (0,1) are used to indicate multiple unit membership, this also provides the basis of a method for handling missing unit identification information in complex hierarchical models. For example, in the case of model (1) we may have complete information about the secondary school attended by each student, but incomplete information about the primary school from which they came. Nevertheless, knowing the locality of the secondary school they currently attend, we may have a reasonable basis for assigning probabilities for attendance at one or more identifiable primary schools.

Suppose now that π_{ij_1} is the weight associated with membership of primary school j_1 for student *i*. In many applications this will simply be the posterior probability of belonging to school j_1 and will generally depend on school size and sample design. We shall also assume that where the actual membership is unknown the student does in fact belong to just one primary school (but see below). For simplicity we assume known membership of just one secondary school for each student. Although we do not know the primary school membership, the level 2 contribution to the variance is still σ^2_{u1}. Since we are ignoring the secondary school weights, this implies that [2] becomes

$$\begin{aligned} y_{i(j_1,j_2)} &= (X\beta)_{i(j_1,j_2)} + \sum_{j_1} u_{j_1}\sqrt{\pi_{ij_1}} + u_{j_2} + e_{i(j_1,j_2)} \\ \mathrm{var}\Big(\sum_{j_1} u_{j_1}\sqrt{\pi_{ij_1}}\Big) &= \sigma^2_{u1}\sum_{j_1}\big(\sqrt{\pi_{ij_1}}\big)^2 = \sigma^2_{u1} \qquad [4] \\ \sum_{j_1}\pi_{ij_1} &= 1 \end{aligned}$$

In the special case where we assume $\pi_{ij_1} = \sqrt{1/n_{j_1}}$, where n_{j_1} is the number of students in the j_1 primary school, representing complete agnosticism about primary school membership, this leads to the intra-primary school correlation between two students who are actually in the same primary school, but neither of whose primary school membership is known, becoming

$$\Big(\frac{\sigma^2_{u1}}{\sigma^2_e + \sigma^2_{u1}}\Big)/n_{j_1} \qquad [5]$$

which reflects the fact that the probability of two randomly chosen students belonging to the same school is $1/n_{j_1}$. If we knew that some students did in fact belong to more than one school then a weighting system as in (2) would need to be used in addition to the following, leading to new weights as a product of the two. Thus we can combine (2) and (4) to deal with both missing identifications and multiple unit membership.

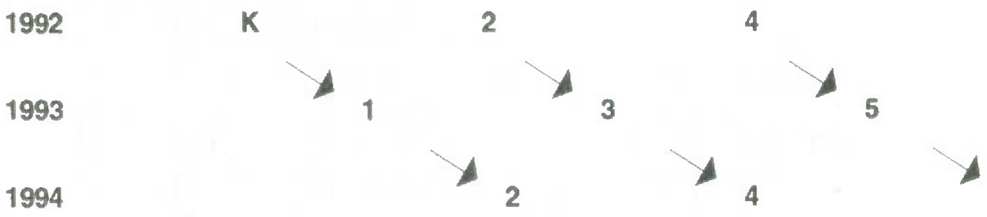
The program *MLwiN* (Goldstein et al., 1998) has special commands which can be used to set up and fit these models.

4. EXAMPLE

To illustrate the application of the above methodology, we make use of data from a three-year longitudinal study of educational effectiveness known as the Victorian Quality Schools Project (Hill, Holmes-Smith, & Rowe, 1993; Hill & Rowe, 1996).

A two-stage, stratified, probability sample of government, Catholic and independent schools in the State of Victoria, Australia, was drawn on the basis of an estimated intra-unit correlation of 0.2 and an average cluster size of 30 (see Ross, 1988a, 1988b). Within these constraints, schools were randomly selected at the first stage of sampling, but with probability proportional to their enrollment size. At the second stage of sampling, the total cohort of students enrolled in the Kindergarten or Preparatory Grade (K), Grade 2, Grade 4, Grade 7 and Grade 9 in each selected school, were included in the sample.

For illustrative purposes, we focus on data relating to teacher ratings of the achievement in English of primary school students. The English scores have been scaled to have a mean of zero and a standard deviation of 1. In the first year of the study (1992), useable data were received from 59 primary schools including 41 government schools, 12 Catholic schools and 6 independent schools, for a total of 6,678 students and 365 teachers. In the second and third years of the study, data were obtained on the same students remaining in the sampled schools as they proceeded to Grades 1, 3 and 5 and 2, 4 and 6 respectively, as indicated diagrammatically below .



Sample attrition rates over the life of the project were relatively high due to a number of factors, one of which was missing data arising from a failure on the part of respondents to answer all questions. However, natural attrition also played a part. Australia is a highly mobile society and a high turnover of students from year to year is common. In addition, policy changes saw the closure of around one in ten Government schools over the three-year period during which the study was in progress and this also had an impact on sample attrition. Finally, five schools dropped out of the project after the first year for a variety of reasons, but mainly on account of workload pressures on teachers. In most cases, the missing data could be regarded as effectively randomly missing rather than systematically related to the characteristics of the students retained in the sample.

In modeling the English achievement of students we assume a bivariate response model of the general form

$$\begin{aligned} y_{ti(j_1,j_2j_3)k} &= (X\beta)_{ti(j_1,j_2,j_3)k} + \beta_z z_{ti(j_1,j_2,j_3)k} \\ &\quad + (v_{tk} + u^{(1)}_{tj_1} + u^{(2)}_{tj_2} + u^{(3)}_{j_3} + e_{ti(j_1,j_2,j_3)k}) \end{aligned} \qquad [6]$$

in which the achievement of student *i*, either at t=1, namely at the end of 1993, or at t=2, namely at the end of 1994, is predicted by a set of student characteristic variables $X_{ti(j_1,j_2,j_3)}$, which may be background or time-varying measures, and a measure of prior achievement $z_{ti(j_1,j_2,j_3)}$ taken at the end of 1992 when t=1 and at the end of 1993 when t=2. The class level terms $u^{(1)}_{tj_1}, u^{(2)}_{tj2}$ and $u^{(3)}_{j3}$ respectively refer to the 1992, 1993 and 1994 class membership of the students. The subscript k indexes schools and V_{tk} is the effect of school k at time t, and $e_{ti(j_1,j_2,j_3)t}$ is the contribution of student i within school k in classes j_1 (1992 class), j_2 (1993 class) and j_3 (1994 class) at time t. The subscript t is thus the indicator for the response. We have not incorporated a subscript t for the 1994 class, since we assume that membership of the 1994 class will affect only the 1994 response. By contrast for the 1992 and 1993 class we have a residual term for both the 1993 response and the 1994 response. Likewise both responses are present at the student and school levels. We also note that the response in 1993 is the prior achievement covariate for the 1994 response. The model [6] can be represented as a five-level model, with observations at t=1 and t=2 nested within students within classes within schools. To accommodate the fact that class composition changed in each of 1992, 1993 and 1994, a cross classification of 1993 classes with 1992 and 1994 classes is introduced into the model by declaring two additional levels in the model, making it a six-level model. The detailed procedure for defining and estimating these models is given in the *MLwiN* user's guide

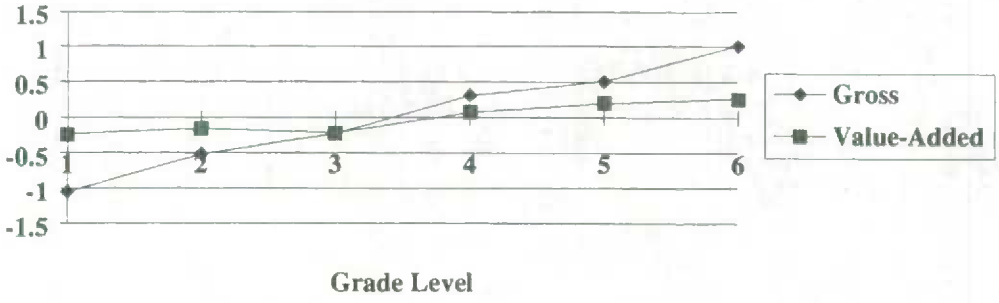
(Goldstein et al., 1998). Note that no level 1 random terms appear in the model, since this defines the bivariate structure (Goldstein, 1995, Chapter 4). Thus, in the random part of the model, there are five variables representing student 1992 class, 1993 class, 1994 class and school effects. Equation [6] assumes that for each student there are two records or sets of observations, one relating to achievement at the end of 1993 and including a 1992 prior achievement measure, and the other relating to achievement at the end of 1994 and including a 1993 prior achievement measure. For each 1993 record, it is assumed that unit identification includes both the 1992 and 1993 class to which each student was assigned and the school to which each student belongs. For each 1994 record, it is assumed that unit identification includes the 1992, 1993 and 1994 class to which each student was assigned and the school identification.

In practice, it was found that there were 460 students with a missing 1992 class identification. To have carried out an analysis using only cases where there was complete identification would have been to reduce efficiency. Also, if missingness is not completely random, by including all available data we will tend to reduce any possible biases.

To incorporate all (4539) students and (6423) records within the one analysis, the assumption was made that the 460 students with missing 1992 class identification information had an equal probability of belonging to any one of the classes within the same school in 1992. Accordingly, we used identification weights calculated as $\sqrt{1/n_{j_2}}$. We note that our procedure is designed to deal with missing *identification* in 1992 records rather than complete missing 1994 records. In the latter case, assuming missingness at random, the data set will be unbalanced with respect to time but we will still obtain efficient (maximum likelihood) estimates (Goldstein, 1995, Chapter 4).

Table 1 summarizes the results of fitting [6] to the data using the above method for handling missing unit identification information within the model. Considering first the fixed parameter estimates, the 'intercept' variables '1993' and '1994' are the achievement levels, expressed in standardised units and adjusted for all other explanatory variables in the model, of students in the earliest years of schooling, namely Grade 1 and Grade 2, in 1993 and 1994. Then follow dummy variables for the other year levels, namely Grades 3 to 6, so that the associated coefficients represent the differences from the Grade 1 and Grade 2 values. Model 1 predicts achievement scores solely on the basis of Grade level, thus providing a 'base' model with which to evaluate the effects of including further explanatory variables. By adding the coefficients for each Grade level to the coefficients for the base Grades (Grade 1 in 1993 and Grade 2 in 1994), estimates can be obtained of average gross achievement levels and (from model 2) of average adjusted or 'value-added' achievement levels across Grades 1 to 6.[4] These are graphed in Figure 1.

Figure 1. *Gross and value-added average scores for Grades 1 to 6*



Turning now to the random part of Model 1, it will be noted that the residual variation at the school level is small and higher for 1994 responses than for 1993 responses, indicating a tendency for the

[4] A more elaborate model which allowed interactions between year and other explanatory variables, that is separate coefficients for each response, was also fitted. The differences in the coefficients between 1993 and 1994 were small, however, and only the simpler model is represented.

gap between schools to grow. The total residual variance for the 1993 response is 0.48 and for the 1994 response is similar at 0.53. The proportion of variance attributable to between school differences is about four per cent for 1993 responses and nine per cent for 1994 responses. It will be noted also from the covariance terms ($\sigma_{v10}, \sigma_{e10}$) that there is a strong, positive correlation at the school level between unadjusted 1993 and 1994 achievement scores (r = 0.849) so that considered jointly, the intra-school correlation is 0.069.

The parameter estimates for residual variance at the class level (levels 3, 4 and 5) of Model 1 indicate that if 1993 and 1994 responses are considered separately, there is an intra-class correlation of 0.353 and 0.330 respectively, indicating substantial differences among classes in teacher ratings of student achievement. Adding 1993 and 1994 effects, this translates into an overall intra-class correlation of 0.376. As one might expect, the effect of 1993 class membership on 1993 responses is significantly greater than its effect on 1994 responses, but the magnitude of the effect on 1994 responses is nevertheless significant.

In Model 2a, student achievement in English is predicted not only by Grade level, but also by four student background characteristics and by Prior Achievement measured one year previously. The parameter estimates indicate that female students make greater progress than males and students from high occupational status families make greater progress than students from families where the highest breadwinners are unemployed or in unskilled occupations. Non-English Speaking Background students make less progress than their English-speaking counterparts and having experienced a Critical Event during the year, such as an extended illness or some psychological trauma, is associated with a negative effect on student progress. Prior Achievement is by far the most important predictor of student progress.

Inclusion of the student background variables and the measure of Prior Achievement result in a significant improvement in model fit as indicated by the log likelihood ratio. Looking at the total (1993 + 1994) effects we note that judged one year at a time, the evidence points to large within-school-between-class differences. The largest effects of 1993 class membership are on 1993 responses, while the largest effects on 1994 class membership are on 1994 responses, however, 1993 class membership has a significant effect on 1994 responses. The effect of 1992 class membership on 1993 and 1994 responses is small and insignificant with respect to the effect on 1994 responses.

Model 2b is identical to Model 2a, but fitted to the data for only those students with complete class identification information. In other words, the 460 students with missing 1992 class identification information were omitted from the analysis. This was done to investigate any biasing effects from excluding these students from the analysis. The parameter estimates for both the fixed and the random parts of the model are in most instances very similar, suggesting that students with missing 1992 class information were broadly representative of the full sample, although there are differences between the intercepts for the different year levels.

5. DISCUSSION

This above approach to handling multiple membership and missing identification problems can be generalized to a wide range of data analytic problems and practical situations. It can be used in both univariate and multivariate linear and non-linear models involving complex hierarchical structures. Its main application, however, would appear to be in connection with longitudinal studies where problems of missing data related to the classification of students by classes or schools are frequently encountered as students change school or the composition of classes changes from one year to the next. This is particularly important for school effectiveness studies where the standard approach has been simply to omit students who change school during the course of a study, even when this involves a majority of the students. Another application is in household studies where individuals move across households over time and in some cases identification of households may be missing.

An important issue is how to determine the weights in a multiple membership model or unit membership probabilities in a missing identification model. In practice it would be advisable to carry out sensitivity analyses by varying the weights and probabilities. For example, we might choose a logarithmic transformation of the time spent in a school to determine the weights. More generally we may suppose that the weight is a function of further explanatory variables, for those individuals who belong to more than one unit. Thus we might assume that the weight is a linear function of time spent and other individual level covariates. This leads to a model involving the product of unknown fixed coefficients and random effects.

A special case is where the covariate is a grouping variable based upon time spent. One difficulty with this approach lies in constraining the weights to add to unity for each individual. One possible solution is to scale the weights so that *on average* they added to unity. The estimation problems associated with this approach are currently being investigated.

The model can be extended to the case where units divide into sub-units or amalgamate into combined units. Thus, for example, schools may combine during the course of a longitudinal study or households may split with members leaving to form new households. To incorporate such possibilities we form a superset of units including all the original units together with the new sub-units and combined units. Membership can then allocated to each unit, for example, according to the time spent in each unit, whether an original or new unit.

ACKNOWLEDGMENTS

We are most grateful to Ken Rowe for comments on a draft. This work was partly supported by the Economic and Social Research Council (UK) under its programme for the Analysis of Large and Complex Datasets. This paper is based upon a longer paper (Hill and Goldstein, 1998).

REFERENCES

Castles, I. (1986). *Australian standard classification of occupations: Statistical classification.* Australian Bureau of Statistics. Canberra: C. J. Thompson Government Printer.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*,74, 430-1

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold; New York: Halsted Press.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., et al. (1998). *A user's guide to MLn for windows (MLwiN).* London, Institute of Education:

Hill, P.W., Holmes-Smith, P., and Rowe, K.J. (1993). *School and teacher effectiveness in Victoria: Key findings from Phase 1 of the Victorian Quality Schools Project.* Melbourne: Centre for Applied Educational Research, The University of Melbourne (ERIC Clearing House, Document No. ED 367 067).

Hill, P.W., and Rowe, K.J (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7, 1-34 .

Hill, P. W. and Goldstein, H. (1998). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural statistics (to appear).*

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19*, (4), 337-350.

Rasbash, J., & Woodhouse, G. (1995). *MLn Command Reference*. London: Institute of Education, University of London.

Rasbash, J., Healy, M.J.R., Browne, W., and Cameron, B. (1998*). MLwiN; a visual interface for multilevel modelling*. London, Institute of Education.

Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

Ross, K.N. (1988a). Sampling. In J.P. Keeves (E.), Educational research, methodology and measurement: An international handbook (pp.527-537). Oxford: Pergamon Press.

Ross, K.N. (1988b). Sampling errors. In J.P. Keeves (E.), Educational research, methodology and measurement: An international handbook (pp.537-541). Oxford: Pergamon Press.

Table 1. *Parameter estimates (and Standard Errors) for two bivariate response cross-classified models with missing identification codes. Model 2b is identical to model 2a but with subjects excluded who have a missing 1992 class identification.*

Parameters		Model 1 (N=6423)	Model 2a (N=6423)	Model 2b (N=5963)
Fixed:				
1993 intercept		**-1.054 (0.010)**	**-0.246 (0.049)**	**-0.223 (0.051)**
Grade 3	t=1	**0.815 (0.057)**	**0.019 (0.063)**	**0.103 (0.064)**
Grade 5		**1.561 (0.055)**	**0.444 (0.062)**	**0.342 (0.066)**
1994 intercept		**-0.515 (0.051)**	**-0.164 (0.051)**	**-0.032 (0.056)**
Grade 4	t=2	**0.823 (0.064)**	**0.236 (0.069)**	**0.061 (0.078)**
Grade 6		**1.527 (0.066)**	**0.422 (0.071)**	**0.365 (0.083)**
Gender (Female)		-	**0.067 (0.010)**	**0.056 (0.010)**
Non-English Speaking		-	**-0.045 (0.020)**	**-0.028 (0.021)**
Occupational Status		-	**0.059 (0.006)**	**0.058 (0.007)**
Critical Events		-	**-0.045 (0.012)**	**-0.036 (0.012)**
Prior Achievement		-	**0.700 (0.010)**	**0.721 (0.010)**
Random:				
School				
σ^2_{v1} **(1994)**		**0.052 (0.019)**	**0.027 (0.016)**	**0.008 (0.014)**
σ^2_{v0} **(1993)**		**0.018 (0.010)**	**0.044 (0.017)**	**0.032 (0.014)**
σ_{v10} **(1993,1994)**		**0.026 (0.011)**	**0.010 (0.012)**	**0.012 (0.010)**
Class				
$\sigma^2_{u(j3)0}$ **(1994)**		**0.181 (0.020)**	**0.184 (0.019)**	**0.221 (0.024)**
$\sigma^2_{u(j2)1}$ **('93 class for '94 response)**		**0.012 (0.006)**	**0.095 (0.011)**	**0.102 (0.013)**
$\sigma^2_{u(j2)0}$ **('93 class for '93 response)**		**0.175 (0.017)**	**0.176 (0.017)**	**0.145 (0.015)**
$\sigma_{u(j2)10}$ **(1993,1994)**		**0.004 (0.008)**	**-0.129 (0.013)**	**-0.119 (0.013)**
$\sigma^2_{u(j1)0}$ **('92 class for '94 response)**		**0.002 (0.004)**	**0.002 (0.003)**	**0.001 (0.002)**
$\sigma^2_{u(j1)1}$ **('92 class for '93 response)**		**0.009 (0.004)**	**0.085 (0.010)**	**0.102 (0.011)**
Student				
σ^2_{e1} **(1994)**		**0.279 (0.008)**	**0.132 (0.004)**	**0.130 (0.004)**
σ^2_{e0} **(1993)**		**0.281 (0.007)**	**0.151 (0.004)**	**0.150 (0.004)**
σ_{e10} **(1994,1993)**		**0.203 (0.006)**	**-0.030 (0.004)**	**-0.035 (0.004)**
-2*log(likelihood)		**10494**	**8149**	
Intra-school correlations				
1994 responses		**0.094**	**0.061**	
1993 responses		**0.037**	**0.096**	
1994 + 1993 responses		**0.069**	**0.079**	
Intra-class correlations				
1994 responses (1994 classes)		**0.330**	**0.418**	
1993 responses (1993 classes)		**0.353**	**0.386**	
1994 + 1993 responses		**0.376**	**0.606**	
σ_{e10} **(1993,1994)**		**0.197 (0.006)**	**-0.0** **29 (0.004)**	**-0.039 (0.004)**
-2*log(likelihood)		**107**	**8289**	

SESSION 3

CORRECTING FOR NONRESPONSE

Proceedings of the Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

WEIGHTING VERSUS MODELLING IN ADJUSTING FOR NON-RESPONSE IN THE BRITISH LABOUR FORCE SURVEY: AN APPLICATION TO GROSS FLOWS ESTIMATION

P.S. Clarke[1] and P.F. Tate[2]

ABSTRACT

The British Labour Force Survey (LFS) is a quarterly household survey with a rotating sample design that can potentially be used to produce longitudinal data, including estimates of labour force gross flows. However, these estimates may be biased due to the effect of non-response. Weighting adjustments are a commonly used method to account for non-response bias. We find that weighting may not fully account for the effect of non-response bias because non-response may depend on the unobserved labour force flows, i.e., the non-response is non-ignorable. To adjust for the effects of non-ignorable non-response, we propose a model for the complex non-response patterns in the LFS which controls for the correlated within-household non-response behaviour found in the survey. The results of modelling suggest that non-response may be non-ignorable in the LFS, causing the weighting estimates to be biased.

1. INTRODUCTION

The British Labour Force Survey (LFS) is a household survey gathering information on a wide range of labour force characteristics. Since 1992 it has been conducted on a quarterly basis, with a sample rotation scheme replacing one-fifth of the sample each quarter. The survey is designed to produce cross-sectional data, but it has been recognised that linking together data on each individual across quarters produces a rich source of longitudinal data. An important use of linked LFS data is for the estimation of labour force gross flows. However, these estimates may be biased, mainly by the effects of non-response and misclassification.

This paper describes work dealing with non-response bias in the context of estimating labour force gross flows from LFS data. Tate (1997) investigates the characteristics of non-responding persons in the LFS, and suggests a method to compensate for non-response by incorporating the results of this investigation into the weighting procedure used to produce population level estimates from the survey. However, there are two problems which could potentially limit the effectiveness of this approach: first, non-response at successive interview rounds may be non-ignorable, that is, the propensity to non-respond may depend on the unobserved labour force status of an individual; and second, the household-based nature of LFS may lead to household-level non-response where non-response behaviour is correlated within households.

A class of models for estimating gross flows in the presence of non-ignorable non-response is presented in Little (1985). However, this work assumes that non-response is an individual-level process where individuals respond or non-respond independently of other individuals. Clarke and Chambers (1997) propose a model that controls for household-level non-response. We assess the suitability of the weighting approach in this situation by comparing labour force gross flows estimates obtained using weighting with those from fitting household-level non-ignorable non-response models. A linked data-set from two successive quarters of the LFS is used in this analysis, namely, the summer and autumn quarters of 1995.

[1] Department of Social Statistics, University of Southampton, Highfield, Southampton S017 1BJ, U.K.

[2] Office for National Statistics, RG/11, 1 Drummond Gate, London SW1V 2QQ, U.K.

2. WEIGHTING ADJUSTMENTS FOR NON-RESPONSE BIAS

Non-respondents may be atypical of the population as a whole, and so their loss from the sample can introduce bias. To some extent, non-response bias in the LFS is compensated for in the course of applying the normal weighting procedure, which weights the sample so that the marginal distributions of the control variables corresponding to sex, age-group and region are consistent with the population (estimated using census data and population projections produced by the Office for National Statistics and the Government Actuary's Department). It is clear that this process will compensate for bias arising from differential non-response by sex, age-group and region, but bias associated with characteristics not used in the weighting procedure may not be compensated for.

Tate (1997) found that, consistently across waves, young people aged 18-29 (especially 18-24), single people, those living in London or in rented accommodation (especially privately rented), and the unemployed or those in temporary employment are under-represented in the LFS sample, that is, are more likely to non-respond. Characteristics which are independently associated are identified using logistic regression and CHAID analysis (Magidson, 1993). This analysis reveals that being a young adult and living in privately rented accommodation are the principal determinants of non-response (there are also some relatively minor additional effects due to being single, in temporary employment or economically inactive). A separate analysis for sample members lost through non-contact and moving house shows the same factors determine non-response for both groups.

The analysis described above identifies the cross-sectional characteristics of sample members that are associated with greater non-response. However, there exists a possibility that non-response might be greater when the sample member is changing economic activity state between interviews, independently of which state they start or finish in. If this is the case, the proportion of people making transitions between different economic activity states would be lower at later waves. The analysis finds this not to be the case, suggesting that there is no additional probability of non-response associated with changing economic activity state. The existing weighting system for the LFS already incorporates age-group, hence incorporating tenure into the weighting procedure will compensate for further non-response bias. Tenure is integrated into the weighting procedure using a calibration method via the software package CALMAR (Sautory, 1992). The customary variables of sex, age-group and region provide control totals, and then the prior weights adjust each category of tenure to the proportions in a cross-section of the LFS sample. We apply the calibration method to a sample of 32,282 households interviewed at both quarters in our sample; the prior weights are used to adjust the tenure categories to the proportions in the first cross-sectional sample. The trial weighting was restricted to the population of working age during both quarters.

Comparing the weighted data thus produced with the unweighted linked data and the cross-sectional data from

cross-sectional data, it does not fully reproduce the population distributions by marital status or economic activity. This is despite the previous finding that tenure, together with age, captures most of the non-response effects. Thus, the method is reapplied with prior weights calculated on the basis of tenure plus marital status,

are needed to adequately compensate for non-response. However, the inclusion of economic activity in the derivation of the prior weights makes the unsatisfactory implicit assumption that the distribution of the cross-sectional sample by economic activity is unbiased. It is also apparent that using prior weights which incorporate economic activity in the first quarter does not satisfactorily reproduce the cross-sectional distribution of economic activity at the second quarter.

3. MODEL-BASED ADJUSTMENTS FOR NON-RESPONSE BIAS

We are concerned with estimating labour force gross flows from the LFS where individuals' labour force status may be unobserved at any stage. In Section 2 we reveal that non-response is non-ignorable because it depends on individuals' missing labour force status. It is further concluded that weighting cannot fully control for non-response bias because the weighting adjustments cannot account for sample loss due to economic activity at the second quarter. Hence, we adopt a model-based approach to compensate for the effects of non-ignorable non-response in the LFS. Theoretical details about modelling non-response in sample surveys are given in Little (1982).

An attempt to estimate labour force gross flows from data subject to non-ignorable non-response is made in Little (1985). However, there are two reasons why Little's approach is not wholly valid in this situation. First, non-response is treated as an individual-level process where each individual's non-response behaviour is independent. In fact, the LFS is a household survey and so non-response within households may be correlated. For example, the LFS interview procedure involves asking one person for permission to interview the household. If that person refuses to grant permission, then each household member non-responds. Second, no distinction is made between different reasons for non-response. As well as refusals, non-response occurs in the LFS due to households being non-contactable, individuals moving house and sample rotation. In addition, sample rotation leads to labour force flows not being observed. The underlying mechanisms for each non-response state would appear to be different, and so must be distinguished between by the non-response model.

3.1 Non-response Patterns in the Labour Force Survey

We now give details about the LFS design that are of particular relevance to modelling non-response. A full description of the LFS survey design is given in ONS (1997).

The LFS interview procedure consists of two states. Stage one involves contacting an eligible household member in order to gain permission to interview the household; stage two involves interviewing the eligible household members.

Non-response at stage one falls into one of three categories: outright refusal, circumstantial refusal and non-contact. Outright refusal occurs when the household, or rather the individual from who permission is sought, refuses to take part in the survey. Circumstantial refusal is less terminal, arising when the household does not agree to be interviewed because the timing, or other conditions, is inconvenient. The third category of interview non-response, non-contact, refers to the situation where it is not possible to contact an eligible household member at all. In the event of an outright refusal, no further attempt is made to obtain an interview, that is, the household is dropped from the survey. In the event of circumstantial refusal or non-contact, a further attempt is made to contact the household at the next wave. Any further non-response at subsequent waves results in the household being dropped.

Stage two non-response occurs when individuals non-respond within an otherwise responding household. It may be impossible to elicit information from certain individuals within a household because, for example, other household members are not willing to provide proxy responses. At subsequent waves, an attempt will be made to interview non-responding individuals if the household responds.

A final complication arises due to the sample changing between quarters. As well as households leaving the sample by being rotated-out, or entering the sample by being rotated-in, households can enter or leave by moving-in or moving-out of a house. A household that moves-out of the sample leaves a slot in the sample at that address which may be filled by a household that moves-in. Individuals can also move-in or move-out of households

3.2 Modelling Household-level Non-response in the Labour Force Survey

Represent the complete sample of *63,486* households by S, indexed by h. The first quarter is herein referred to as t_1 and the second quarter as t_2. The number of eligible individuals in h between t_1 and t_2 is denoted by n_h, which we take to be known and fixed between both quarters. An eligible individual is of working age and resident on the British mainland. Labour force status has three categories: E='Employed', U='Unemployed' and N='Not in Labour Force'. To make the following results comparable with the weighted estimates, it is necessary to control for sex and age-group when estimating the labour force flows. Sex has the usual categorisation, and age-group has three groups: '16-24', '25-44' and '45 and over'. There are six categories formed by the cross-classification of sex and age-group, indexed by $x=1,2,...,6$.

The number of individuals in h with labour force flow (a,b) and sex-age-grouping x is assumed to be the random variable $N_h(abx)\geq 0$, where $\sum_{a,b,x}N_h(abx)=n_h$. The non-response status of h at t_j is denoted by R_{hj}, for $j=1,2$, with 8 outcomes: I='Interview', NC='Non-contact', CR='Circumstantial Refusal', OR='Outright Refusal', MI='Move-in', MO='Move-out', RI='Rotated-in' and RO='Rotated-out'. The realisations of the non-response indicators R_{h1} and R_{h2} determine the non-response flow of h. The non-response states are divided into two groups: in-sample and in-transit. States I, NC, CR and OR are in-sample non-response states because the household is resident in the sampled address and is approached for an interview. The other states, MI, MO, RI and RO are called in-transit states because the household is in the process of leaving the address or entering the address, and as such is not approached for an interview. For example, RO refers to a household that was rotated-out of the sample after the previous quarter, and MI is a household that moves-in to a selected address in the following quarter.

The individual frequencies for h are represented by random vector $\mathbf{N}_h=(N_h(EE1),N_h(EU1),...,N_h(NN6))$, and the household non-response flow is represented by random vector $\mathbf{R}_h=(R_{h1},R_{h2})$. The realisations of these random quantities are denoted by $\mathbf{n}_h$ and $\mathbf{r}_h$, respectively. The joint model for the labour force flows, the sex-age-grouping and the non-response flows is specified as

$$\Pr(\mathbf{N}_h=\mathbf{n}_h, \mathbf{R}_h=\mathbf{r}_h)=\Pr(\mathbf{N}_h=\mathbf{n}_h)\Pr(\mathbf{R}_h=\mathbf{r}_h \mid \mathbf{N}_h=\mathbf{n}_h),$$

where the first factor of the right-hand-side is called the labour force flows model and the second factor is called the non-response flows model.

We assume that the labour force flows and sex-age groupings within each household follow a multinomial distribution, that is,

$$\mathbf{N}_h \sim \mathrm{MN}(n_h,\omega)$$

where $\omega=(\omega(EE1),\omega(UE1),...,\omega(NN6))$ is a vector of probabilities, $\omega(abx)>0$ is the probability of having labour force flow (a,b) and sex-age-grouping x, and $\sum_{a,b,x}\omega(abx)=1$. The main implications of the multinomial assumption are twofold: individuals' behaviour is independent within households; and since ω is constant, households are assumed to be homogeneous with respect to their individuals' behaviour. These are strong assumptions that are unrealistic in practice. We also make the further assumption that non-response and sex-age-grouping are conditionally independent given labour force status, which enables the non-response model to be constructed as follows.

As discussed previously, there are two problems to overcome when modelling non-response in the LFS: first, a satisfactory way of modelling the probability of a household non-responding as a function of its individual characteristics; and second, the complex non-response patterns which occur in the LFS must be accommodated into the model. To tackle the first point, we follow Clarke and Chambers (1997) and model the probability of a household non-response flow as a weighted average of the individual labour force flows frequencies for the household, namely,

$$\pi(uv|\boldsymbol{n}_h) = \frac{1}{n_h}\sum_{a,b} n_h(ab)\psi(uv|ab), \tag{1}$$

for all u,v, where $\pi(uv|\mathbf{n}_h)=\Pr(\mathbf{R}_h=(u,v)|\mathbf{N}_h=\mathbf{r}_h)$ and $\{\psi(uv|ab)\}$ are the coefficients of the weighted average. Since $\{\pi(uv|\mathbf{n}_h)\}$ are probabilities, setting $n_h=1$ in (1) gives $\psi(uv|ab)$ the interpretation of the probability of a household of size one (i.e., an individual) having non-response flow (u,v) given labour force flow (a,b). The task of (1) is to control for clusters of individuals with the same non-response pattern. Hence, individuals who non-respond in responding households are treated as separate non-responding households.

To parameterise the complex non-response patterns which occur in the LFS, we consider the non-response mechanism to be a three-stage process (itself preceded by the allocation of individual labour force flows by the labour force flows model). The first two stages determine whether the non-response is an in-transit or in-sample state. Stage one: the household is determined to be *RI* or *RO* or neither. Stage two: given that the household is not *RI* or *RO*, it is determined to be *MI* or *MO* or neither. The first two stages are conditional because flows between in-transit non-response states at t_1 and t_2 are impossible or cannot be identified by the survey. For example, a household cannot be *RI* at t_1 and *RO* at t_2 because the design stipulates it must be in-sample for five consecutive quarters. Furthermore, it is not possible to identify households that are *RI* at t_1 and *MO* at t_2 because they are never in the survey, entering an address after the first interview and leaving before the second, and similarly for other flows between in-transit non-response states. Thus, stage three determines the household in-sample non-response state, either for the quarter the household was in-transit, or for both quarters. Each stage is parameterised in the following way.

Sample rotation is, by definition, determined entirely by the survey design and is independent of labour force flows. Thus, the probability of *RI* at t_1 is α_1, the probability of *RO* at t_2 is α_2, and the probability of neither is $1-\alpha_1-\alpha_2$. Conditional on the household neither being *RI* nor *RO*, we allow the probability of the household moving-in or moving-out to depend on its labour force flows; thus, the probability of *MI* at t_1 is $(1-\alpha_1-\alpha_2)\beta_1(ab)$, the probability of *MO* at t_2 is $(1-\alpha_1-\alpha_2)\beta_2(ab)$ and the probability of neither *MI* nor *MO* is $(1-\alpha_1-\alpha_2)(1-\beta_1(ab)-\beta_2(ab))$. In-sample non-response is unconditional on the preceding in-transit non-response states. If a household is in-sample during both quarters, most flows between *I*, *NC*,*CR* and *OR* are possible; and if the household is in-sample at one quarter only, all in-sample non-response states are possible. The only constraint on flows between in-sample non-response states is due to *OR* households at t_1 being dropped at t_2. To parameterise the in-sample non-response structure, we factorise the in-sample non-response probabilities into: $\lambda_u(ab)$, the probability of in-sample non-response state u at t_1 given labour force flow (a,b) $(u=I,NC,CR,OR)$; and $\theta_{v|u}(ab)$, the probability of v at t_2 given u at t_1 and labour force flow (a,b) $(u,v=I,NC,CR,OR)$, where $\theta_{v|OR}(ab)=1$ if $v=OR$ and 0 otherwise. For households in-sample at t_2 the marginal in-sample probabilities are $\theta_v(ab)$. Examples of this parameterisation are: $\psi(RI,I|\mathrm{ab})=\alpha_1\theta_I(ab)$, $\psi(NC,MO|ab)=(1-\alpha_1-\alpha_2)\beta_2(ab)\lambda_{NC}(ab)$ and $\psi(I,CR|ab)=(1-\alpha_1-\alpha_2)(1-\beta_1(ab)-\beta_2(ab))\lambda_I(ab)\theta_{CR|I}(ab)$, which are substituted into (1).

3.3 Non-response Models for the Labour Force Survey

Model fitting and estimability for this class of models is discussed in Clarke and Chambers (1997). To ensure estimability, we must constrain the non-response model in accordance with some plausible assumptions about the non-response mechanism. The first model considered is ignorable model I_A where the probability of non-response is independent of labour force status: $\beta_j(ab)=\beta_j$, $\lambda_u(ab)=\lambda_u$ and $\theta_{v|u}(ab)=\theta_v$. This non-response model has only 10 parameters and so is not overparameterised. We then consider three non-ignorable models, where non-response depends on labour force status at the current time. Model N_{A1} is the same as I_A except that $\lambda_u(ab)=\lambda_u(a)$ and $\theta_{v|u}(ab)=\theta_v(b)$, a total of 22 non-response parameters. Model N_{A2} is the same as N_{A1} except that $\beta_1(ab)=\beta_1(a)$ and $\beta_2(ab)=\beta_2(b)$, a total of 26 parameters. The final non-ignorable model is N_B where the probability of in-sample non-response depends on being interviewed at t_1: $\theta_{v|I}(b)=\lambda_v(a)$ and $\theta_{v|u}(ab)=\theta_v(b)$ if $u\neq I$, also a total of 26 parameters.

4. RESULTS AND DISCUSSION

The estimated probabilities of each labour force gross flow are shown in Table 1. The first two columns are estimates from the unweighted data and those weighted using prior weights based on tenure, marital status and economic activity; the next four columns are estimates under the four estimable non-response models, I_A, N_{A1}, N_{A2} and N_B.

Table 1 - Gross flows estimates: unweighted, weighted and for the non-response models

Gross flows	Unweighted	Weighted	Model I_A	Model N_{A1}	Model N_{A2}	Model N_B
EE	70.62	69.84	68.89	66.36	66.13	66.52
EU	1.08	1.13	1.15	1.52	1.46	1.46
EN	1.53	1.55	1.58	1.50	1.50	1.51
UE	1.61	1.77	1.66	2.09	2.19	1.79
UU	3.78	4.33	4.30	6.92	7.13	7.17
UN	1.00	1.09	1.09	1.22	1.32	1.28
NE	1.40	1.42	1.39	1.32	1.31	1.31
NU	1.06	1.11	1.15	1.52	1.43	1.41
NN	17.92	17.75	18.79	17.55	17.53	17.55

The main differences between the estimates relate to the categories for employed in both quarters and unemployed in both quarters. For both categories, the estimates are similar for the weighted data and model I_A, but change considerably when moving to the various non-ignorable models (the employed proportion decreasing and the unemployed increasing). The changes between the ignorable and the non-ignorable model estimates are less pronounced for the transitional flows between different economic activity states, with the exception of transitions into unemployment. The only aspect that does not conform to this pattern is the relatively high proportion not in the labour force at both quarters estimated by the ignorable model.

The results of the work on weighting suggest that, although more of the non-response bias can be related to differential non-response by age or tenure, adjusting for these characteristics alone, or even together with additional variables which contribute to non-response to a lesser degree, is not enough to reduce non-response bias satisfactorily. Furthermore, there is evidence of non-ignorable non-response in the LFS. Indeed, the salient difference between the non-ignorable model estimates and the others is that unemployed individuals are being under-represented in the LFS sample. This is a feasible result, indicating that the model-based approach is at least producing substantively believable adjustments.

Before more concrete conclusions about the existence and size of the effects of non-ignorable non-response in the LFS can be drawn, it is necessary to carry out further work. The parameters of non-ignorable non-response models are weakly identified (see, for example, Little (1982, p. 246)) and thus point estimates may be extreme, that is, the effect of non-ignorable non-response may be overstated. For example, for model N_{A1} the probability of *CR* at t_1 for unemployed is 0.56, and for not in labour force it is 0. It is unlikely that the true probabilities are so extreme, and so one avenue may be to constrain the non-response probabilities to lie in regions of believable values and re-estimate the labour force gross flows. Furthermore, the instability of the non-response parameters implies the parameter estimate variances are large. We plan to refit our models to further linked data-sets from the LFS to ascertain whether these results are the result of random variability or the result of a systematic effect due to non-ignorable non-response. Lastly, if non-ignorable non-response is found to be a significant phenomenon in the LFS, we plan to perform simulation studies to assess the robustness of any proposed weighting adjustments based on our models to mis-specification.

REFERENCES

Clarke, P.S., and Chambers, R.L. (1997), Estimating labour force gross flows from surveys subject to household-level nonignorable nonresponse. Presented at the IASS/ISO Satellite Meeting on Longitudinal Studies, Jerusalem.

ONS (1997). *LFS User Guide - Volume 1: Background and Methodology*, Socio-economic Division, Office for National Statistics, U.K.

Little, R.J.A. (1982), Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, **77**, pp. 237-250.

Little, R.J.A. (1985), Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistics Institute*, Proceedings of the 45th Session: Invited Papers, Section 15.1, pp. 1-18.

Magidson, J. (1993). *SPSS for Windows: CHAID, Release 6.0*, SPSS Inc.

Sautory, O. (1992), Calibration on known marginal counts for sample surveys: practical experiences at INSEE. Paper presented to the Workshop on Uses of Auxiliary Information in Surveys, Orebro, Sweden.

Tate, P.F. (1997), Utilising longitudinally linked data from the Labour Force Survey. Presented at the IASS/ISO Satellite Meeting on Longitudinal Studies, Jerusalem.

Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

CALCULATION OF WEIGHTS FOR THE EUROPEAN PANEL OF HOUSEHOLDS: A COMPARISON OF SOME INDICATORS BASED ON VARIABLES SELECTED TO CORRECT FOR HOMOGENEOUS CATEGORICAL NON-RESPONSE

Christine Chambaz[1]

ABSTRACT

At the end of 1993, Eurostat launched a 'community' panel of households. The first wave, carried out in 1994 in the 12 countries of the European Union, included some 7,300 households in France, and at least 14,000 adults 17 years or over. Each individual was then followed up and interviewed each year, even if they had moved. The individuals leaving the sample present a particular profile. In the first part, we present a sketch of how our sample evolves and an analysis of the main characteristics of the non-respondents. We then propose 2 models to correct for non-response per homogeneous category. We then describe the longitudinal weight distribution obtained from the two models, and the cross-sectional weights using the weight share method. Finally, we compare some indicators calculated using both weighting methods.

KEY WORDS: Panel, longitudinal weights, cross-sectional weights.

1. INTRODUCTION

The European panel of households was launched in 1994 in the 12 countries which at that time made up the European Union. Its main purpose was to study the employment and income dynamics of *individuals*. The survey is composed of two questionnaires, one for the household (composition, housing, etc.) and the other for individuals 17 years or over (income, employment, health, relationships, etc.).

Contrary to what its name would imply, the European panel of households is a *panel of individuals*. The evolution of the sample was determined in reference to individuals, who fall into two categories:

- the individuals making up the households selected in year 1 are the panel's basic sample. They will be contacted annually for the duration of the panel, even if they move (within metropolitan France). They make up the population of *panel individuals*. By definition, all individuals of households responding to wave 1 of the panel are panel individuals. By agreement, children born over the course of subsequent waves to mothers in the original individual panel are also panel individuals.

- other adults in the households with at least one panel individual are interviewed, but only if they are living with that panel individual (adult or child). They make up the population of *non-panel individuals*. By definition, this population exists only as of wave 2 of the panel, and should be evolving over the coming waves.

The European panel of households is not an actual panel, but a cohort study, in that the sample is not renewed (Deville, 1998).

[1] INSEE, 'Revenus et patrimoine des ménages' Division, Stamp F350, 18 boulevard A. Pinard, 75675 Paris Cedex 14, FRANCE.

Because there are two categories of population, two types of weightings are calculated:

- a longitudinal weighting for panel individuals, so that changes in the status of those individuals can be studied. This longitudinal weighting is derived from the basic individual weight, defined as the initial weights corrected for changes in the sample. For a given period *(t,t+n)*, we can calculate as many longitudinal weightings as there are follow-up subperiods.

- a cross-sectional weighting, to be used to do a cross-sectional analysis of individual responses. This weighting is common to all adult respondents in a household, whether or not they are panel individuals. By taking non-panel individuals into account for the cross-sectional analysis, we can partly compensate for a shrinking sample as panel individuals leave the scope of the survey. In principle, it must ensure that the population of households interviewed is representative at all times.

The weightings were calculated based on the method recommended by Eurostat (Eurostat, 1995). We will first analyze individual non-response of adult panel individuals (17 years and over) and estimate the probabilities of non-response in homogeneous groups (part 2). Two non-response models are tested to assess the impact of choice of model on the distribution of weights and the estimation of some indicators. We then calculate the longitudinal basic weights in wave t (Wt) by correcting the basic weights in wave (t-1) (Wt-1) based on those non-response probabilities (part 3). We then calibrate the longitudinal basic weights to the age and gender structure of the population in t, calculated from the Employment Survey, and from those weights, derive the cross-sectional weights calculated using the weight share method (part 4). Finally, we compare some indicators calculated using either set of weightings to the total population and then to the most mobile sub-populations (part 5). The analyses are essentially based on wave 2 of the panel. Reference is made to wave 3 where necessary. This document was taken in large part from a paper presented at the Insee 1998 Journées de Méthodologie Statistique (Chambaz and Legendre, 1998).

2. ANALYSIS OF INDIVIDUAL NON-RESPONSE AND CORRECTION PER HOMOGENEOUS GROUP

2.1 Changes in the sample between waves 1, 2 and 3

In wave 1 (W1), 7,344 households were interviewed, which represented approximately 76% of the households in scope for the survey. They were composed of 18,916 individuals, including 14,524 adults eligible to respond to the individual questionnaire.

In wave 2 (W2), only 'panel' individuals in the households *which responded in W1* are supposed to be re-interviewed. However, some have left the scope of the survey, i.e. are deceased, the entire household has moved, or they have emigrated. On the other hand, newborns have entered the survey population. In total, 18,909 panel individuals must be tracked, including 14,636 individuals eligible to respond to the individual questionnaire.

Response rates are relatively high, with 12,986 panel adults (88.8%) having agreed to be interviewed. The attrition rate, defined as the gap between the number of panel individuals responding in W2 and the number responding in W1 is thus 9.4%.

In wave 3 (W3), all individuals entering the sample in W2 must be tracked, *whether or not they responded in V2*. This is a major change compared to wave 2.

Taking into account departures from the sample scope and births, 18,912 panel individuals must be tracked, including 14,724 adults. The response rates are again relatively high, although lower than in W2[2]: 12,533 panel adults (85.1 %) agreed to be interviewed. The attrition rate between W2 and W3 is 3.5 %.

2.2. Selection of an individual non-response model and correction of non-response per homogenous groups (CNRHG)

The non-response analyzed here involves panel individuals only ; non-panel individuals by definition have a nil basic weight. Children are considered to have responded if the household was contacted. For them, the non-response correction involves correcting their basic weight using the basic non-contact rate of households including children (rate weighted per number of children in those households). By agreement, newborns are assigned half the weight of their mothers. Response for adults is defined as response to the individual questionnaire. This population will not be referred to in the rest of this document.

Information is gathered in two stages: first the household is contacted and then the individuals in the household are interviewed. We therefore could have first estimated the probability of contacting the individual through his or her household, and then a probability of non-response based on contact. We preferred to directly estimate the probability of non-response by introducing into the explanatory model variables strongly correlated to the probability of contact: nationality, type of household, socio-professional category, even moving of the household, etc.

Two explanatory models for the non-response of panel adults were estimated, in W3 as in W2. For each model, exogenous variables were gathered during the preceding wave, and are thus available for both responding and non-responding individuals. In one of the models, we also introduced a variable describing geographic mobility *since* the preceding wave, based on information gathered by the investigator.

For waves 2 and 3, the most discriminant variables from the point of view of non-response are much the same, i.e. in the two models calculated, the type of household, nationality of the household's reference person, his or her activity or socio-professional category, the type of community, and the age of the individual, and even (wave 3) housing-related expenses. However, when moving following a family break-up is introduced as a variable, it appears however to be by far the most discriminant, particularly in wave 2.

The non-response correction was calculated by modifying the basic weights (W(t-1) using the following formula: [see last page for translation]

$$\text{Basic weights } \mathrm{W}(\mathrm{t}) = \frac{\text{Basic weights } \mathrm{W}\,(\mathrm{t}-1)}{1-\text{non}-\text{response rate }(\mathrm{t})}$$

To limit weight dispersion, the non-response rates were calculated for the homogeneous categories defined by the intersection of the variables that appear to be the most discriminant.

For example, for wave 2, approximately 30 categories were established. Categories with only a few individuals are grouped with the closest category to calculate the rate of non-response. The proximity of categories was determined using coefficients estimated in the model. The rate of non-response varies, based on category, from 4.6% to 38.8%, without taking into account moving households for the CNRHG, and between 2.9% and 58.9% when taken into account.

In wave 3, as in wave 2, the distribution of basic weights for adult panel individuals is slightly more dispersed when we incorporate the moving and break-up variable within the CNRHG.

[2] This higher rate of non-respoonse is automatically created by the rules for tracking individuals, which stipulate that persons who refused to respond in W2 must be tracked.

The comparison of weights for waves 1 and 2 or 2 and 3 also show an increased range of weight ratios in this case. For example, in W2, this ratio is on average 1.1267, i.e. between 1.0297 and 2,4317[3], compared to the average of 1.1265 when the CNRHG does not incorporate the moving variable, in which case the ratio is then between 1.0486 and 1.6339.

3. CALIBRATING TO MARGINS AND CALCULATING CROSS-SECTIONAL WEIGHTS

3.1. Calibrating to margins of W2 basic longitudinal weights

The purpose of calibrating to margins is to ensure greater cross-sectional representivity of our sample, and was done before calculating cross-sectional weights.

Calibration has some drawbacks in the European panel of households. Because the sample is not renewed, it is impossible to represent individuals joining families through immigration. This limitation may however be ignored, because the panel is over just six years and immigration levels are relatively low.

Only the population used to calculate the cross-sectional weights (i.e. respondant panel adults) was recalibrated. The calibration is minimal, based only on individual structures per 10-year age group x gender and the number of individuals per household, as estimated using the March 1995 Employment survey.

In wave 2, when moving and break-up variables are incorporated in the CNRHG, the recalibrated weights vary in a ratio of 1 to 12.3. The 1st to 99th percentiles ratio is 4.2; the 5th to 95th percentiles ratio falls to 2.2. The ratio of basic weights after and before calibration is on average 0.99; it varies from 0.90 to 1.10. The distribution after calibration is thus slightly more dispersed (the ratio of 1 to 12.3 comes from the maximum weight of a single individual).

3.2. Calculation of cross-sectional weights using the weight share method

Based on Eurostat recommendations, the weight share method was selected for calculating cross-sectional weights. This principle involves sharing the total of the adults' basic weights among all adult members of the household, whether or not they are panel individuals. All responding adult members of a household are thus assigned the same cross-sectional weight (Lavallée, 1995).

We can thus calculate as many cross-sectional weights as there are series of basic weights or longitudinal weights. In wave 2, the longitudinal weights are strongly related to the W1-W2 period. However, beginning in wave 3, two longitudinal weights can be calculated based on whether the non-response in W2 of individuals who respond in W3 is considered total or partial non-response. In the first case, a W1-W2-W3 longitudinal weight will be calculated for just panel individuals who respond to all three waves of the survey; in the second case, we can calculate a longitudinal weight for all W1 individuals responding in W3, whatever their response status in W2. Based on the choice made, individuals who respond in W3 but not in W2 will either have a longitudinal basic weight or not. A household in which all of panel individuals fall into this category therefore may or may not have a cross-sectional weight. So as not to lose from the analysis those households in which no individual responded in wave 2, the basic weights used to calculate cross-sectional weights in W3 were calculated by correcting W1 basic weights.

[3] Maximum value not reached per individual at maximum weight.

3.3. Household distributions of W2 cross-sectional weights

We calculated the cross-sectional weights before and after calibrating the W2 basic weights for the panel adults for each of the two series of longitudinal basic weights.

In the case of an adjustment without incorporating the behaviour variables within the CNRHG, the cross-sectional weights before calibration have a maximum ratio of 1 to 24.8. The ratio of the 1st to the 99th percentiles is thus 6.7, falling to 2.2 between the 5th and 95th percentiles. After calibration, the maximum ratio of the cross-sectional weights reaches 26.6, and the ratios of the 1st to 99th percentiles and 5th to 95th percentiles remain more or less the same (6.7 and 2.4). The ratio of pre- to post-calibration cross-sectional weights is on average 1.00, varying between 0.88 and 1.09.

In the other case, where the CNRHG incorporates behaviour variables, the cross-sectional weights are further dispersed before calibration, but the calibration changes them less and instead tends to reduce dispersion. Before calibration, the cross-sectional weights have a maximum ratio of 1 to 36.8; the 1st to 99th percentiles ratio is thus 6.1, with a 5th to 95th percentiles ratio of 2.3. After calibration the maximum ratio of cross-sectional weights is lower (35.7), and the 1st to 99th percentiles and 5th to 95th percentiles ratios remain the same (6.1 and 2.3).

4. EFFECT OF WEIGHTING ON SOME INDICATORS

4.1 Population with a basic weight strongly affected by whether or not behaviour variables (moving and breakup) are introduced

Individual basic weights are considered strongly affected by the introduction of household behaviour variables if they differ by more than 5%. 1.8% of adult panel individuals have a basic weight at least 5% higher when those variables are incorporated. On the contrary, 5.7% have a weight at least 5% lower.

In 8 out of 10 cases, individuals whose weights increase strongly are renters (1/3 of the total population of panel adults). Sixty-three percent (63%) are actively employed, 12% unemployed and 15% are students. Forty-two percent (42%) changed jobs between October 1994 and October 1995. Among those actively employed, 1/3 have fixed-term employment. They are younger than the rest of the population: 54% are under 25 and 32% between 25 and 35. These are basically single people (41%) and childless couples (37%). One-third have post-secondary education. They are over-represented in urban communities outside Paris.

The population with the lowest weight when behaviour variables are incorporated is both similar and different: fewer are actively employed (43%) but more are unemployed (8.5%) than our sample as a whole. Seventeen percent (17%) changed jobs between the two waves, fewer than the previous group, but more than our sample as a whole. Twenty-five percent (25%) of those actively employed have fixed-term contracts. This population is older than the preceding population, but are younger than our sample as a whole: 43% are under 25 and 15% between 25 and 35. These are basically single-parent families (19%) and blended families (34%). Twenty five percent (25%) are in training or going to school. Close to one in three live in Paris.

4.2 Cross-sectional indicators of standard of living and poverty

When looking at the population as a whole, the usual distribution indicators of standard of living and poverty indicators appear relatively unaffected by the weighting used (Table 1). Inequality and poverty seem slightly more significant when the moving variable in incorporated within the CNRHG, and when the sample is calibrated, but the differences remain relatively minor (1.4% for standard of living, 0.3 points for the poverty threshold), a rather reassuring observation.

However, when the indicators are declined based on age segments, or the type of household is taken into account, the differences are more marked. Based on the weighting system used, the estimated poverty rate for those aged 15-29 varies between 12.7% and 13.3%. Similarly, the estimated poverty rate for single people is strongly related to the set of weightings used. Differential attrition according to categories may thus have non-negligeable impacts on analyses of inter-category disparities.

Table 1: Some cross-sectional indicators of living standard(*) and poverty for individuals in 1994

	Population as a whole		Individuals between 15 and 29 years		Single persons	
	Average standard of living	Poverty rate	Average standard of living	Poverty rate	Average standard of living	Poverty rate
No behaviour variables (moving, break-up) for the CNRHG						
No calibration	103 717	8.8	92 558	12.7	91 801	17.1
Calibration	103 603	9.0	92 251	12.9	91 733	17.2
Behaviour variables (moving, break-up) incorporated for the CNRHG						
No calibration	103 421	9.0	91 670	13.1	90 771	17.8
Calibration	103 282	9.1	91 426	13.3	90 669	17.8

(*) Standard of living is defined as the ratio of total household income to number of consumption units (CU). The scale used to calculate CU gives a weight of 1 to the adult head of the household, 0.5 to other adults and 0.3 to children under 14 years of age. Individuals are defined 'poor' if their standard of living is lover than half the median standard. Estimation of the poverty threshold thus varies with the estimation of the median, and is conditional on the weighting system.

5. CONCLUSION

The choice of non-response model may strongly influence weight distributions. However, when the results of the analyses conducted are compared with the two series of weights from different models, overall the results are close. The weights constructed in each system diverge significantly for a few very specific segments of the population only. However, the statistics produced on these specific sub-populations appear to be affected by the choice of model. The observed variances are statistically non-significant, but in the absence of precise calculations of the estimates done, may result in contradictory interpretations. For example, assume that the 1993 poverty rate was 8.7%. Based on the weighting system selected, one can conclude stability (8.8%), or an increase (9.1%) in the percentage of poor people in the population.

REFERENCES

Chambaz C. and N. Legendre (1998). Calcul des pondérations dans le panel européen des ménages . *Journées de Méthodologie Statistique,* awaiting publication.

Deville J.-C. (1998). Les enquêtes par panel: en quoi diffèrent-elles des autres enquêtes? follow-up to Comment attraper une population en se servant d'une autre? *Journées de Méthodologie Statistique,* awaiting publication.

Eurostat (1995). Groupe de Travail Panel Communautaire de Ménages , Paris, September 18 and 19, 1995, Longitudinal Weighting . *Doc. PAN 51/95,* Eurostat, July.

Lavallée P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method, *Survey Methodology,* Vol. 21, No. 1, Statistics Canada, June 1995.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

A NEW LONGITUDINAL PROCESSING SYSTEM FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Pat Doyle[1]

ABSTRACT

The U.S. Bureau of the Census implemented major changes to the design of the Survey of Income and Program Participation (SIPP) with the panel begun in 1996. The revised survey design emphasized longitudinal applications and the Census Bureau attempted to understand and resolve the seam bias common to longitudinal surveys. In addition to the substantive and administrative redesign of the survey, the Census Bureau is improving the data processing procedures which yield microdata files for the public to analyze. The wave-by-wave data products are being edited and imputed with a longitudinal element rather than cross-sectionally, carrying forward information from a prior wave that is missing in the current wave. The longitudinal data products will be enhanced, both by the redesigned survey and new processing procedures. Simple methods of imputing data over time are being replaced with more sophisticated methods that do not attenuate seam bias. The longitudinal sample is expanding to include more observations which were nonrespondents in one or more waves. Longitudinal weights will be applied to the file to support person-based longitudinal analysis for calendar years or longer periods of time (up to four years).

Keywords: longitudinal surveys, post data collection processing.

1. INTRODUCTION

In conjunction with the conversion of the Survey of Income and Program Participation (SIPP) to Computer-Assisted Personal Interviewing, the Census Bureau directed several changes to the survey and its data processing to improve the data as a basis for longitudinal analysis. In particular, the post data collection processing system is being redesigned to take better advantage of the longitudinal features of the survey. The redesigned system will enhance the products we produce for both cross-sectional and longitudinal estimation.

In this paper we describe our plans for the new processing system. This report focuses on weighting, editing, and imputation procedures currently being designed and developed as well as issues of file format, layout, and access. In particular, we describe the universe for files, the weights, the scope of files, our approach to processing, and our strategies for item imputation and weighting.

An extended version of this paper will exist on the Census Bureau's SIPP web cite. It will represent our source of information for the public to anticipate the features of the longitudinal data products from SIPP. Users unfamiliar with SIPP are referred to that report and other documents on that cite for more information.

2. UNIVERSE FOR LONGITUDINAL DATA PRODUCTS

The longitudinal files produced from SIPP are complicated to understand and use because they contain more

[1] This report describes work of the SIPP longitudinal work group. It was compiled by Pat Doyle with assistance of the group at large. Contact Pat Doyle, U.S. Bureau of the Census, Demographic Surveys Division, Room 3376-3, Washington DC 20233, Phone: 301-457-3795, Fax: 301-457-2306, E-mail: patricia.j.doyle@ccmail.census.gov. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

observations than just those on which longitudinal estimation should be based. They contain records for persons in the longitudinal sample as well as other persons for whom data are collected in SIPP.

The Census Bureau typically provides its data users with data products that are "complete" (i.e., without missing data). In other words, we reduce the bias due to unit nonresponse through the use of adjusted weights and we assign imputed values to all unreported items. SIPP longitudinal data products will be created using the same philosophy although not all observations will be complete for the full reference period of a longitudinal panel. Also, the approach to developing longitudinal SIPP data products will be somewhat different from the approach used for most Census Bureau data products.

2.1 Longitudinal Sample

The 1996 panel longitudinal files will offer the opportunity to analyze data over the full panel or over calendar years within the time period of the full panel. To support these multiple reference periods for longitudinal analysis multiple weights will be produced, one for each period. Also, the longitudinal sample corresponding to those weights will vary depending on the reference period.

Regardless of the analytic period, the longitudinal sample will be defined as persons who receive positive longitudinal weights, i.e., those persons who meet one of the following conditions during the reference period for a particular weight:

> They are in the sample and with valid data for a fixed point in time (i.e., the initial cohort) and they continue to be interviewed during the analytic period for as long as they were in the SIPP universe
>
> They do not qualify under the first condition, they are in the initial cohort, they remain in the SIPP universe, and any noninterview or two consecutive noninterview waves are bounded by interview waves.

Military personnel residing exclusively with military personnel will be excluded from the set of persons with positive longitudinal weights.

2.2 Other People Represented in The Files

Only subsets of persons interviewed in SIPP get a positive longitudinal weight and, hence, form the longitudinal sample. However, the longitudinal files to be publicly disseminated from SIPP will contain information on all persons ever in the survey. In other words, at any point in time in the longitudinal file information exists on the longitudinal sample persons (e.g., their total income and age) as well as on persons with whom they reside. Hence, attributes of the sample persons' family, program units, and household can be constructed for analysis corresponding to any point in time during the reference period.

For cases with missing interviews (be they part of the longitudinal sample or not), we perform imputations for noninterviews occurring during the period they are in the SIPP universe. For some but not all of these, we perform longitudinal imputation for missing waves of data. Instances when we do impute the data longitudinally include: persons with bounded waves of missing data, and persons in households where other persons have bounded waves of missing data. Bounded waves of missing data are single waves or two consecutive waves of noninterview with an interview immediately before and after. For missing waves of data that do not qualify for longitudinal imputation, any information that was cross-sectionally imputed during the wave processing is carried forward to the longitudinal file.

3. SCOPE OF LONGITUDINAL FILES

The longitudinal files are flexible in the definition of the sample with which and the analytic period over which longitudinal analysis can be conducted. However, the data contained in the file, the period over which these data are processed, and the organization of these data are fixed.

3.1 Content

The longitudinal file reflected information from the SIPP core questionnaire, selected recodes designed to facilitate longitudinal analysis, and weights for use in estimation. The data file will cover the basic demographic, income, labor force, and program content included in the core. The extent to which we are able to preserve all of the details of the core will depend on physical constraints imposed by the available technologies for distributing data. Since we have not yet specified the medium for data distribution, those are constraints are unknown. We are interested in feed back on the file size and layout constraints faced in the world today and expected in the next few years.

We face additional constraints on content for two reasons. First, the potential exists for 52 replicates of each variable for the 48 months of the reference period plus the extra months needed to complete calendar years 1996 and 1999. Second is a large number of records (roughly 120,000) to be included in the final file. Because of these content constraints, the longitudinal data products from the early panels of SIPP added very few recodes designed to facilitate analysis. For example, earlier files reflected very few attributes of households and families and formatted the unearned income variables based on a generic presentation of unearned income, with only implicit identification of sources of income.

File size and content not withstanding, we are planning to expand the recodes for the 1996 panel files to include monthly attributes of families. Furthermore, we will add to the information on the file to make it easier for users to create their own family and household attributes. Our plans for identifying individual income sources are not yet final.

3.2 Organization

Ultimately the final longitudinal product from a SIPP panel is the full-panel file and it reflects core information covered throughout the life of the panel. However, we will also issue an interim longitudinal product for the 1996 panel covering calendar years 1996 and 1997 data from the first 7 waves of the 1996 panel.

While each longitudinal data product is one physical file, it is organized to support varying definitions of the longitudinal reference period. Specifically, it is organized to support longitudinal studies of events occurring during the full reference period covered by the file. It is also organized to support calendar year studies. For each period, the appropriate longitudinal sample consists of the group of persons with positive weights.

The data are organized by variable and within variable, by month (or wave). The months correspond to the months of the reference period (numbered consecutively) which is a varying calendar period depending on rotation group. Conversion of reference month replicates of each variable to calendar months is often needed for analysis but the conversion involves a straightforward calculation based on data in the file.

4. PROCESSING METHODS

With the 1996 Panel we are introducing a new approach to producing both the cross-sectional and longitudinal data products. The new approach for this panel changes not only the production of the longitudinal files but also the construction of the cross-section files.

4.1 Cross-Section Files

After Wave 1, the cross-section files will reflect longitudinal imputation although they will contain cross-sectional weights and reflect essentially a cross-sectional image of the U.S. In other words, we will impute for missing data in the current wave based on knowledge of events in the prior wave (i.e., forward imputation).

Some items in the cross-section files will be imputed using a longitudinal hot deck or direct substitutions methods described further in a subsequent section. Some items that are conditional on characteristics of the reference period will not be directly imputed because those characteristics (like number of weeks in each reference month of the wave) will vary by rotation group. Finally, to maintain correlation of information within each persons' record, some items will be derived from other data elements rather than imputed directly.

Because the procedure is designed to work for item nonresponse which can be any small subset of the total items, there are no conditional procedures embedded in the imputation logic to ensure that, for a given recipient in the hot deck, that the same donor will be used for all missing items. Hence, use of a derivation procedure (rather than an imputation procedure) is necessary to ensure that information within one record relates to other information in the same record in a plausible fashion. For example, we need to ensure that someone is working or not working but not both at a fixed point in time.

4.2. Longitudinal files

As with the earlier panels, the longitudinal data products from the 1996 panel will be generated from the cross-section files. Currently, our plans for the longitudinal files will parallel the methods used for the earlier panels, except they will require use of new longitudinal imputation methods discussed in more detail in the next section. For each item of missing data on each person, we will look to see if there is reported information in the subsequent wave (or waves, depending on the context). If so, we will determine whether to substitute the information imputed based on the prior wave, with a new information based on the subsequent waves.

However, we are still reviewing the interactions of the imputation process across modules. Because the longitudinal hot deck uses information from across the instrument to impute a given item, we need to consider the impact of changes in one module in the longitudinal processing may have on imputations of items in the other modules. For example, when considering an imputation of Wave t, if the wave 2 correlates (on which the cross section imputation was based) change during longitudinal imputation, do we still want to rely on prior wave 1 data as imputed on the cross section files. Would not we want to redo the imputation in light of the new values for the correlates, even though we might otherwise accept the forward imputation already performed. This situation suggests an iterative process of imputation but this approach is computationally intensive and, therefore, may not be feasible.

Instead of iteration we are considering the following approach: first establish an order in which to provide imputation that takes the major interdependencies into account, then, taking each section in the designated sequence, reimpute all the imputed data from the wave files. The approach for reimputation would be the same approach as used in the wave file processing with one exception. Rather than imputing information only from the prior wave, also allow the imputed value to be derived from a subsequent wave.

The method of choosing to replace a forward imputed item with a backward imputed item derived from reported data in a subsequent wave has not been firmly decided. At this time we are considering letting that choice be determined based on the potential of the subsequent wave to inform the imputation and on the potential for reducing the bias in spell lengths. Overall, however, we plan to do the following:

> If the missing information was not reported in the prior wave (and thus cross sectionally imputed in the current wave) then it will be imputed based on the reported information in the subsequent wave.
>
> If the missing information was not reported in either the prior or the subsequent wave it will remain cross-sectionally imputed as it is on the wave files.
>
> If the information was reported in both the prior and subsequent waves then we will potentially derive imputed estimates from both waves.

4.3 Unique Missing Data Problem

The 1996 panel presents a rather unique missing data problem which must be addressed either through imputation or reweighting. There was a U.S. government furlough that delayed the start of the 1996 panel from February to April resulting in a gap in 1996 calendar year data for half of the sample. 1996 is a critical year at least in the analysis of welfare reform in the U.S. since it represents a year of transition from an old to a new system of assisting our low income population. Therefore, we have decided to compensate for the missing data through imputation rather than reweighting so that we can keep the full longitudinal sample.

A comparable missing data problem will occur at the end of the 1996 panel because the Census Bureau has decided to end the last wave of the year 2000 panel with the January 2000 interview, thus missing end of calendar year 1999 data for 2 rotation groups. So that the full longitudinal sample can be used as a basis for estimation, we will impute the missing months of information at the end as well.

5. IMPUTATION METHODS

Here we summarize three types of longitudinal imputation strategies and how we plan to use these in carrying out the longitudinal imputation for both the cross-sectional and longitudinal data products.[2] The fuller version of this paper summarizes some literature on how well each if these imputation methods perform.

The first imputation method is **direct substitution** of information from either a prior or subsequent wave. There are several variations on this method: simple carry over, modified carryover, and random carry over. Methods vary in determining which information is actually used in the substitution. Simple carry over uses information reported for the last month of the prior wave to as the value to assign to one or more missing months in the current wave. We will use simple carry over for characteristics that should not change over time or change slowly (like the demographic information). Modified carry over methods use information directly from the prior or subsequent wave, the determination of which will be based on observed empirical distribution of the months at which the transition occurred. Random carry over uses one or the other wave depending on the outcome of a random draw from the uniform distribution.

We are also using a **longitudinal hot deck** procedure for our longitudinal imputation of items that can change relatively often within the reference period of the survey.. In general, a hot deck procedure is designed to locate a "donor" from among persons reporting a particular item and then using their data as the basis of the imputation. The determination of the "donor" varies by hot deck method as does the way in which the donor's information is used to determine the imputed amount.

The hot deck method considered for this study is referred to as a longitudinal hot deck because we will select the donor based on longitudinal rather than current-wave characteristics in the survey as follows:

> Lets say we have person x in wave t who is missing a key variable in wave t but not in wave t-1. To find a donor for person x, we first look back to wave t-1 to determine his/her characteristics in wave t-1, including the value of the missing variable of interest and other correlated variables. Using the values of those wave t-1 variables as dimensions of the hot deck, we find a match from among persons in Wave t-1 who (1) resemble the person in wave t-1 and (2) who have reported data for the variable of interest in waves t-1 and t. When the optimum match is found (based on what ever matching criteria is chosen for this hot deck), that person becomes the donor for variable of interest for person x in wave t.

When the donor is found, we can either impute the amount the donor reported in wave t or we can compute the amount for wave t from the amount in wave t-1 based on the observed change in values across waves in the donor's record. Once a donor is used, it is flagged and it is only used again if no other donors after it exist.

[2] Edits are performed as well but these are primarily for consistency within a wave. In general, we are not performing wave-to-wave consistency edits on the 1996 panel except to correct an error in reporting basic demographics discovered in later waves.

As just illustrated, donors can be chosen based on characteristics in the prior wave from among persons who report the missing item in the current wave. Similarly, we can apply the same technique to the find the donor in the subsequent wave from among the set of persons reporting the variable of interest in Waves t and t+1.

Other methods of longitudinal imputation use a model to predict current wave values based on prior wave values using **regression techniques**. Such models can be designed to predict the actual missing values or to predict change from another (typically prior) wave. Such models could be estimated directly from person-level data or aggregate data. We are not currently planning to use these methods but they are still on the table as options should they seem appropriate for a given item.

A special case of imputation is when we have an entire interview that needs to be imputed, i.e., a **missing wave.** As noted earlier, we include persons with missing waves when the missing wave is bounded by two reported interviews (that are at most two waves apart). Most but not all missing data in a missing wave record are imputed using the direct substitution method. The exceptions are household-level attributes which have been edited to consistency with other information from successful interviews in the household, information specific to a calendar month (such as the number of weeks in a month) or conditional on the length of the calendar month (such as weeks worked in that month).

With some exception, we randomly determine the transition point between months where data are derived from the prior wave and the subsequent wave. The exceptions relate to persons who move, where the transitions are tied to the move date itself, and to persons who attain the age of 15 in the course of the missing wave, where the transition is tied to the birthday. (More limited information is collected for persons under age 15.) Information collected by wave (rather than by month) is derived from the prior or subsequent wave depending on the outcome of a random draw.

6. WEIGHTING

The weights assigned to persons in a designated longitudinal sample (i.e., longitudinal weights) are the product of 3 components, an initial weight, a noninterview adjustment factor, and second stage adjustments. Multiple weights are assigned to a file, one for each longitudinal reference period defined for a given file (generally this is the full reference period of the file plus each calendar year period identifiable within that period). The computation of weights is identical across the reference periods but the starting values and subsequent adjustments vary because the composition of the longitudinal sample varies. The general procedures to compute the three components follow.

6.1 Initial Weight

Traditionally he initial weight represents the inverse of the probability of selection in the first wave of the longitudinal reference period (frequently Wave 1), adjusted for household nonresponse in that wave and for first stage factors appropriate for cross-sectional weighting (Kalton, 1998). For 1996, we are just using the initial weight adjusted for noninterview based on region, characteristics of area of residence, race, tenure, household size, and rotation group.

6.2 Post Wave 1 Noninterview Adjustment

Th initial weight is further adjusted for person-level noninterviews occurring after the initial wave (i.e, only those that are not otherwise compensated for through the missing-wave imputation process discussed earlier). This adjustment accounts for all persons who received a cross-sectional weight in the initial wave who do not receive a longitudinal weight for this reference period.

The person-level noninterview adjustment factors have varied over the tenure of SIPP in an attempt to better compensate for the nonrandom nonresponse that occurs after Wave 1 (Kalton, 1998). More recently the controls for this noninterview adjustment have been race and ethnicity, labor force activity (self-employed or not and in the labor force or not), household income, education of the reference person, welfare program participation, ownership of financial assets, and tenure. For 1996 we are introducing some new factors: within PSU stratum code, Census division, number of imputed items, and average monthly poverty ratio.

6.3 Second Stage Adjustments

The second stage adjustments are intended to bring the weighted estimates from SIPP in line with independent estimates. Specifically, the SIPP longitudinal weights are controlled to Current Population Survey estimates of households and independent estimates of the size of the population of Hispanic Origin. The process differs slightly for persons age 15 and over versus persons under age 15. The adjustment process is controlled so that the final weights are not extremely large or extremely small and so that the adjustments are based on data on population groups derived from cell sizes of at least 30 unweighted cases.

REFERENCES

Kalton, Graham (ed) "SIPP Quality Profile 3rd Edition." Washington DC: U.S. Bureau of the Census, 1998.

SESSION 4

GROSS FLOWS I

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

ESTIMATION OF GROSS FLOWS FROM COMPLEX SURVEYS ADJUSTING FOR MISSING DATA, CLASSIFICATION ERRORS AND INFORMATIVE SAMPLING

Danny Pfeffermann[1] and Natalia Tsibel[1]

ABSTRACT

This article extends and further develops the method proposed by Pfeffermann, Skinner and Humphreys (1998) for the estimation of gross flows in the presence of classification errors. The main feature of that method is the use of auxiliary information at the individual level which circumvents the need for validation data for estimating the misclassification rates. The new developments in this article are the establishment of conditions for model identification, a study of the properties of a model goodness of fit statistic and modifications to the sample likelihood to account for missing data and informative sampling. The new developments are illustrated by a small Monte-Carlo simulation study.

KEY WORDS: Goodness of fit, informative sampling, missing data, parameter identifiability.

1. INTRODUCTION

Gross flows estimates are often produced from longitudinal data observed over several time periods. A familiar problem with the use of this kind of data is the existence of classification errors that are known to bias the gross flows estimators. Many of the methods proposed in the past to deal with this problem assume the existence of validation data obtained from re-interview surveys, which permits estimation of the misclassification rates. See, for example, the articles by Abowd and Zellner (1985), Poterba and Summers (1986), Chua and Fuller (1987) and Singh and Rao (1995). All these articles consider the estimation of gross flows between states of labor markets, using data collected from Labor Force Surveys (LFS). The present study uses a similar framework.

The major problem with the use of methods that rely on the existence of validation data is that very often such data are not available, or that they are not suitable for adjusting the gross flows estimates, (Forsman and Schreiner, 1991). This lead various researchers to consider models that relate the true states and the observed states at different occasions, see, e.g., Van de Pol and Langeheine (1990), Vermunt (1996) and Humphreys and Skinner (1997). All these studies assume that data are available in the form of contingency tables.

In a recent article, Pfeffermann, Skinner and Humphreys (1998, hereafter PSH) propose a different approach for adjusting flow estimates for bias. The proposed approach is based on fitting separate multinomial logistic models to the true state transition probabilities and the classification error probabilities, using values of auxiliary variables at the individual level. The notable advantage of this approach is that it does not require validation data for the estimation of the misclassification rates. It permits the identification of factors that are related to the misclassification and true transition rates and it

[1] Danny Pfeffermann and Natalia Tsibel, Department of Statistics, Hebrew University, Jerusalem, Israel, 91905.

does not require independence between the classification errors at different time points, an assumption underlying the other methods mentioned before. The major disadvantage of the proposed approach is that it requires the estimation of many more parameters. Another issue is the robustness of the flow estimates to possible departures from the postulated working model, given that the latter refers to the unobservable true states and classification errors.

In this article we further develop this approach focusing on four main topics:

1- We establish necessary conditions for the identification of the model fitted to the sample measurements,
2- We derive the sample likelihood for the case of randomly missing data,
3- We propose modifications to the likelihood that account for informative sampling schemes,
4- We establish the asymptotic distribution of a goodness of fit statistic employed in PSH.

Section 2 reviews briefly the models and estimation procedures employed by PSH. Sections 3-6 discuss the four extensions listed above and Section 7 presents simulation results illustrating some of the developments of the previous sections.

2. THE PSH APPROACH

The method proposed by PSH consists of fitting separate multinomial logistic models at the unit level for the true state transition probabilities and the misclassification probabilities, using values of auxiliary variables. Let Y_{kt} denote the true state of individual k at time t, taking values $1 \ldots L$, and let y_{kt} denote the corresponding observed state. The covariate variables values associated with unit k are denoted by x_{kt}. Notice that different covariates may be included in different models, (see Section 3), but for expository purposes we use in this section the uniform notation x_{kt} for all the models.

A- Model for Initial True States

$$\pi_l^{k1} = \Pr(Y_{k1} = l \mid x_{k1}) = \exp(x'_{k1}\beta_l)/\{1+\textstyle\sum_{j=1}^{L-1}\exp(x'_{k1}\beta_j)\},\ l = 1 \ldots L-1, \quad \pi_L^{k1} = 1 - \sum_{l=1}^{L-1}\pi_l^{k1} \qquad (2.1)$$

B- Model for True Transition Probabilities

$$\pi_{ij}^{kt} = \Pr(Y_{kt} = j \mid Y_{k,t-1} = i, x_{kt}) = \exp(x'_{kt}\gamma_{ij})/\{1+\textstyle\sum_{m=1}^{L-1}\exp(x'_{kt}\gamma_{im})\},\ i = 1 \ldots L,\ j = 1 \ldots L-1,$$

$$\pi_{iL}^{kt} = 1 - \textstyle\sum_{j=1}^{L-1}\pi_{ij}^{kt} \qquad (2.2)$$

C- Model for Observed States

$$q_{li}^{kt} = \Pr(y_{kt} = i \mid Y_{kt} = l, x_{kt}) = \exp(x'_{kt}\alpha_{li})/\{1+\textstyle\sum_{j=1}^{L-1}\exp(x'_{kt}\alpha_{lj})\},\ l = 1 \ldots L,\ i = 1 \ldots L-1,$$

$$q_{lL}^{kt} = 1 - \textstyle\sum_{i=1}^{L-1} q_{li}^{kt} \qquad (2.3)$$

The model defined by (2.3), but not the models defined by (2.1) and (2.2) may include the previously observed state among its covariates.

Assuming that the sampling design used for the sample selection is non-informative so that the model defined by (2.1) - (2.3) holds for the sample data, and that observations y_{kt} are obtained for $t = 1 \ldots T$ time points independently between units k, the sample log-likelihood is obtained as $\log(l) = \sum_{k=1}^{n}\log[l(k)]$ where

$$l(k) = \Pr(y_{k1} = i_{k1} \ldots y_{kT} = i_{kT}) = \sum_{l_1=1}^{L} \cdots \sum_{l_T=1}^{L}\Big[\prod_{t=2}^{T} q_{l_t i_{kt}}^{kt}\,\pi_{l_{t-1}l_t}^{kt}\Big] q_{l_1 i_{k1}}^{k1}\,\pi_{l_1}^{k1} \qquad (2.4)$$

and the summation $\Sigma_{l_1=1}^{L} \ldots \Sigma_{l_T=1}^{L}$ is over all possible realizations of the true states $Y_{k1} \ldots Y_{kT}$. Notice that the construction of the likelihood imposes a Markovian structure on the observed and true state probabilities. See Equations (2.2a)-(2.2c) in PSH.

The likelihood defined by (2.4) depends on models for unobservable components which raises the question of the robustness of the estimates to possible model misspecification. As a partial solution to this problem, PSH propose to maximize the likelihood (2.4) subject to the constraints that the predicted (expected) true stocks (marginal proportions) at each time point under the model agree with the corresponding observed stocks. These constraints are based on the proposition that the observed stocks are approximately unbiased for the true stocks, even in the presence of misclassification, a proposition underlying the other methods mentioned in the Introduction. (See e.g., Singh and Rao, 1995 for discussion.) Simulation results reported in PSH indicate good performance of the constrained estimators. An alternative procedure proposed to us by W. Fuller (private communication) is to constrain the predicted true stocks to agree with the expected (predicted) observed stocks under the model. The resulting model is known in the literature as the *unbiased response error model*, Fuller (1987, Section 3.4.2). The latter constraints have the advantage of not depending on sample statistics but we surmise that by constraining the predicted true stocks to agree with the observed stocks, the resulting predictors are more robust to possible model misspecifications.

3. PARAMETER IDENTIFIABILITY

A fundamental question underlying the estimation of the vector parameters $\{\beta_l\}, \{\gamma_{ij}\}$ and $\{\alpha_{li}\}$ by maximization of the likelihood (2.4) is whether there is more than one global maximum. When this is the case, the vector parameters (or some of them) are no longer *identifiable*, (Kendall and Stuart, 1979, Vol. 2, p.44). In this section we define necessary conditions for parameter identifiability under the model defined by (2.1)-(2.3).

Let $\tilde{X}_j(\pi)$ denote the X-variables included in the logistic models (2.2). (We assume the same variables $\tilde{X}_j(\pi)$ for all *j=1...L*.) Similarly, let $\tilde{X}_l(q)$ denote the X-variables included in the logistic model (2.3), (assumed to be the same for all *i*).

THEOREM 1: The parameters $\{\beta_l\}, \{\gamma_{ij}\}$ and $\{\alpha_{li}\}$ are not identifiable under either one of the following conditions:

1- $\tilde{X}_1(\pi) \equiv \ldots \equiv \tilde{X}_L(\pi)$ and $\tilde{X}_1(q) \equiv \ldots \equiv \tilde{X}_L(q)$,

2- The classification probabilities are independent of the true states; $q_{1i}^{kt} = \ldots q_{Li}^{kt} = q_i^{kt}$ for all $i = 1 \ldots L, t = 1 \ldots T$,

3- Observations are available for only $T \leq 2$ time points, (at most one set of transitions).

<u>*Proof of 1*:</u> Here we illustrate the non-identifiability of the parameters by considering the case *L=2* and commenting on the case *L=3*. When *L=2* there are five unknown vector parameters, $\beta_1, \gamma_{11}, \gamma_{21}, \alpha_{11}, \alpha_{21}$. Consider the following transformation,

$$\beta_1 \mapsto -\beta_1 \ , \ \gamma_{11} \mapsto -\gamma_{21} \ , \ \gamma_{21} \mapsto -\gamma_{11} \ , \ \alpha_{11} \mapsto \alpha_{21} \ , \ \alpha_{21} \mapsto \alpha_{11} \qquad (3.1)$$

which corresponds to permutation of the true states such that state 1 becomes state 2 and vice versa. For example, by (2.2), $\pi_{11}^{k2} = \exp(x_{k2}'\gamma_{11})/\{1+\exp(x_{k2}'\gamma_{11})\}$ and $\pi_{21}^{k2} = \exp(x_{k2}'\gamma_{21})/\{1+\exp(x_{k2}'\gamma_{21})\}$. By changing $\gamma_{11} \leftrightarrow -\gamma_{21}$, the new state transition probabilities, assuming $\tilde{X}_1(\pi) \equiv \tilde{X}_2(\pi)$ are, $\vec{\pi}_{11}^{k2} = \exp(-x_{k2}'\gamma_{21})/\{1+\exp(-x_{k2}'\gamma_{21})\} = \pi_{22}^{k2}$ and $\vec{\pi}_{21}^{k2} = \exp(-x_{k2}'\gamma_{11})/\{1+\exp(-x_{k2}'\gamma_{11})\} = \pi_{12}$. We thus have the following changes in the model probabilities implied by the transformation (3.1),

$$\vec{\pi}_1^{k1} = \pi_2^{k1} , \vec{\pi}_{11}^{kt} = \pi_{22}^{kt} , \vec{\pi}_{21}^{kt} = \pi_{12}^{kt} , \vec{q}_{11}^{kt} = q_{21}^{kt} , \vec{q}_{21}^{kt} = q_{11}^{kt} . \qquad (3.2)$$

Next we show that the likelihood contributions $l(k)$ defined by (2.4), and hence the log-likelihood $\log(l) = \Sigma_{k=1}^{n} \log[l(k)]$ are invariant to the parameter transformation (3.1). To simplify the exposition we suppose $T=2$ and $i_1=i_2=1$. (The same is true for general T and any realization $i_1 \ldots i_T$ of the observed states.) Before the transformation,

$$l(k) = \pi_1^{k1} q_{11}^{k1} \pi_{11}^{k2} q_{11}^{k2} + \pi_1^{k1} q_{11}^{k1} \pi_{12}^{k2} q_{21}^{k2} + \pi_2^{k1} q_{21}^{k1} \pi_{21}^{k2} q_{11}^{k2} + \pi_2^{k1} q_{21}^{k1} \pi_{22}^{k2} q_{21}^{k2}\,. \qquad (3.3)$$

After the transformation,

$$l(k) = \pi_1^{k1} q_{11}^{k1} \vec{\pi}_{11}^{k2} q_{11}^{k2} + \vec{\pi}_1^{k1} q_{11}^{k1} \vec{\pi}_{12}^{k2} \vec{q}_{21}^{k2} + \pi_2^{k1} \vec{q}_{21}^{k1} \vec{\pi}_{21}^{k2} \vec{q}_{11}^{k2} + \vec{\pi}_2^{k1} \vec{q}_{21}^{k1} \vec{\pi}_{22}^{k2} \vec{q}_{21}^{k2} \qquad (3.4)$$

which by virtue of (3.2) is the same as (3.3). Since the two likelihood functions are the same, there are clearly two global maximums.

When $L=3$, there are six different sets of parameter values corresponding to the six possible permutations of the true states, all yielding the same likelihood. Consider, for example, the 'clockwise permutation' $1 \mapsto 2, 2 \mapsto 3, 3 \mapsto 1$ such that $\vec{\pi}_1^{k1} = \pi_3^{k1}, \vec{\pi}_2^{k1} = \pi_1^{k1}, \ldots, \vec{\pi}_{12}^{kt} = \pi_{31}^{kt}, \vec{\pi}_{23}^{kt} = \pi_{12}^{kt}, \ldots, \vec{q}_{32}^{kt} = q_{22}^{kt}, \vec{q}_{33}^{kt} = q_{23}^{kt}$. As in the case $L=2$, if Condition 1 is satisfied these changes in the model probabilities correspond to a simple linear transformation of the vector parameters, leaving the likelihood function intact.

Proof of 2: Under Condition 2,

$$l(k) = \Sigma_{l_1=1}^{L} \ldots \Sigma_{l_T=1}^{L} [\prod_{t=2}^{T} q_{l_t i_{kt}}^{kt} \pi_{l_{t-1} l_t}^{kt}] q_{l_1 i_{k1}}^{k1} \pi_{l_1}^{k1} = (\prod_{t=1}^{T} q_{i_{kt}}^{kt}) \Sigma_{l_1=1}^{L} \ldots \Sigma_{l_T=1}^{L} [\prod_{t=2}^{T} \pi_{l_{t-1} l_t}^{kt}] \pi_{l_1}^{k1} = (\prod_{t=1}^{T} q_{i_{kt}}^{kt})$$

since $\Sigma_{l_1=1}^{L} \ldots \Sigma_{l_T=1}^{L} \prod_{t=2}^{T} [\pi_{l_{t-1} l_t}^{kt}] \pi_{l_1}^{k1} = \Sigma_{l_1=1}^{L} \ldots \Sigma_{l_T=1}^{L} \Pr(Y_{k1} = l_1 \ldots Y_{kT} = l_T) = 1$. Thus, the likelihood does not depend on the parameters $\{\beta_l\}$ and $\{\gamma_{ij}\}$ which therefore are obviously not identifiable.

Proof of 3: Suppose first that the parameters $\{\beta_l\}, \{\gamma_{ij}\}$ and $\{\alpha_{li}\}$ consist only of intercept terms. The number of unknown parameters is then $(L-1)+2L(L-1)=(L-1)(2L+1)$. On the other hand, with $T=2$ and fixed sample size n, there are only $(L-1)+L(L-1)=(L-1)(L+1)$ sufficient statistics, (assuming 'full data' in the sense that observations are available for all possible transitions), such that the model is not identifiable. When $T=3$ and the data is full, the number of sufficient statistics is $(L-1)(L+1)+ L^2(L-1)=(L-1)(L^2+L+1)$ which is larger than the number of parameters.

In the more general case where some or all of the vector parameters are of dimension larger than 1, the total number of vectors is $(L-1)(2L+1)$ so that assuming sufficient observations for each transition, all the vector parameters are estimable based on only $T=3$ time points.

Further discussion

A simple solution to the parameter identifiability problem is to impose different regression variables for the logistic models in (2.2) and/or (2.3). This, however, is not always possible as for example in the case where the various models only contain intercept terms. An alternative solution is suggested by the fact that whereas permutation of the true states under the first condition leaves the predicted observed stocks and flows unaffected, it does affect the predicted true stocks and flows. (See PSH for the computation of the respective predictions under the model.) Thus, the alternative solution is to consider all possible permutations of true states and select the one that yields predicted true stocks that are most consistent with "prior knowledge". For example, if it is believed that the true stocks in the various time points should follow a similar order to that of the observed stocks, the permutation of true states that yields orders of the true stocks that are closest to the orders of the observed stocks should be selected. We mention in this respect the constrained maximization of the likelihood referred to at the end of Section 2. Constraining the parameter values to yield predicted true stocks that coincide with the corresponding observed stocks or predicted observed stocks automatically defines a single choice for the parameter values and hence a single set of predicted true stocks and flows..

4. THE LIKELIHOOD IN THE CASE OF RANDOMLY MISSING DATA

In this section we consider situations where classifications of units are missing "completely at random" for some time points. Such situations occur for example when the data are collected in surveys that use rotating panel sampling schemes. For instance, the U.S. LFS uses a rotating scheme of 4 months in sample, 8 months out of the sample and then 4 more months in the sample. In Israel the sampling scheme is 2 quarters in, 2 quarters out and then 2 quarters in again. As shown below, this kind of missing data can easily be handled by an appropriate integration of the "complete data" likelihood.

Let $T_{mis}^{k}=\{t \mid i_{kt} \text{ missing}\}$ define the time points for which the classification of unit k is not observed. In what follows we suppose that the explanatory variables are known for all *t=1...T*. This will always be the case with variables that are fixed over time or that change systematically, like Age for example. Some extrapolations may be needed in practice for other variables. Denote, as before, by $l(k)=\Pr(y_{kt}=i_{kt}, t \notin T_{mis}^{k})$ the likelihood contribution by unit k. With data missing completely at random, the likelihood *l(k)* is calculated as

$$l(k)=\sum_{l_1=1}^{L}\cdots\sum_{l_T=1}^{L}\{\sum_{i_{kt}=1,t\in T_{mis}^{k}}^{L}[(\prod_{t=2}^{T} q_{l_t i_{kt}}^{kt}\pi_{l_{t-1}l_t}^{kt})q_{l_1 i_{k1}}^{k1}\pi_{l_1}^{k1}]\}\,. \tag{4.1}$$

The sample log-likelihood is obtained as $\log(l)=\sum_{k=1}^{n}\log[l(k)]$, similarly to the complete data case.

COMMENT: A much more difficult problem occurs when the missing data process is informative like for example when the probabilities of response depend on the true states in some unknown way. This problem is not considered in the present article.

5. INFORMATIVE SAMPLING

The likelihood functions defined by (2.4) and (4.1) assume a non-informative sampling scheme. Suppose, however, that units are selected to the sample with unequal selection probabilities that depend on the true or reported classifications at the first time point that they join the sample. Clearly, ignoring the sampling process in such cases may bias the parameter estimators and hence the prediction of the true flows. A possible way to deal with this problem is by use of the pseudo likelihood (PL) approach. The approach consists of estimating consistently the 'census likelihood equations', (the equations that would have been obtained in the case of a census), and solving the estimating equations. In what follows we describe the approach in some more detail with reference to our model and show how to estimate the variances of the resulting estimators. For further discussion on the use of the PL approach with many examples, see Binder(1983) and Skinner, Holt and Smith(1989).

Let $\log[L_U(\theta)]=\sum_{i=1}^{N}\log[l(i,\theta)]$ denote the census log-likelihood function indexed by the vector parameter $\theta=\{\beta_l\},\{\gamma_{ij}\},\{\alpha_{li}\}$. The census maximum likelihood estimator (mle), $\hat{\theta}_c$, maximizes $\log[L_U(\theta)]$ and in regular cases solves the likelihood equations,

$$U(\theta)=\sum_{i=1}^{N}u_i(\theta)=\sum_{i=1}^{N}\frac{\partial\log[l(i,\theta)]}{\partial\theta}=0\,. \tag{5.1}$$

The pseudo mle (pmle) is defined as the solution $\hat{U}(\theta)=0$ where $\hat{U}(\theta)$ is design consistent for $U(\theta)$. The estimator $\hat{U}(\theta)$ in common use is the Horvitz-Thompson estimator obtained by weighting each sample score $u_i(\theta)$ by the weight $w_i=1/\Pr(i\in s)$ so that $\hat{\theta}_{pmle}$ is the solution of

$$\hat{U}_{H-T}(\theta)=\sum_{i=1}^{n}w_i u_i(\theta)=0\,. \tag{5.2}$$

The estimator $\hat{\theta}_{pmle}$ is generally design consistent for $\hat{\theta}_c$ which is consistent for θ under the model.

The variance of $\hat{\theta}_{pmle}$ cannot be estimated using large sample likelihood theory. Instead, we follow Binder (1983) and extract an estimate for the design variance, $\hat{V}_D(\hat{\theta}_{pmle})$ from $\hat{V}_D[\hat{U}_{H-T}(\hat{\theta}_c)]$ as follows: Expand, $\hat{U}_{H-T}(\hat{\theta}_{pmle}) = 0 \cong \hat{U}_{H-T}(\hat{\theta}_c) + \frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}(\hat{\theta}_{pmle} - \hat{\theta}_c)$, such that $\hat{U}_{H-T}(\hat{\theta}_c) \cong -\frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}(\hat{\theta}_{pmle} - \hat{\theta}_c)$ and $V_D[\hat{U}_{H-T}(\hat{\theta}_c)] \cong [\frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}] \times V_D(\hat{\theta}_{pmle}) \times [\frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}]^T$. Changing sides,

$$V_D(\hat{\theta}_{pmle}) \cong \{\frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}\}^{-1} \times V_D[\hat{U}_{H-T}(\hat{\theta}_c)] \times \{[\frac{\partial \hat{U}_{H-T}(\theta)}{\partial\theta \mid \theta = \hat{\theta}_c}]^T\}^{-1}. \quad (5.3)$$

The estimator $\hat{V}_D(\hat{\theta}_{pmle})$ is obtained by replacing $\hat{\theta}_c$ by $\hat{\theta}_{pmle}$ in the right hand side of (5.3). Notice that for large populations and small sampling fractions, $\hat{V}_D(\hat{\theta}_{pmle})$ estimates also the "total variance" over all possible population (census) values and all possible sample selections.

6. A MODEL GOODNESS OF FIT STATISTIC

PSH use the following measures to assess the goodness of fit of models fitted to the sample data,

$$\delta_t = 1 - \frac{1}{2n}\sum_{k=1}^{n}\sum_{i=1}^{L} | I(y_{kt} = i) - P_{kt}^i |; \quad t = 1 \ldots T \quad (6.1)$$

where $I(y_{kt} = i)$ is the indicator function and $P_{kt}^i = \Pr(y_{kt} = i \mid x_{kt})$. The measure δ_t compares the observed states at time t with the probabilities to observe these states under the model. A similar measure can be defined to compare the observed transitions with their estimated probabilities. The measure δ_t has the following properties:

P1- $0 \le \delta_t \le 1$; P2- $\delta_t = 1 \Leftrightarrow \forall k,\ \mathrm{y}_{\mathrm{kt}} = i \Leftrightarrow P_{kt}^i = 1$; P3- $\delta_t = 0 \Leftrightarrow \forall k,\ \mathrm{y}_{\mathrm{kt}} = i \Rightarrow P_{kt}^i = 0$;
P4- $P_{kt}^i \equiv \frac{1}{L} \Rightarrow \delta_t = \frac{1}{L}$; P5- $P_{kt}^i = \frac{n(t,i)}{\sum_{i=1}^{L} n(t,i)} \quad \delta_t = \frac{\sum_{i=1}^{L} n^2(t,i)}{[\sum_{i=1}^{L} n(t,i)]^2} \ge \frac{1}{L}$; $n(t,i) = \sum_{k=1}^{n} I(y_{kt} = i)$.

The properties P1- P5 follow straightforwardly from the alternative expression,

$$\delta_t = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{L}[I(y_{kt} = i)P_{kt}^i] \quad (6.1a)$$

Written this way, δ_t is seen to represent the "average probability" of the observed states.

Another desirable property of the measure δ_t is that it is possible to extract its asymptotic distribution under a given model. This permits the construction of confidence intervals for its expectation. Define $Q_{kt} = \sum_{i=1}^{L} I(y_{kt} = i)P_{kt}^i$ such that $\delta_t = \frac{1}{n}\sum_{k=1}^{n} Q_{kt}$. The random variables (rv) Q_{kt} take the values $P_{kt}^1 \ldots P_{kt}^L$ with probabilities $P_{kt}^1 \ldots P_{kt}^L$. Hence,

$$E(Q_{kt}) = \sum_{i=1}^{L}(P_{kt}^i)^2 = \mu_{kt} \ ; \ \mathrm{Var}(\mathrm{Q}_{\mathrm{kt}}) = \sum_{i=1}^{L}(P_{kt}^i)^3 - [\sum_{i=1}^{L}(P_{kt}^i)^2]^2 = \sigma_{kt}^2 \quad (6.2)$$

Also, $Q_{1t}...Q_{nt}$ are uniformly bounded, independent, and $V_n^2 = Var(\sum_{k=1}^n Q_{kt}) \xrightarrow[n\to\infty]{} \infty$. Thus, $[\delta_t - \frac{1}{n}\sum_{k=1}^n \mu_{kt}]/[V_n/n]$ is distributed asymptotically as *N(0,1), t=1...T.* We mention that it is also simple to calculate $Cov(\delta_t, \delta_\tau)$ *for* $t \neq \tau$ and hence the asymptotic joint distribution of $\delta_{(T)} = (\delta_1 ... \delta_T)$.

In a recent article, Estrella (1998) proposes a new goodness of fit measure for dichotomous dependent variables, defined as a function of the log-likelihoods under the full model and the model with an intercept term only. We intend to compare the properties of this measure and δ_t in the near future.

7. MONTE-CARLO SIMULATION RESULTS

In order to illustrate some of the new developments of this article, we simulated population measurements for T=4 time points from a reduced model fitted previously to data on employment status of heads of households, collected by the Israel LFS. For further simplification we classified the employment status into the states of "employed" (E) and "Other" (O). Sample data for samples of size *n=1500* were obtained in two different ways: A- By simulating directly from the population model which corresponds to noninformative sampling from "large populations" and B- By generating a (single) population of *N=6000* values and selecting simple stratified samples with the strata defined by the "observed" employment status at time *t=1*. The sampling proportions in the two strata are n_E/N_E*=750/4269* and n_O/N_O*=750/1731*. This sampling process is informative.

The model used for the simulations is the same as defined by (2.1)-(2.3) (with L=2) and is defined by the first three columns of Table 1. The explanatory variables are defined as follows where we denote by *I(.)* the indicator function: Intercept(Int), Age, Education(Ed)=*I(more than 12 years at school),* Child=*I(has children aged 5-14),* Married(Mar)=*I(married),* Self=*I(self supplied information),* Method(Met)=*I(personal interview).*

TABLE 1. *Simulation Results: Coefficient Estimates, SD's and SD Estimates.*

Population model			Noninformative Sampling				Informative Sampling			
Model Eq's	Coeff.	Val.	Means mle	SD mle	M(SDE) inf. mat	M(SDE) design	Means "mle"	Means pmle	SD pmle	M(SDE) design
q_{11}^{kt}	Int	1.5	1.50	0.08	0.07	0.07	1.32	1.52	0.08	0.07
	Self	3.2	3.30	0.54	0.57	0.59	3.51	3.39	0.41	0.39
q_{21}^{kt}	Int	-0.9	-0.89	0.09	0.12	0.12	-0.88	-0.90	0.11	0.08
	Met	-2.2	-2.26	0.57	0.65	0.68	-2.69	-1.97	0.62	0.84
π_{11}^{kt}	Int	0.8	0.73	0.64	0.61	0.64	0.47	0.70	0.60	0.61
	Mar	1.7	1.72	0.27	0.29	0.30	1.95	1.73	0.33	0.32
	Age	2.6	2.81	1.53	1.35	1.42	3.41	2.85	1.21	1.32
π_{21}^{kt}	Int	-0.9	-0.97	0.63	0.76	0.85	-0.83	-0.86	0.58	0.43
	Child	1.8	1.82	0.32	0.33	0.36	1.74	1.85	0.25	0.19
	Age	-2.9	-2.80	1.38	1.65	1.86	-2.65	-3.16	1.20	0.90
π_1^{k1}	Int	0.7	0.69	0.08	0.08	0.08	0.13	0.68	0.05	0.05
	Ed	3.4	3.50	0.51	0.62	0.63	3.08	3.45	0.45	0.32
	δ_4(u)	0.65	0.65	0.01			0.61	0.65	0.01	
	δ_4(c)	0.75	0.75	0.01			0.73	0.75	0.01	

Table 1 contains summary statistics computed over 100 samples selected for each sampling process. The columns entitled "means" show averages of the coefficient estimates. The columns entitled SD show the corresponding standard deviations. The columns entitled M(SDE) show averages of the SD estimates with "inf.mat" denoting estimates obtained from the inverse information matrix and "Design" denoting the design variances obtained from (5.3). The last 2 rows show the means and SD of the goodness of fit measures δ_4, defined by (6.1). We distinguish between the case where the probabilities $\Pr(y_{k4} = i)$ are calculated as $P^i_{k4} = \Pr(y_{k4} = i \mid x_{k4})$, for which we denote the measure by $\delta_4(u)$, and the case where these probabilities are calculated as $\tilde{P}^i_{k4} = \Pr(y_{k4} = i \mid x_{k4}; y_{k1}, y_{k2}, y_{k3})$, for which we denote the measure by $\delta_4(c)$. Both sets of probabilities are easily calculated from (2.4).

The results emerging from Table 1 can be summarized as follows:

1- For the case of noninformative sampling, the mle's of the model coefficients are essentially unbiased, with all the means being within two standard errors of the true values. Notice the large SD's of some of the coefficient estimates, indicating the high degree of variability in the population data. The two methods used to estimate the SD's yield similar estimates which in most cases are close, in average, to the empirical SD's. The goodness of fit measure $\delta_4(c)$ is seen to be significantly higher than $\delta_4(u)$, illustrating the much-improved predictions of the observed classifications at time t=4 when conditioning on the previously observed states. We mention in this respect that when fitting to the same samples the model with only intercept terms, the mean measures are $\bar{\delta}_4(u) = 0.63$ and $\bar{\delta}_4(c) = 0.72$, suggesting that for these data the inclusion of the covariates improves the prediction power only marginally. On the other, the mean log-likelihood values computed for the two models are –2627.7 for the full model and –2944.1 for the model with intercepts only, indicating the high significance of the covariates.

2- For the case of informative sampling, the "mle's" obtained by maximizing the unweighted likelihood are highly biased for many of the coefficients, a direct effect of the sample selection bias. These biases are largely reduced by use of the pmle's that maximize the probability weighted likelihood equations, with almost all the means being either within or on the border of two standard errors of the true values. The design variance estimates again perform well on average in at least most of the cases. The weighted measures of goodness of fit, obtained by weighting each of the quantities $Q_{kt} = \sum_{i=1}^{L} I(y_{kt} = i)\hat{P}^i_{kt}$ (equation 6.1a) inversely proportional to the unit's selection probability are similar to the corresponding unweighted measures obtained for the case of noninformative sampling, and they are unbiased for the corresponding measures computed for the population data.

REFERENCES

Abowd, J.M. and Zellner, A. (1985), Estimating Gross Labor Force Flows. *Journal of Business and Economic Statistics*, 3, 254-283.

Binder, D.A. (1983), On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-292.

Chua, T. and Fuller, W.A. (1987), A Model for Multinomial Response Error Applied to Labor Flows. *Journal of the American Statistical Association*, 82, 46-51.

Estrella, A. (1998), A New Measure of Fit for Equations with Dichotomous Dependent Variables. *Journal of Business and Economic Statistics*, 16, 198-205.

Forsman, G. and Schreiner, I. (1991), The Design and Analysis of Re-interview: An Overview. In *Measurement Errors in Surveys* (eds. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman). New York: Wiley.

Fuller, W.A. (1987), *Measurement Error Models*. New York: Wiley.

Humphreys, K. and Skinner, C.J. (1997), Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error. *Survey Methodology*, 23, 53-60.

Kendall, M and Stuart A. (1972), *The Advanced Theory of Statistics*. London: Charles Griffin and Company.

Pfeffermann, D., Skinner, S. and Humphreys, K. (1998), The Estimation of Gross Flows in the Presence of Measurement Error Using Auxiliary Variables. *Journal of the Royal Statistical Society*, Series A, 161, 73-82.

Poterba, J.M. and Summers, L.H. (1986), Reporting Errors and Labor Market Dynamics. E*conometrica*, 54, 1319-1338.

Singh, A.C. and Rao, J.N.K. (1995), On the Adjustment of Gross Flow Estimates for Classification Error with Application to Data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989), *Analysis of Complex Surveys*. Chichester: Wiley.

Van de Pol, F. and Langeheine, R. (1990), Mixed Markov Latent Class Models. In *Sociological Methodology 1990* (ed. C.C. Clogg), pp. 213-247. Oxford: Basil Blackwell.

Vermunt, J.K. (1996), *Log-Linear Event History Analysis*. Tilburg: Tilburg University Press.

Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

MOBILITY MATRICES AND COMPUTATION OF ASSOCIATED PRECISION

Nathalie Caron[1] and Christine Chambaz[2]

ABSTRACT

The study of social mobility, between labour market statuses or between income levels, for example, is often based on the analysis of mobility matrices. When comparing these transition matrices, with a view to evaluating behavioural changes, one often forgets that the data derive from a sample survey and are therefore affected by sampling variances. Similarly, it is assumed that the responses collected correspond to the 'true value.'

The purpose of our study is, first of all, to show measurement errors indirectly by means of an example. After defining the transition matrices, we then estimate the accuracy of the transition probabilities between states. Two types of transition matrices are considered in turn, according to the way in which the different possible states are defined: these states are defined first based on the modalities of a *categorical* variable during the survey, and then based on an estimated *distribution* of the survey data. The calculation of accuracy is more complex in the second instance. The study is carried out using data from the Economic Survey of Households. In the example considered, states correspond to the standard of living distribution quartiles. They are first assumed to be exogenous (which corresponds to the case in which the modalities are derived from a *categorical variable* from the survey). The inaccuracy related to to how they are defined is then taken into account to estimate the variance. The confidence intervals obtained for the elements of a transition matrice are shorter when the inaccuracy related to the definition of states is taken into account.

KEY WORDS: mobility matrices, measurement errors, accuracy.

1. INTRODUCTION

The study of inequalities is often based on the analysis of snapshot distributions of social states, such as a socio-professional class or standard of living quantile. However, those states are not immutable: a person whose standard of living on date t places him or her among the poorest 10% of the population may on date t' ($t'>t$) have a higher standard of living and thus fall within another distribution decile. Measuring social welfare may therefore be significantly impacted by whether or not opportunities for social mobility are incorporated.

Dardanoni (1993) proposes a particularly interesting approach to the analysis of social mobility. He considers asymmetrical utility functions, giving more weight to the future of individuals who initially fall within the most disadvantaged categories. He then establishes a criterion for categorizing societies sharing the same fixed distribution of social states from a social welfare perspective. This criterion is used to compare matrices where each element $\tilde{p}_{i,j}$ corresponds to the probability that an individual's situation is lower than or equal to i in t, and lower than or equal to j in $t' > t$. The thorough comparison of these matrices assumes that observed variances can thus be initially tested to estimate their variances. The purpose of this

[1] INSEE, Unité Méthodes Statistiques, Stamp F410, 18 boulevard A. Pinard, 75675 Paris Cedex 14, France.

[2] INSEE, Division Revenus et Patrimoine des ménages, Stamp F350, 18 boulevard A. Pinard, 75675 Paris Cedex 14, France.

study is therefore to propose a method for estimating the variance in transition probabilities, in particular probabilities $\widetilde{p}_{i,j}$ used in the dominance criteria established by Dardanoni.

The survey selected for this study is the triannual PCV (Permanente sur les Conditions de Vie des ménages) (Household Standard of Living Survey), half of the sample of which is renewed each year, and specifically the intersections of modalities for responses to the same question given by a single household in two successive surveys (May 1996 and May 1997). In this study, a household is defined as identical if its address is the same and if its head of household is the same (first name, date of birth and gender has remained the same). The households weights were adjusted by calibration to the margins to correct for sampling fluctuations and attrition.

First, we show measurement errors indirectly using a variable with limited variance between two successive interviews. We then estimate the accuracy of transition probabilities between states. Two types of transition matrices are considered in turn, according to the way in which the different possible states are defined: these states are defined first based on the modalities of a *categorical variable* during the survey, and then based on an estimated *distribution* of the survey data. This paper is an excerpt of a more lengthy paper to be published in 1998
(Caron, Chambaz, 1998).

2. TRANSITION PROBABILITY OR MEASUREMENT ERROR?

All statistical analyses are based on data measured with a certain degree of inaccuracy. A well-known source of inaccuracy is selected households who refuse to respond to the survey, which may, in the absence of a correction for non-response, produce biased estimators. Another less blatant source comes from two other types of errors liable to occur between the time the question is asked and the time the results file is available: measurement errors and variable coding and capture errors. In this section, we illustrate the existence of measurement errors using the education level variable (see also Caron, 1993).

The education level of the reference person is a variable constructed from statements made by the individual about their highest degree of general education, technical education and post-secondary education. These three education variables are each coded using code-cards during the survey. The only possible variations therefore correspond to an increase in the education level, and are located on the upper diagonal of the transition matrice. However, in our sample of 2,299 reference persons who have not moved, 281 (12.2%) gave answers that placed them lower in May 1997 than in May 1996 (Table 1). Three hundred and seventy-one (371) reference persons in our sample (16.1%) reported an increase in the education level between May 1996 and May 1997.

This proportion is approximately 10 times higher than that obtained from the first two waves of the European Panel of Households, in which questions on education are asked differently.

The types of errors illustrated in this section also affect the subsequently-analyzed standard of living transition matrices. Given that the main goal of this study is not to estimate the variability attributable to these errors, all observed situation changes will be treated as actual changes.

Table 1: Variance in reference person's education level

Estimated education	Estimated education level in May 1997						
level in May 1996	0	1	2	3	4	5	6
Level 0 (no diploma)	**361**	80	25	13	6	4	12
Level 1 (DFEO)	49	**326**	36	10	2	7	10
Level 2 (CAP-BEP)	15	30	**318**	50	10	3	11
Level 3 (BEPC)	13	14	30	**170**	15	19	8
Level 4 (tech. degree)	0	2	7	13	**50**	16	5
Level 5 (general degree)	4	3	4	18	14	**75**	29
Level 6 (post-grad)	7	4	8	10	**11**	**25**	**347**

Source: PVC survey May 1996 and 1997

Field: households surveyed both times where the reference person has not changed

Note: impossible a priori configurations are shaded.

3. ESTIMATING THE ACCURACY OF THE TRANSITION MATRICE COMPONENTS

3. 1. Reminders and notes

This survey uses a multi-stage stratified sampling plan. Several accuracy calculations conducted from this survey produced a design effect close to 1, meaning that in the initial approach, the sampling plan can be considered as a one-time random sampling.

3. 2. Methodological aspect

N represents the total number of households in the population and n represents the number of respondent households. The extrapolation weights in the file are represented as w_k.

3.2.1 States defined using based on the modalities of a categorical variable

It is assumed that the states are defined using the p modalities of variable X, represented as VAR1, VAR2,...., VARP. We will use $\hat{p}_{ij}$ to represent the estimated proportion of heads of households responding with the *i*th modality of the variable X during the first interview and with the *j*th modality during the second interview. The variance estimation calculations subsequently developed for the estimated proportion $\hat{p}_{11}$ are easily generalized for the estimated proportion $\hat{p}_{ij}$.

Using for each household k:

$y_k = 1$ if the head of household reported VAR1 at both interviews

$y_k = 0$ if not

and using $\hat{N}$ to represent the estimated number of households from the survey, we obtain

$$\hat{p}_{11} = \frac{\hat{Y}}{\hat{N}}$$

where $\hat{Y}$ is the estimator of the sum of *Y* obtained with weights w_k.

As a result, $\hat{p}_{11}$ is written as the ratio of the estimated sums of the two variables. Using the linearization technique applied to a ratio, we show that the variance for $\hat{p}_{11}$ is that of the sum of the variable *U* defined for a household k by:

$$u_k = \frac{1}{\hat{N}}\left(y_k - \frac{\hat{Y}}{\hat{N}}\right)$$

which, by approximating the sampling plan as a without replacement random sampling plan, gives

$$V(\hat{p}_{11}) = V\left(\frac{\hat{Y}}{\hat{N}}\right) = V(\hat{U}) = \left(1 - \frac{n}{N}\right)\frac{N^2}{n}\frac{1}{n-1}\sum_{k=1}^{n}(u_k - \bar{u})^2 \approx \frac{\hat{N}^2}{n} s_u^2$$

where $\hat{U}$ is the estimated sum of the variable U and s_u^2 is the modified empirical variance of the variable U calculated from the sample.

Therefore, the level 0.95 confidence interval of the ratio $\hat{p}_{11}$ is:

$$\left[\hat{p}_{11} - 1.96\sqrt{\hat{V}(\hat{U})}\,;\,\hat{p}_{11} + 1.96\sqrt{\hat{V}(\hat{U})}\right]$$

3.2.2. States defined based on an estimated distribution of the survey data

The accuracy calculation differs from that established in the previous example because the states, and as a result the boundaries separating the different states, are derived from an estimated distribution of the survey data. We will use $\hat{T}_{\alpha\beta}$ to represent the estimated proportion of individuals whose variable X value is lower than the estimated α distribution fractile and whose variable Y value is lower than the estimated β distribution fractile. In this section we will estimate the $\hat{T}_{\alpha\beta}$ variance. The results obtained would not be particularly valid for the proportion $\hat{p}_{22}$ defined in the transition matrix presented above, that is for estimations of proportions of individuals located in the transition matrix at the intersection of the two states, one of which is defined by two boundaries estimated from the sample.

Let x_α (resp. y_β) be an estimate of the fractile x_α^* (resp. y_β^*) of the variable X distribution with order α, based on the first interview (resp. β for the second interview). By using M to represent the discrete snapshot measurement with unit value in each point of the population and $\hat{M}$ to represent the measurement describing the weighted sample (using weights w_k), the statistic of interest $T_{\alpha\beta}$ and its estimator $\hat{T}_{\alpha\beta}$ can be written respectively as:

$$T_{\alpha\beta} = \frac{\int^{x^*_\alpha}\int^{y^*_\beta} dM(x,y)}{N} = \frac{\int^{x^*_\alpha}\int^{y^*_\beta} dM(x,y)}{\int\int dM(x,y)} \quad \text{and} \quad \hat{T}_{\alpha\beta} = \frac{\int^{x_\alpha}\int^{y_\beta} d\hat{M}(x,y)}{\hat{N}} = \frac{\int^{x_\alpha}\int^{y_\beta} d\hat{M}(x,y)}{\int\int d\hat{M}(x,y)}.$$

According to J.-C. Deville (1997, 1998), the variance of $\hat{T}_{\alpha\beta}$ is the sum of the linearized variable related to $T_{\alpha\beta}$ which corresponds in the case studied to the influence function $T_{\alpha\beta}$. By noting that the parameter of interest $T_{\alpha\beta}$ is a ratio, the linearized variable of $T_{\alpha\beta}$ is: $IT_k = \frac{1}{N}\left(A_k - T_{\alpha\beta}\right)$ where A_k corresponds to the linearity of the numerator obtained using the influence function.

After calculating A_k and grouping the various terms (see Caron, Chambaz, 1998), the influence factor IT is written as:

$$IT_k = \frac{1}{\hat{N}}\left(\left(\mathbb{I}\left(x_k < x_\alpha, y_k < y_\beta\right) - \hat{T}_{\alpha\beta}\right) - A\left(\mathbb{I}\left(x_k < x_\alpha\right) - \alpha\right) - B\left(\mathbb{I}\left(y_k < y_\beta\right) - \beta\right)\right)$$

where

- A (resp. B) is an estimate of the distribution function in y_β (resp. x_α) conditional upon $x = x_\alpha$ (resp. $y = y_\beta$), i.e.:

$$A = \frac{\sum\limits_{\substack{y_k < y_\beta \\ x_k \in [x_\alpha - \varepsilon; x_\alpha + \varepsilon]}} w_k}{\sum\limits_{x_k \in [x_\alpha - \varepsilon; x_\alpha + \varepsilon]} w_k} \quad \text{and} \quad B = \frac{\sum\limits_{\substack{x_k < x_\alpha \\ y_k \in [y_\beta - \varepsilon; y_\beta + \varepsilon]}} w_k}{\sum\limits_{y_k \in [y_\beta - \varepsilon; y_\beta + \varepsilon]} w_k}$$

The element ε must be chosen in such a way that the estimations of the distribution function are the least sensitive as possible to sample fluctuations.

- x_k (resp. y_k) represents the income for household k during the first interview (resp. second).
- $\mathbb{I}(m) = 1$ if the occurrence m is checked, or 0 if not.

We can analyze the consistency of the linearized variable obtained, for example by verifying that β (resp. α) tends towards 1, i.e. by choosing as a variable of interest the proportion of individuals whose variable ***X*** value (resp. ***Y***) is lower than the α fractile (resp. β), the linearized variable is also zero, resulting in a zero variance. In this configuration, the proportion is a known quantity which is also equal to α (resp. β).

As a result, the level 0.95 confidence interval for the ratio $\hat{p}_{11}$ is:

$$\left[\hat{p}_{11} - 1.96\sqrt{\hat{V}(I\hat{T})}; \hat{p}_{11} + 1.96\sqrt{\hat{V}(I\hat{T})}\right]$$

where $\hat{V}\left(I\hat{T}\right) = \left(1 - \frac{n}{N}\right)\frac{N^2}{n} s_{IT}^2$ using s_{IT}^2 for the modified empirical variable of variable ***IT*** calculated for the sample.

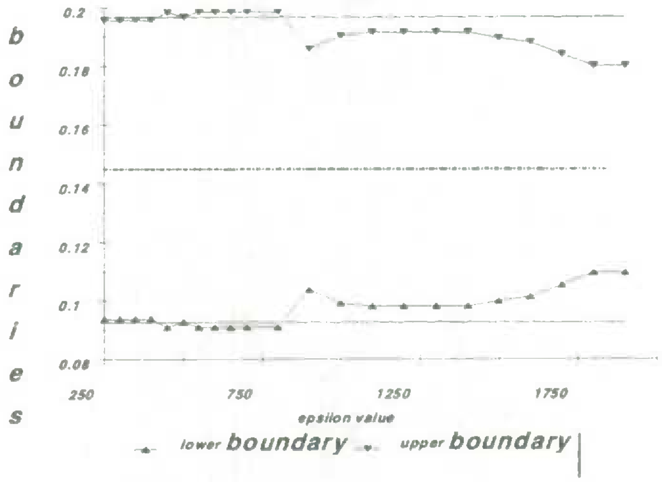
3.3. Application

In this section, we will apply the results obtained above to the proportion of stable households in the first standard of living[3] distribution quartile for the two interviews. We compare the variance estimations obtained by assuming first that the values of the quartiles are fixed and secondly that the values of the quartiles are estimated from the survey. The calculation was done for the northern Pas-de-Calais region (158 observations).

The confidence interval obtained by assuming that the quartiles estimated from the survey correspond to the true boundaries is wider than that obtained by integrating the random nature related to the estimation of those quartiles. However, we observe instability in the estimation of the variance when parameter ε is small. Such instability is related to the estimations of the conditional distribution functions based on very small numbers (one or even two observations for the initial values ε). The width of the confidence interval decreases when ε increases until it stabilizes when the estimations of the conditional distribution functions are more robust. Specifically, the confidence interval for low ε values may be wider than that obtained when fixed quartiles are assumed.

[3] Standard of living is defined as the ratio of household income to number of equivalent adults in the household. The equivalency scale used assigns a weight of 1 to the first adult in the household, 0.5 to other adults and 0.3 to children under 14 years of age.

Graph: proportion of heads of households in the first standard of living distribution quartile for the two interviews ($\hat{p}_{11}$)- North Region - Pas-de-Calais[4]



Value of the first quartile:
- May 1996 : 34 092 francs
- May 1997 : 36 108 francs

4. CONCLUSION

In this study, the change in the situation matrices considered are based on states defined using the distribution of a variable observed in the sample. We compared the length of two confidence intervals: the interval obtained when assuming that the situation definitions are exogenous and the interval obtained taking into account the true determination of states. We observe that the confidence intervals obtained for the elements of a transition matrice are shorter when the inaccuracy of the definition of the states is taken into account. The same result can be established for the stability probabilities in a situation (see Caron and Chambaz, 1998). However, the variance estimation calculations conducted in the latter case have two limitations. On the one hand, they are based on two estimates of conditional distribution functions. Although those may prove unstable when they are based on very few observations (1, 2 or 3), this can be offset by slightly increasing the number of observations used in the estimation of distribution functions. The estimation of the variance thus reaches an almost stable level, corresponding to the sought value. On the other hand, as we explained at the beginning, all observed changes in situation were treated as true changes in situation. To complete our results, we would have to be able to evaluate the variability derived from measurement errors and link it to the calculation of the variance estimation done here.

To conduct comparison tests of mobility matrices, two types of approaches are possible: one consists of emphasizing simplicity in the variance calculations by assuming that the states are defined exogenously for the survey; the other, however, consists of conducting the most thorough calculations possible by incorporating inaccuracy related to the definition of the states, given that application is more complicated. When there is a significant number of transition matrices to be compared, an interim approach could consist of doing thorough calculations for both matrices, evaluating the order of importance of the over-estimation of the variance obtained using exogenous states, and then doing the simplest calculation by extrapolating

[4] **Legend**: The lower boundary (upper resp.) corresponds to the lower boundary (upper resp.) of the confidence interval when it is assumed that the values of the quartiles are estimated from the survey. The three solid lines correspond, from bottom to top: to the lower boundary of the confidence interval when it is assumed that the values of the quartiles are known, to the estimator, and to the upper boundary of the confidence interval when it is assumed that the values of the quartiles are known.

that over-estimation for all other comparisons.

REFERENCES

Caron, N. (1993) : "Réflexion sur les erreurs de mesure: l'exemple de l'enquête de conjoncture auprès des ménages", working paper n°F9308 in the INSEE Direction des Statistiques Démographiques et Sociales.

Caron, N. and Chambaz, C. (1998) : "Matrices de mobilité et calcul de la précision asssociée", working paper, Survey Methodology, awaiting publication.

Dardanoni, V. (1993) : "Measuring social mobility", *Journal of Economic Theory*, **61**, P.372-394.

Deville, J.-C. (1997) : Course notes - ENSAE.

Deville, J.-C.(1998) : "Estimation de variance pour des statistiques complexes : Techniques de résidus et de linéarisation", awaiting publication.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

A LATENT CLASS MODEL FOR THE TRANSITION FROM SCHOOL TO WORKING LIFE IN THE PRESENCE OF MISSING DATA

Giulio Ghellini[1]
Andrea Regoli[2]

ABSTRACT

A longitudinal study on a cohort of pupils in the secondary school has been conducted in an Italian region since 1986 in order to study the transition from school to working life. The information have been collected at every sweep by a mail questionnaire and, at the final sweep, by a face-to-face interview, where retrospective questions referring back to the whole observation period have been asked. The gross flows between different discrete states – still in the school system, in the labour force without a job, in the labour force with a job – may then be estimated both from prospective and retrospective data, and the recall effect may be evaluated. Moreover, the conditions observed by the two differenl techniques may be regarded as two indicators of the 'true' unobservable condition, thus leading to the specification and estimation of a latent class model. In this framework, a Markov chain hypothesis may be introduced and evaluated in order to estimate the transition probabilities between the states, once they are corrected or the classification errors. Since the information collected by mail show a given amount of missing data in terms of unit nonresponse, the 'missing' category is also introduced in the model specification.

KEY WORDS: Gross flows; Response errors; Memory errors; Latent Markov models.

1. INTRODUCTION

A great deal of attention is recently being paid to the data quality from longitudinal studies. Such a growing interest has mainly taken the shape of analyses for measuring, treating and controlling the presence of non sampling errors (in the form of measurement errors or nonresponses). Less marked has been the contribution to the specification of statistical models combining both the phenomenon under analysis and the error generating mechanism (see, among the others, Abowd and Zellner, 1985, and Poterba and Summers, 1986).

This paper is intended to contribute to the latter kind of research, by focusing on the transition between discrete states. The availability of longitudinal data may arise either from a panel or a retrospective survey. The former suffers mainly from wave nonresponse (besides measurement errors due to conditioning effects), bringing the problem of selection bias to the foreground. The latter is mainly plagued by memory errors that are even more severe when the reference period is farther back in time.

Specifically this paper deals with the opportunity of adjusting the measurement errors presumably included in the data from a longitudinal Italian survey (called LEVA project): this survey is designed to follow some cohorts of pupils in order to study their transition from school to working life over a period of six/seven years after the compulsory education. The information on the schooling/working condition is collected every year by a mail questionnaire; the last interview is planned as a face-to-face interview and includes retrospective questions over the whole time period. On the basis of this information, a stochastic approach for estimating the school to work transition has been followed, which exploits both observations on the

[1] Giulio Ghellini, Dip. Metodi Quantitativi, P.za S. Francesco 8, 53100 Siena, Italy, giulio@turing.dmq.unisi.it

[2] Andrea Regoli, ISTAT, Via A. Ravà 150, 00142 Roma, Italy, regoli@istat.it

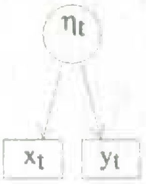
same set of units. The two observations correspond to two manifest indicators of the schooling/working condition, which is considered to be a variable that is not directly observable.

2. MARKOV MODELS AND WAVE NONRESPONSE

The methodology applied in this paper is defined in the framework of latent Markov models. These models, proposed by Lazarsfeld and Henry (1968) and used by Van de Pol and Langeheine (1992) for the flows estimation with longitudinal data, assume that at each time period a 'true' condition, observable with errors, is associated with every sample unit. In other words, the actual condition is to be considered a latent variable, only observable through one or more fallible indicators at the same time period.

In this study, at each time period, the latent condition - employed, unemployed, student - is supposed to be measured through a pair of indicators, the condition declared respectively at the repeated and at the retrospective survey (figure 1).

Figure 1 - The latent variable and the indicators



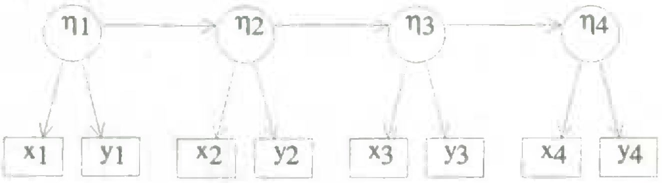
η_t: latent condition, time t

x_t: condition observed at the mail, repeated survey, time t

y_t: condition observed at the retrospective survey, time t

Moreover, the latent classification is assumed to follow a first order Markov chain, such that the latent condition at time t+1 depends on the latent condition at time t only (and not on the latent condition at the previous times, figure 2).

Figure 2 - The latent Markov model



Besides the initial probabilities of every latent category, the model estimates: *a)* the conditional transition probabilities between the latent states and *b)* the response probabilities, i.e. the probabilities of observing the state 'i' when the true state is 'j'.

Maximum likelihood estimates of the parameters can be obtained through a version of the EM algorithm implemented in the software PANMARK (Van de Pol, Langeheine and de Jong, 1991).

If another observed category (missing condition) is added to the three just recalled (employed, unemployed, student) for the repeated survey, the probability of a missing response can be estimated conditional on each latent state. This allows us to verify whether the nonresponse behaviour could be assumed independent of the latent condition or, on the contrary, the belonging to one of the three latent conditions could significantly modify the chance of nonresponse.

As already mentioned, the memory errors that may afflict the retrospective data could plausibly lead to an underestimation of the observed transitions (see, for instance, Mathiowetz and Duncan, 1988). The true, latent transitions are expected to be adjusted from this kind of errors, so that from the retrospective survey the estimated flows turn out to be larger than the observed flows.

On the other hand, the observed flows may be affected by classification errors, determining spurious transitions. The estimated transitions are expected to take this kind of errors into account and adjust for it in the direction of turning the spurious changes into actual permanencies (see, among the others, Meyer, 1988 and Singh and Rao, 1991).

The estimated response probabilities allow to measure the degree of correspondence between the latent condition and what is observed through both the retrospective and the repeated survey. With reference to the repeated survey, the estimates permit the evaluation of both measurement errors (in the form of classification errors) and wave nonresponses.

3. THE DATA SET

The data come from the LEVA project (see figure 3 for the design of the survey; see also Bernardi, Ghellini and Penello, 1996), specifically from the first cohort of pupils.

Figure 3 Survey design of the LEVA project

Observed Cohorts	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
I Cohort	I C ❶ May	II M ❶ May	III M ❶ Oct/Dec	IV M❶ Oct/Dec	V M ❷ Oct/Dec	VI M ❸ Oct/Dec	VII H ❶ Oct/Dec			
II Cohort				I C ❶ May	II M ❶ Oct/Dec	III M ❹ Oct/Dec	IV M ❶ Oct/Dec	V M ❷ Oct/Dec	VI M ❸ Oct/Dec	VII H ❶ Oct/Dec
III Cohort							I C ❶ May	II M ❶ Oct/Dec	III M ❶ Oct/Dec	IV M ❶ Oct/Dec
IV Cohort										I C ❶ May

Survey techniques:	C: interview in classroom	Observed subgroups:	❶: the whole sample
	M: interview by mail		❷: school leavers not employed
	H: Interview at home		❸: students
			❹: not students

The information has been collected by both three mail interviews (may 1987, october 1988 and october 1989) and a face-to-face interview (late 1992). The first mail interview asks for both the present condition and the condition referring back to october 1986, while the face-to-face interview contains retrospective questions over the whole period 1986-1992.

The initial sample is composed of 2,422 pupils. The analysis has kept the information on 1,618 units (about two thirds of the whole sample), corresponding to those participating both at the final interview and at least at one yearly mail survey.

The nonresponse rate is higher for the first and the third mail survey (respectively 16.7 and 14.8 per cent) and lower for the second (5.3 per cent), due to differences in the follow-up strategies for the initial nonrespondents.

In accordance with the expectations, the observed transitions from the retrospective survey show a less stressed dynamics, especially for the labour force conditions (employed and unemployed). With reference to the flows from the second to the third wave (see table 1) the percentage of the employed who stay in the same condition after one year is 90.4 from the mail survey and 97.0 from the retrospective survey. For the unemployed there are even more marked differences: from the yearly survey, only the 37.5 per cent of them remain unemployed at the third wave, as against the 48 per cent from the retrospective final survey.

Table 1- Permanencies in the observed conditions at the different waves (percentages)

Observed condition	Mail 1→2	Retrospective 1→2	Mail 2→3	Retrospective 2→3	Mail 3→4	Retrospective 3→4
Employed	98,3	100,0	90,4	97,0	96,8	97,4
Unemployed	62,8	86,8	37,5	48,0	47,6	57,5
Student	97,2	97,7	86,6	88,4	89,9	88,9

4. MODEL SPECIFICATION AND PRELIMINARY ANALYSES

The above discussion has led to the specification and the estimation of a first order Latent Markov Model. The latent variable is the school/work condition (employed, unemployed, student) at the four waves, each time measured by two manifest indicators, i.e. the condition observed through both the repeated mail and the retrospective survey.

The model parameters are: a) the initial probabilities of being in a specified condition; b) the response probabilities[3]; c) the transition probabilities, i.e. the conditional probabilities of moving from one condition at the beginning of the time period to another one at the end of the interval.

Since the pattern of the observed condition shows many empty cells, the chi-square approximation for the likelihood ratio (LR) statistics is not valid in order to evaluate the goodness-of-fit of the model. Nevertheless such approximation can be used for the comparison of nested models as a criterion for the specification search (Van de Pol, Langeheine and de Jong, 1991).

An evaluation of the estimated response probabilities (table 2) highlights the unemployment as the condition which is more often missclassified.

Table 2 – Estimated response probabilities

Mail 1

		Observed condition			
		E	U	S	N.R
Latent condition	E	0.36	0.25	0.11	0.28
	U	0.13	0.47	0.11	0.28
	S	0.00	0.00	0.85	0.15

Mail 2

		Observed condition			
		E	U	S	N.R
Latent condition	E	0.55	0.08	0.09	0.28
	U	0.21	0.49	0.03	0.28
	S	0.00	0.00	0.85	0.15

Mail 3

		Observed condition			
		E	U	S	N.R
Latent condition	E	0.81	0.07	0.04	0.08
	U	0.29	0.53	0.10	0.08
	S	0.00	0.01	0.95	0.04

Mail 4

		Observed condition			
		E	U	S	N.R
Latent condition	E	0.69	0.02	0.01	0.28
	U	0.29	0.38	0.05	0.28
	S	0.02	0.02	0.88	0.08

Retrospective 1

		Observed condition		
		E	U	S
Latent condition	E	0.90	0.01	0.09
	U	0.00	0.73	0.27
	S	0.00	0.00	1.00

Retrospective 2

		Observed condition		
		E	U	S
Latent condition	E	1.00	0.00	0.00
	U	0.00	0.76	0.24
	S	0.00	0.00	1.00

[3] The estimated response probabilities are the entries of two kinds of matrices. The former is a 3x4 matrix whose entries correspond to the probabilities of observing the 4-state condition through the repeated survey, given the 3-state latent condition. The latter is a 3x3 matrix whose entries are the probabilities of observing the 3-state condition through the retrospective survey, given the 3-state latent condition.

Retrospective 3

		Observed condition		
		E	U	S
Latent condition	E	0.88	0.03	0.09
	U	0.02	0.63	0.35
	S	0.00	0.00	1.00

Retrospective 4

		Observed condition		
		E	U	S
Latent condition	E	1.00	0.00	0.00
	U	0.11	0.69	0.20
	S	0.00	0.01	0.99

E: employed, U: unemployed, S: student, N.R.: nonresponse

Moreover, the probability of not responding to the repeated survey is lower if you are still in the schooling system. In a formal way, the independence of the response behaviour from the true condition (an assumption which is equivalent to the MCAR assumption; Rubin, 1976) may be checked by introducing the restrictions of equality for the entries in the response probability matrices for the repeated survey corresponding to the probabilities of a missing response given the three latent conditions. In such a situation, the matrix to be estimated for each wave of the repeated survey is of the following kind:

$$\begin{bmatrix} A & B & C & D \\ E & F & G & D \\ I & J & K & D \end{bmatrix},$$

where the equality restriction is highlighted by the same letter (D). The comparison is based on the difference between the LR statistics of the two nested models (the general and the restricted model), showing that such restrictions are to be rejected (ΔLR=145.15 with 6 degrees of freedom).

If the equality restriction is only imposed on the two parameters for the employed and the unemployed, i.e. if the matrix to be estimated is of the following kind:

$$\begin{bmatrix} A & B & C & D \\ E & F & G & D \\ I & J & K & L \end{bmatrix},$$

the LR difference shows that this restriction can't be rejected (ΔLR=1.58 with 3 degrees of freedom).

The estimated latent transition matrices (see table 3) show that the permanencies in the same condition (the diagonal entries in the matrices) are usually higher than the observed ones. The largest flows are estimated from the unemployment to the employment. Generally these matrices are of a lower triangular kind, showing nearly no moves out of the employment state nor moves towards the student condition.

Table 3 – Estimated transition probabilities

1→2

	E	U	S
E	1.00	0.00	0.00
U	0.13	0.87	0.00
S	0.00	0.01	0.99

2→3

	E	U	S
E	1.00	0.00	0.00
U	0.43	0.53	0.04
S	0.08	0.07	0.85

3→4

	E	U	S
E	0.98	0.02	0.00
U	0.28	0.72	0.00
S	0.03	0.03	0.94

E: employed, U: unemployed, S: student

The information on some background variables included in the available dataset allows to introduce and evaluate the presence of heterogeneity both in the dynamics of transition and in the response behaviour. Previously conducted analyses have underlined a set of variables to be considered as relevant for the study of the transition. They comprise: the gender, the regularity in the compulsory schooling path and the educational level of parents or eldest siblings.

In the framework of latent Markov models, the observed heterogeneity may be taken into account if a mixed model, with two or more different chains, is specified (see Poulsen, 1982).
Some preliminary analyses in this direction show that, on the basis of gender and schooling path, the hypothesis of heterogeneity (in transition and response probabilities) between groups cannot be rejected. Specifically, the males who have left the school after the first wave of the survey have a higher estimated probability of being employed at the fourth wave than the females in the same situation (0.79 against 0.49). Finally, the pupils who have passed through a schooling failure during the compulsory school have a higher estimated probability of remaining unemployed in the whole period than the others (0.49 against 0.23).

REFERENCES

Abowd, J.M. and Zellner, A. (1985), Estimating Gross labor-Force Flows, *Journal of Business & Economic Statistics*, 3, 254-283.

Bernardi L., Ghellini G. and Penello C. (1996), A panel survey design for the study of school careers and work path, paper presented at the *Fourth International Social Science Methodology Conference*, Univ. of Essex, 1-4 July, 1996 (mimeo).

Lazarsfeld, P.F. and Henry, N. W. (1968), *Latent Structure Analysis*, Boston.: Houghton Mifflin.

Mathiowetz, N.A. and Duncan G.J. (1988), Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment, *Journal of Business and Economic Statistics*, 6, 221-229.

Meyer, B.D. (1988), Classification- Error Models and Labour Market Dynamics, *Journal of Business and Economic Statistics*, 6, 385-390.

Poterba, J.M. and Summers L. H. (1986), Reporting Errors and Labor Market Dynamics, *Econometrica*, vol.54, 6, 1319-1338.

Rubin, D.B. (1976), Inference and missing data, *Biometrika*, 63, 581-592.

Singh, A. C. and Rao, J.N.K. (1991), Classification Error Adjustment for Gross Flow Estimates, Technical Report, n. 183, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Van de Pol, F.J.R., Langeheine, R. and de Jong W. (1991), PANMARK User manual: Panel Analysis Using Markov Chains, Version 2.2, Netherlands CBS.

Van de Pol, F.J.R. and Langeheine, R. (1992), Analysing Measurement Error in Quasi-Experimental Data: an Application of Latent Class Models to Labour Market Data, *Working Paper of he European Scientific Network on Household Panel Studies*, Paper n° 57, Colchester, University of Essex.

SESSION 5

GROSS FLOWS II

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

TOOLS FOR INFERENCE ON DYNAMICS OF LOW INCOME STATUS

Milorad S. Kovačević[1]

ABSTRACT

The log-linear modelling of categorical longitudinal survey data on income is studied. An emphasis is on inference about change. Special attention is paid to modelling of longitudinal data from two waves. A small illustration is based on data from the Canadian Survey of Labour and Income Dynamics.

KEY WORDS: Categorical Longitudinal Data; Log-Linear Modelling; Gross Flows; Transition Rates

1. INTRODUCTION

A goal of this paper is to show how the methods developed for analysis of categorical cross-sectional data may be applied to data from complex longitudinal surveys in order to analyse change. For such data, it is necessary to account for the temporal orderings and within-subject correlations, as well as to allow for the complex design of a longitudinal survey.

We consider a case where the original response variable y is categorized into $K \geq 2$ discrete, mutually exclusive and exhaustive categories, which we interchangeably call states. For example, after-tax family income can be categorized into two categories, "below the low income cut-off (LICO) corresponding to the family type" and "above the LICO." Here, the assumption is that LICOs come from an independent data source. A more general analysis of change for K>2 is presented as well. For example, we consider the family income categorized into K ordered categories as in (1) where $\xi_{t;p}$ is the estimated 100pth percentile of the income distribution at wave t for the corresponding family type

$$y_{ti}^{*} = \begin{cases} 1, & y_{ti} \leq \xi_{t;1/K}, \\ 2, \text{ if} & \xi_{t;1/K} < y_{ti} \leq \xi_{t;2/K}, \\ \dots, & \dots \\ K, & y_{ti} > \xi_{t;1-1/K} \end{cases} \tag{1}$$

In this case the category boundaries are sample dependent and thus add some variability to estimates of gross flows and transition rates. A balanced sample is assumed, i.e., that there is a longitudinal sample of n_L individuals that responded in all T waves. Also we assume that there is no classification error.

The data used for an illustration in this paper represent individuals of all ages in the 1992 Canadian population (the reference population) who were alive and lived in Canada in 1994. The illustration part of the paper is partly a reproduction of the analysis presented in a Statistics Canada document by Noreau, Webber, Giles and Hale (1997). An example of a 2x2 table of estimated counts is taken from their paper. Another table found in their paper presents the gross flows of persons and transition rates between the after-tax family income quintile groups. These tables are given in the Appendix.

[1] Milorad S. Kovačević, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada, e-mail: kovamil@statcan.ca

The next section deals with gross flows and transition rates, their estimation and variance estimation of their estimates. Section 3 contains a review of the basics of log-linear modelling of survey data. An approach to modelling change using data from two waves is developed in Section 4. In section 5 a summary is given with a list of possible directions for future research.

2. ESTIMATION OF GROSS FLOWS AND TRANSITION RATES

Any analysis is usually confined to a domain of interest in the population. In the case of a longitudinal population, a domain definition necessarily depends on time. For example, we define **the principal domain of the SLID in 1994** as: the members of the 1992 reference population that were alive in 1994, who lived in one of the ten provinces, but not in institutions, Indian reserves, nor in Armed Forces barracks. The longitudinal weights in a year of interest should add up to the size of the reference population properly reflecting changes in the size of domains of the reference population over time.

Straightforward summary information about a categorical variable y_t^* observed in both waves, takes the form of a KxK table of weighted counts $\{\hat{N}_{ab}\}$ $(a=1,...,K, b=1,...,K)$ that sum up to the estimated size of the domain of interest of the longitudinal population $\Sigma_a \Sigma_b \hat{N}_{ab} = \hat{N}_D$. The first index a indicates the state of the categorized variable y^* at the first time point, and the second index b indicates the state of the same variable at t=2. Estimates $\{\hat{N}_{ab}\}$ estimate the sizes of mutually exclusive subdomains of the domain of interest of the longitudinal population, taking form (2):

$$\hat{N}_{ab} = \sum_{i \in s_L} w_i I\{y_{1i}^*=a\} I\{y_{2i}^*=b\} , \qquad (2)$$

and are called estimates of gross flows. The corresponding cell proportions are estimated as $\hat{P}_{ab} = \hat{N}_{ab} / \hat{N}_D$.

For measuring change it is generally more convenient to use transition rates which estimate the conditional probabilities of staying in the same state or shifting from one state to the other. In the case of a KxK table the transition rates are obtained as ratios of estimated counts to the corresponding row margins, that is $\hat{P}_{b|a} = \hat{N}_{ab} / \hat{N}_{a+}$, where $\hat{N}_{a+} = \sum_b \hat{N}_{ab}$.

For inferential purposes it is necessary to estimate the covariance matrix of the table of estimated counts or transition rates. The number of degrees of freedom (df) with which these variances and covariances are estimated is important for further inferential applications of the variance estimators. For the linearized variance estimator the number of degrees of freedom is approximated by the total number of PSU's reduced for the number of strata, assuming sampling of PSU's with replacement. In the SLID example, variance and covariance estimates are obtained using the linearized covariance estimator with 237 df at Canada level.

Regarding the definition of the states (categories) we distinguish two cases: (i) category boundaries are obtained from sources different than the actual sample, and (ii) category boundaries are estimated from the same sample. In the second case the variability of the boundaries has to be accounted for in estimation of variances of gross flows and transition rates. Using linearization we are able to obtain an expression for the variance estimator in a closed form. If using the case of resampling methods, for each replicate we need to reestimate the boundaries as well.

3. LOG-LINEAR MODELS

Most categorical data analysis methods have been developed in the context of simple random sampling and stratified sampling assuming independence among sampling units. If these methods are applied unmodified to data obtained in complex sampling schemes that include stratification, clustering and unequal probabilities

of selection, results may be erroneous. A variety of modifications of the classical methods has been developed. These different approaches, with the risk of oversimplification, can be classified into two groups:
(i) The first group represents modifications of the usual Pearson X^2(or likelihood ratio test statistics L_R) in order to achieve an asymptotic χ^2 distribution for a given complex design. The most notable result in this group is known as the Rao-Scott theory, given in papers by Rao and Scott (1981, 1984, 1987) and Rao and Thomas (1987, 1988). In the same group is Fay's (1985) jackknifing X^2 test which essentially estimates the Rao-Scott first-order corrected X^2 or L_R by means of resampling.

(ii) The second group of methods is based on the application of the generalized least squares method (Grizzle, Starmer and Koch (1969), and Koch, Freeman and Freeman, (1975)). The assumption on availability of a consistent estimate of the covariance matrix of the estimated cell counts (or proportions) is essential. Then, the Wald statistic is used for testing. However, due to the low precision of covariance estimation for complex surveys, especially when the number of cells in the table is large, the Wald statistic may not perform adequately, resulting in a high rate of rejection under a null hypothesis. A good study of this phenomenon is given in Hidiroglou and Rao (1987).

In order to present the basic ideas for analysis of change we adopt the second approach. The transition tables contain a small number of cells relative to the number of degrees of freedom of the estimated covariance matrix.

A regular log-linear model, with the unknown proportions π replaced by their design-consistent estimators $\hat{p}$, is:

$$\log\hat{p} = X\lambda + \varepsilon. \tag{3}$$

The error term ε reflects the correlations among estimated proportions, as well as their heteroscedasticity. It is centred at zero and has the covariance matrix V_ε which is approximated by expression (4) (using Taylor linearization):

$$V_\varepsilon \approx D^{-1}V(\hat{p})D^{-1}, \tag{4}$$

where $D = diag\{\pi\}$. Now, the estimator of λ (the vector of parameters) can be obtained by the weighted least squares method as

$$\hat{\lambda} = (X'\hat{V}_\varepsilon^{-1}X)^{-1}X'\hat{V}_\varepsilon^{-1}\log\hat{p},$$

with its covariance matrix estimated by a consistent estimator

$$\hat{V}(\hat{\lambda}) = (X'\hat{V}_\varepsilon^{-1}X)^{-1}.$$

The residual term

$$W_{g-r} = (\log\hat{p} - X\hat{\lambda})'\hat{V}_\varepsilon^{-1}(\log\hat{p} - X\hat{\lambda})$$

is asymptotically χ^2 distributed under the null hypothesis of goodness of fit for the model. The number of df of W_{g-r} is *g-r*, where *g* is the number of cells size of the table, and *r* is the number of parameters in the model.

The next section contains some examples of log-linear models for analysis of change in 2x2, and more general models for *K*x*K* tables.

4. MODELLING OF CHANGE

4.1. Change in a 2x2 table

To analyse change in low income (LI) status one needs to model categorical longitudinal data in such a way that a test of goodness of fit of such a model is actually making inference about change. We begin with an elementary situation with two waves (*T*=2) and two states (*K*=2). In order to emphasize components of dynamics, Clogg, Eliason and Grego (1990) reparametrized the standard log-linear model. For example in a

2x2 table, cells (1,1) and (2,2) define the "no-change" situation or "persistence" in the LI status (a=b=1) or in the above the LI status (a=b=2), respectively . Similarly, cells (1,2) and (2,1) denote change from "low income" to "above low income", and vice versa. The estimated proportions are $\hat{p} = \{\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{21}, \hat{p}_{22}\}$, a corresponding design matrix X is

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & 0 & -1 & -1 \\ 1 & -1 & 0 & 1 \end{pmatrix},$$

and the parameter vector is $\lambda = (\mu, \lambda^P, \lambda^S, \lambda^A)'$. Here λ^P describes the persistence effect, λ^S expresses the effect of change or the symmetry, and λ^A is a parameter that describes the temporal association between two distributions of counts (at t=1 and t=2).

To infer about transitions between the states, we first observe that the ratio of the transition rates from a given state a at t=1 to states 1 and 2 at t=2 represents the odds $P_{1|a} / P_{2|a} = N_{a1} / (N_{a\cdot} - N_{a1})$, a=1,2. Taking the logarithm of the ratio we end up with an ordinary logit model

$$L_a^{(2)} = \log\left[P_{1|a} / (1 - P_{1|a})\right] = 2\lambda_1^{(2)} + 2\lambda_{a1}^{(12)}, \; a=1,2. \tag{5}$$

This model tells how transition rates vary with different states at t=1. However, since this model is saturated, there are not enough degrees of freedom to reparametrize it to analyse change.

Another approach may be taken for testing a hypothesis on persistence (or on symmetry) in transition probabilities. Assuming large-sample normality, a null hypothesis on persistence: $P_{1|1} - P_{2|2} = 0$ can be tested using the normal test. The variance of the difference can be estimated using the covariance matrix obtained for the table of transition rates: $\hat{Var}(\hat{P}_{1|1} - \hat{P}_{2|2}) = \hat{Var}(\hat{P}_{1|1}) + \hat{Var}(\hat{P}_{2|2}) - 2\hat{Cov}(\hat{P}_{1|1}, \hat{P}_{2|2})$.

Note that the interpretation of a test on persistence in the case of counts is different from the case of transition probabilities. In the first case, H_0: λ^P=0, we test whether the number of people that remain in the same state is identical for both states. In the case of transition rates, the null hypothesis H_0: $P_{1|1} = P_{2|2}$ means that the conditional probabilities of staying in the same state are identical.

As an illustration, consider Table A.1, which contains estimates of gross flows from two waves of SLID. Models of persistence, symmetry and independence were fit to these data. Parameter estimates, with their corresponding standard errors in parenthesis, are displayed in Table 1. Also included are Wald statistics for goodness of fit of each model, computed using the estimated covariance matrix of the estimated counts (Table A.3).

Clearly, there is a substantial difference between the two states in the persistence of gross flows, with fewer people stay in the LI state compared to people staying in the non-LI state. Similarly, fewer flow from the non-LI state to the LI state, than vice versa, making the margins heterogeneous.

Table 1: Results for models of change applied to data on gross flows in Table A.1

Models	$\hat{\mu}$	$\hat{\lambda}^A$	$\hat{\lambda}^P$	$\hat{\lambda}^S$	df	Wald statistic
Independence ($\lambda^A = 0$)	8.522 (.034)	0	-1.290 (.059)	0.185 (.046)	1	765
Persistence ($\lambda^P = 0$)	8.540 (.035)	1.094 (.034)	0	0.321 (.059)	1	921
Symmetry ($\lambda^S = 0$)	7.909 (.040)	0.958 (.035)	-1.164 (.038)	0	1	9.84
Saturated	7.907 (.034)	0.958 (.034)	-1.133 (.038)	-0.186 (.059)	0	

Using the interpretation of the odds ratio, $\exp(4\,\lambda^{A})$, as the ratio between the prevalence of the LI state at t =2 among those that were in the LI state at t =1, and the incidence of the LI state at t =2, we may say that if one was in the LI state in 1993, will be in the LI state in 1994, 45.77 times as likely as one would move there who was not there in 1993.

The test of persistence for the transition rates gives a Z statistic with a value of -13.39 indicating that there is a significant difference in the conditional persistence of the two states over time. Testing for symmetry in a 2x2 transition table is essentially the same as testing for conditional persistence.

4.2 Change in a KxK table

Some of the special models for square tables (see Bishop et. al (1975, p. 281)) can be adapted and used for studies of change in a KxK contingency table when K>2. Without going deeper into explanation of these special models we will briefly discuss their relevance for studying change.

In general, testing of marginal homogeneity (MH) via log-linear modelling is a difficult task since it includes all counts (or proportions) in the body of the table (K^2) and their variances and covariances (K^4). Instead, we perform two tests - one for symmetry and the other for quasi-symmetry and use their difference as a conditional test on MH, provided that quasi symmetry holds.

A standard model that is useful for analysis of gross flows is the quasi-independence model (or the mover-stayer model) introduced by Blumen, Kogan, and McCarthy, 1955. The basic assumption is that the longitudinal respondents are either movers (meaning that they change their response categories from wave to wave), or stayers (meaning that they keep the same response category in two waves.) If we can separate movers from stayers, the contingency table for stayers would have all off-diagonal elements equal to zero, while movers could exhibit total independence. Since there is no way to separate movers from stayers, we may model the conditional independence of the off-diagonal cells through a quasi-independence model.

Quantile groups as defined in (1) are ordinal categories; their ordinality may be used to define a logit model for transition rates for adjacent quantile groups. First, the ordered scores $\{u_a^{(1)}\}$ and $\{u_b^{(2)}\}$ are assigned to the quantile groups at t=1,2. Then,

$$\log\left(P_{b+1|a}/P_{b|a}\right)=\log\left(N_{a,b+1}/N_{ab}\right)=\lambda_{b+1}^{(2)}-\lambda_b^{(2)}+\alpha\,(u_{b+1}^{(2)}-u_b^{(2)})\,u_a^{(1)}. \qquad (6)$$

For unit-spaced scores $\{u_b^{(2)}\}$, model (6) simplifies to $\log\left(P_{b+1|a}/P_{b|a}\right)=\beta_b+\alpha\,u_a^{(1)}$, where $\beta_b=\lambda_{b+1}^{(2)}-\lambda_b^{(2)}$. An interpretation of parameter α is the following: for a person that was in the a+1 group at t =1, the odds of transition to the quantile group b+1 instead of the quantile group b at t=2 differs by a factor of exp(α) from the same odds for a person who was in the a group at t=1. The equivalence of the logit models for transition rates to the log-linear models of proportions or counts provides a clear way for parameter estimation and testing of goodness of fit.

As an illustration of application of these special models consider Table A.2 which is a 5x5 table of estimated counts and proportions for the principal domain in 1994. Variable Y is categorized into 5 categories which are time dependent. This has an effect in estimation of the covariance matrix, see section 2. All tests were significant at the .05 level (see Table 2). A particular interpretation of α =0.6 and exp(α)=1.82, from model (6), is the following: a person that was in the second quintile group in 93, has an odds of moving up to the third quintile group in 94 that is 1.82 times higher than a person who was in the lowest quintile group in 93.

Table 2. Test results for models applied to data from Table A.2

Model	df	Wald Statistic	(χ^2)	Model	df	Wald Statistic	(χ^2)
Independence	6	164.98	(12.59)	Quasi-symmetry	6	19.69	(12.59)
Marginal Homogeneity	4			Conditional MH	4	4.18	(9.48)
Symmetry	10	23.87	(18.31)	Quasi-independence	11	46.11	(19.68)

5. SUMMARY

The idea for this paper was to provide analysts with a set of tools for inference about change based on longitudinal categorical data. Estimation of gross flows and transition rates and their variances is only a starting point on which the next stage, the modelling of change, relies. Models of symmetry, marginal homogeneity and independence are modified and applied to data with temporal dependencies. Examples using real data from SLID are chosen to support the selection of methods.

Methods for inference on change based on more than two waves would be a natural extension of this material. Some models could be directly extended, others require additional theory and modifications to suit the complexity of longitudinal surveys. These methods include hierarchical modelling for multi-way tables, and Markov models for transition rates. Another direction would lead toward incorporating covariates in modelling of change, distinguishing cases of fixed and time-varying covariates.

REFERENCES

Bishop, Y.M.M., S.E.Fienberg and P.W. Holland (1975). *Discrete Multivariate Analysis.* The MIT Press, Cambridge, Massachusetts.

Blumen, I., M. Kogan, and P.J. McCarthy, (1955). *The Industrial Mobility of Labour as a Probability Process.* Cornell Studies, No. 6, Cornell University Press, Ithaca, New York.

Clogg, C.C., S.R. Eliason and J.M. Grego (1990). Models for Analysis of Change in Discrete Variables. In *Statistical Methods in Longitudinal Research, vol.II*, (Ed. Alexander von Eye) Academic Press, Inc.

Fay, R.E. (1985). A Jackknifed Chi-Squared Test for Complex Samples. *Journal of the American Statistical Association*, 80, 148-157.

Grizzle,J.E., C.F. Starmer and G.G. Koch (1969). Analysis of Categorical Data by Linear Models. *Biometrics* 25, 489-504.

Hidiroglou, M.A. and J.N.K. Rao (1987). Chi-Squared Tests with Categorical Data from Complex Surveys. Part I. *Journal of Official Statistics*, 3, 117-132.

Koch, G.G., D.H. Freeman and J.L. Freeman, (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review*, 43, 59-78

Noreau, N., M. Webber, P. Giles and A. Hale (1997). Crossing the Low Income Line. *The Income and Labour Dynamics Working Paper Series*, 97-11, Statistics Canada, Ottawa, Canada.

Rao, J.N.K. and A.J. Scott (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-way Tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K. and A.J. Scott (1984). On Chi-Squared Tests for Multiway Contingency Tables With Cell Proportions Estimated From Survey Data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K. and A.J. Scott (1987). On Simple Adjustments to Chi-Square Tests With Sample Survey Data. *The Annals of Statistics*, 15, 385-397.

Rao, J.N.K. and D.R. Thomas (1987). Chi Squared Tests for Contingency Tables. In *Analysis of Complex Surveys*, (Eds. Skinner, C.J., D.Holt, and T.M.F.Smith) John Wiley and Sons Ltd. Chichester

Rao, J.N.K. and D.R. Thomas (1988). The Analysis of Cross-Classified Categorical Data From Complex Sample Surveys. *Sociological Methodology*, 18, 213-270.

APPENDIX

Table A.1 Gross Flows (in thousands) and Transition Rates Into and Out of Low Income (LI) Status Between 1993 and 1994

1993	1994		
	LI Status	Non-LI Status	Total
LI Status	$\hat{N}_{11}=2231$ $\hat{P}_{1\mid 1}=0.721$	$\hat{N}_{12}=864$ $\hat{P}_{2\mid 1}=0.279$	$\hat{N}_{1\cdot}=3095$
Non-LI Status	$\hat{N}_{21}=1255$ $\hat{P}_{1\mid 2}=0.053$	$\hat{N}_{22}=22484$ $\hat{P}_{2\mid 2}=0.947$	$\hat{N}_{2\cdot}=23740$
Total	$\hat{N}_{\cdot 1}=3486$	$\hat{N}_{\cdot 2}=23348$	$\hat{N}=26835$

Table A.2 Gross Flows (in thousands) and Transition Rates Between Income Quintile Groups in 1993 and 1994

1993	Income quintile group in 1994					
	First	Second	Third	Fourth	Fifth	Total
First	3722 0.697	1152 0.216	292 0.055	118 0.022	52 0.010	5336
Second	920 0.174	3019 0.570	1001 0.189	249 0.047	105 0.020	5295
Third	312 0.058	761 0.142	3012 0.563	1139 0.213	129 0.024	5354
Fourth	234 0.043	325 0.060	835 0.153	3038 0.559	1007 0.185	5439
Fifth	169 0.031	106 0.020	233 0.043	801 0.148	4103 0.758	5411
Total	5358	5362	5374	5345	5396	26835

Table A.3 Covariance Matrix for 2x2 Gross Flows Table A.1

(x 10^6)

Cov	$\hat{N}_{11}$	$\hat{N}_{12}$	$\hat{N}_{21}$	$\hat{N}_{22}$
$\hat{N}_{11}$	23323	1830	3347	-7245
$\hat{N}_{12}$		5990	918	575
$\hat{N}_{21}$			12326	-11294
$\hat{N}_{22}$				412732

Table A.4 Covariance Matrix for 2x2 Transition Rates Table A.1

(x 10^{-4})

Cov	$\hat{P}_{1\mid 1}$	$\hat{P}_{2\mid 1}$	$\hat{P}_{1\mid 2}$	$\hat{P}_{2\mid 2}$
$\hat{P}_{1\mid 1}$	4.377	-4.377	0.053	-0.053
$\hat{P}_{2\mid 1}$		4.377	-0.053	0.053
$\hat{P}_{1\mid 2}$			0.237	-0.237
$\hat{P}_{2\mid 2}$				0.237

Proceedings of Statistics Canada Symposium 98
Longitudinal Analyses for Complex Surveys
May 1998

EVALUATING NONRESPONSE ADJUSTMENT IN THE CURRENT POPULATION SURVEY (CPS) USING LONGITUDINAL DATA[1]

Brian A. Harris-Kojetin[2] and Edwin L. Robison[3]

ABSTRACT

The purpose of the present study is to utilize panel data from the Current Population Survey (CPS) to examine the effects of unit nonresponse. Because most nonrespondents to the CPS are respondents during at least one month-in-sample, data from other months can be used to compare the characteristics of complete respondents and panel nonrespondents and to evaluate nonresponse adjustment procedures. In the current paper we present analyses utilizing CPS panel data to illustrate the effects of unit nonresponse. After adjusting for nonresponse, additional comparisons are also made to evaluate the effects of nonresponse adjustment. The implications of the findings and suggestions for further research are discussed.

KEY WORDS: Panel Nonresponse; Weighting

1. INTRODUCTION

Survey nonresponse occurs at several different levels, with individuals or households not responding at all (unit nonresponse), not responding during one wave or more waves or panels of a longitudinal survey (wave/panel nonresponse), or simply omitting certain survey items (item nonresponse). Ideally, one would hope that nonrespondents are a random cross-section of the sample, reflecting the same demographic, geographic, and economic groups as respondents. If this is the case, then one need not be concerned about obtaining biased results from a survey in which there was some degree of nonresponse. However, it is typically the case that nonrespondents differ from respondents on these characteristics (for reviews see Goyder, 1987; Groves and Couper, 1998). Therefore, even surveys with high response rates may have some degree of bias in their results to the extent that nonrespondents differ from the respondents.

Given that a survey has encountered some level of unit nonresponse, the survey manager is faced with the decision of how to deal with it. Typically weighting adjustments are used to compensate for unit nonresponse (Kalton and Kasprzyk, 1986) and a variety of different procedures and models can be used, e.g., population weighting, sample weighting, ratio, and response propensity (for summaries of different procedures see Kalton and Kasprzyk, 1986; Oh and Scheuren, 1983). The purpose of all these procedures is essentially to increase the weights of the respondent cases to represent the nonrespondents. Nonresponse weighting adjustments require some information about the nonrespondents and respondents as well as assumptions about the differences between respondents and nonrespondents.

[1] Opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

[2] Bureau of Labor Statistics, 2 Massachusetts Avenue N.E., Room 4915, Washington, D. C. 20212 USA.

[3] Bureau of Labor Statistics, 2 Massachusetts Avenue N.E., Room 4915, Washington, D. C. 20212 USA.

Of course, it is the nature of nonresponse that we almost never know exactly what we want to know about survey nonrespondents. Since longitudinal surveys have a panel or wave component, some insights into the characteristics of nonrespondents is available from information they provided on a previous or subsequent panel or wave. For example, monthly labor force surveys with large sample overlaps from month-to-month can use gross flow analysis to examine the "flow" of persons from respondent to nonrespondent status (and vice versa) to understand better the characteristics of nonrespondents and the consequences of nonresponse on labor force statistics (e.g., Stasny and Fienberg, 1985). In previous research using the Current Population Survey (CPS) we found that persons who were nonrespondents to the CPS one-month had higher rates of unemployment, labor force participation, and employment than those who were respondents both months (Tucker and Harris-Kojetin, 1997).

The purpose of the present research was to utilize longitudinal data from the CPS to identify specific differences between respondents and nonrespondents, and then to evaluate the CPS nonresponse adjustment procedures by comparing respondents and nonrespondents after the nonresponse weighting adjustment. Because the normal CPS processing is done separately on each monthly file, we created a longitudinal file that resembled as closely as possible a typical monthly file so that we could simulate the nonresponse adjustment as accurately as possible. In the current paper, we present preliminary analyses comparing the characteristics of nonrespondents to respondents using information obtained in months in which there was an interview. We then used similar procedures to the CPS nonresponse adjustment procedures to see how well the adjustment compensates for the differences observed between respondents and nonrespondents on demographic and labor force characteristics.

2. METHODS

2.1 Design of the CPS

The CPS is the monthly household labor force survey for the United States conducted by the U.S. Census Bureau for the U.S. Bureau of Labor Statistics. Approximately 50,000 eligible households are sampled each month in a two-stage clustered design. Households selected for the sample are interviewed for 4 consecutive months, are not interviewed 8 months, and then are interviewed again for 4 consecutive months. Furthermore, in any given month, one eighth of the sample is composed of households participating for the first time (month-in-sample 1; MIS 1), one-eighth the second time (MIS 2), etc. All households, except those in for the first time and the fifth time, were in sample the previous month; and, therefore, ¾ of the households are the same from month-to-month.

2.2 Data used in the Present Study

In order to replicate as closely as possible the typical monthly nonresponse adjustment to CPS, we created a longitudinal CPS file that consisted of eight "cohorts" who were in the CPS sample for the full 8 months over a 22-month period. Data for the present investigation were drawn from the eight cohorts that completed all of the eight months-in-sample (MIS 1-8) since their initial selection during or after April 1996. The first cohort began in April 1996 and finished in July 1997, and the last cohort began in November 1996 and completed in February 1998. This file is similar to the monthly CPS files that contain eight rotation groups. This data set includes a total of 69,618 households.

We included in our analysis file 51,289 households (about 74% of all the households in the sample during this time) that were eligible for interview all of the months that they were in sample. Of these households, 7148 (13.9%) had some combination of interviews and nonresponses (i.e., refusals, noncontacts, other noninterviews), while 43,294 households (84.4%) were interviewed all months they were in sample. There were an additional 847 (1.7%) households that were nonrespondents all months they were in sample that are not included in any analyses. These complete nonrespondents represented less than 11% of the nonresponding cases. A total of 123,539 persons (87%) were in households that were interviews all their months in sample and 18,604 persons (13%) were in households that had at least one interview and at least one nonresponse. For purposes of the present paper, we do not distinguish between persons who were within the household all eight months and those who may not have been living in the household one or more of those months.

2.3 Nonresponse Weighting Adjustment

The current CPS unit nonresponse adjustment is performed within sample PSUs that are assigned to clusters that are similar in Metropolitan Statistical Area (MSA) status and size, and generally are within the same state. Similarly, non-MSA PSUs are also assigned to clusters with other non-MSA PSUs. Each MSA cluster is split into separate cells for "central cities" and "not central cities," while non-MSA clusters are split into separate cells for urban and rural areas. There are currently a total of 250 nonresponse-adjustment-weighting cells in the CPS. These cells may be collapsed if there are fewer than 50 unweighted interviewed households in a particular cell or if the nonresponse adjustment factor exceeds certain limits (see below). Typically, less than 10 cells are collapsed in a month. In the current dataset, after collapsing, there were a total of 242 cells.

The unit nonresponse adjustment factor is normally computed in CPS by dividing the total number of weighted, eligible households within a nonresponse adjustment cell by the weighted number of interviewed households. This has the effect of weighting up the respondent households to equal the total number of (weighted) eligible households. This was also performed in the present investigation except that we treated the complete respondents (interviewed all months they were in sample) as the respondents and treated the partial respondents (interviewed at least once with at least one nonresponse) as the nonrespondents. This adjustment factor was then multiplied by the baseweights to create a nonresponse-adjusted weight.

3 RESULTS

3.1 Comparison of Complete and Partial Respondents on Demographic Characteristics

In the initial analyses cases were weighted to reflect the probability of selection. All statistical tests and calculation of standard errors were conducted in WesVarPC (Westat, 1997) and take into account the complex design of the CPS sample. Due to space limitations, only some characteristics are shown in Table 1. The second column shows the percentage distribution for each characteristic for the complete respondents and the third column shows the distribution for the partial respondents. There were significant differences between partial and complete respondents for region of the country, urbanicity, and poverty of the area (all p's $< .01$). There was a relatively greater proportion of households that were partial respondents in the Northeast and West and in central cities, and in high poverty areas, while a relatively greater proportion of households were complete respondents the Midwest, South, in rural areas, and in low poverty areas. Partial respondent households were also relatively more likely than complete respondent households to be occupied by renters, have fewer persons living there (1 or 2), and to be headed by a single person (all $p's < .01$). In addition, members of partial respondent households were relatively less likely to answer a question on total family income than members of complete respondent households.

There were also significant differences between complete and partial respondents on a variety of person-level demographic characteristics. Households that were partial respondents were relatively more likely to contain persons who were Black, Hispanic, who had graduated college, who were 25-54 years of age, who had never been married, and were not related to the householder than households that were complete respondents. Complete respondent households were relatively more likely than partial respondent households to contain persons who were white, under 19 years old or over 65 years old, who had not earned a High School Diploma, and who were married.

Table 1. Comparison of Partial and Complete Respondents and Comparison of Estimated Total (Nonresponse Adjusted Complete Respondents) with Actual Totals (Complete and Partial Respondents). (Percent Distribution with standard errors in parentheses; * $p < .05$, ** $p < .01$)

Characteristic	Partial Respondent	Complete Respondent	Estimated Total	Actual Total
Number of Persons				
1	27.44	21.64**	21.67	22.45**
	(.82)	(.27)	(.28)	
2	33.74	31.43**	31.34	31.76*
	(.56)	(.18)	(.19)	
3	16.38	17.63*	17.62	17.45
	(.56)	(.22)	(.23)	
4	13.60	16.78**	16.78	16.34*
	(.50)	(.17)	(.18)	
5 or more	8.84	12.52**	12.59	12.00**
	(.27)	(.17)	(.18)	
Household Income				
DK/Refused	16.57	4.82**	4.84	6.47**
	(.80)	(.10)	(.10)	
Race				
White	78.94	84.25**	83.96	83.56
	(.62)	(.25)	(.25)	
Black	15.60	11.01**	11.13	11.60
	(.47)	(.30)	(.30)	
American Indian	1.12	.95	.95	.97
	(.20)	(.05)	(.05)	
Asian/Pacific Islander	4.34	3.80	3.96	3.87
	(.48)	(.21)	(.22)	
Age				
0-15 years	20.57	22.80**	22.82	22.51*
	(.47)	(.12)	(.13)	
16-19 years	5.48	6.26**	6.24	6.16
	(.24)	(.09)	(.09)	
20-24 years	7.09	6.95	6.94	6.97
	(.38)	(.07)	(.07)	
25-34 years	17.18	14.84**	14.88	15.14**
	(.31)	(.10)	(.10)	
35-44 years	18.30	16.17**	16.18	16.44*
	(.33)	(.11)	(.11)	
45-54 years	13.56	12.48**	12.47	12.62
	(.24)	(.14)	(.15)	
55-64 years	8.25	8.04	8.02	8.06
	(.18)	(.09)	(.09)	
65 + years	9.56	12.47**	12.44	12.09*
	(.21)	(.16)	(.16)	
Education				
< High School	18.90	22.44**	22.45	21.97*
	(.45)	(.20)	(.21)	
HS Diploma only	32.17	31.56	31.41	31.64
	(.43)	(.28)	(.29)	
Some College	25.11	25.42	25.38	25.38
	(.38)	(.16)	(.17)	
Bachelors Degree +	23.82	20.59**	20.75	21.02
	(.62)	(.28)	(.27)	

In the next phase of analyses, the nonresponse adjustment cells were used to create weights for the complete respondents to bring them to a level that would reflect the total sample without including the partial respondents. These estimated totals are shown in column 4 of Table 1. Because we know these characteristics for the partial respondents, we can compare the estimated totals from the nonresponse-weighted complete respondents to the actual totals of the complete and partial respondents (shown in the last column of Table 1). We used the standard errors of the estimated totals to construct 95 percent confidence intervals and tested whether the point estimate of the actual totals fell within the confidence interval of the estimated total. All of the percentages in actual totals for region of the country, urbanicity, and poverty of the area fell within the confidence intervals of the estimated totals. This is quite expected since the nonresponse adjustment uses geographic information to create the weighting adjustment cells. However, as can be seen in the last two columns of Table 1, the nonresponse adjustment of the complete respondents underestimates the percentage of one and two person households and the percentage of households that did not provide income data. The actual totals for the different racial groups all fell within the confidence intervals of the estimated totals, but the nonresponse adjustment of the complete respondents overestimates the percentage of children less than 15 years old and adults over 65 years but underestimates the percentage of adults age 25-44 years. The nonresponse adjusted complete respondents also overestimates the percentage of persons with less than high school education (see Table 1).

3.2 Comparison of Complete and Partial Respondents on Labor Force Characteristics

Unlike most of the demographic characteristics noted above, a person's labor force classification may change from month-to-month during the time that they are in the CPS sample; therefore, one cannot determine with certainty a person's labor force classification during the months the person was a nonrespondent. However, for illustrative purposes, we can compare the differences in labor force classification between the complete respondents and partial respondents by using the labor force status of the partial respondents each time they were interviewed and had a labor force status. Given space limitations we will ignore the month-in-sample and count complete respondents eight times, while partial respondents may be counted at most seven times, since they had at least one nonresponse. As can be seen in Table 2, partial respondents were relatively more likely to be in the labor force, to be employed, and to have a higher unemployment rate than complete respondents. As can be seen in the last two columns of Table 2, the nonresponse adjustment of the complete respondents underestimates the percentage of persons who are in the labor force and the percentage employed.

Table 2. Comparison of Partial and Complete Respondents and Comparison of Estimated Total (Nonresponse Adjusted Complete Respondents) with Actual Totals (Complete and Partial Respondents). (Percent Distribution with standard errors in parentheses).

Labor Force Status	Partial Respondent	Complete Respondent	Estimated Total	Actual Total
All Months-in-Sample				
Civilian Labor Force	67.99	65.29**	65.23	65.56*
	(.52)	(.14)	(.13)	
Employed	64.71	62.36**	62.28	62.60**
	(.48)	(.13)	(.12)	
Unemployment Rate	4.83	4.48	4.52	4.52
	(.17)	(.08)	(.08)	

*$p < .05$, ** $p < .01$

4 DISCUSSION

The present results have possible implications for post-survey nonresponse adjustment procedures as well as field procedures. Current CPS nonresponse adjustment procedures utilize only geographic information. The present results show that this is insufficient in reducing nonresponse bias for some demographic and labor force characteristics. A major obstacle in improving unit nonresponse adjustment is the fact that very little information is currently obtained in the field about nonrespondents. Changes in field procedures that put more emphasis on obtaining more characteristics of the household and persons living there is essential in furthering research and understanding of nonresponse and improving nonresponse adjustment (Groves and Couper, 1995; Madow et al., 1983). In addition, further research should also be conducted on utilizing prior information obtained from the household for use in making nonresponse adjustments. Although this would not help with the cases their first month-in-sample, these data are potentially very useful for nonresponse in other MIS, and there are many more possibilities for enhanced and refined nonresponse adjustments. The current results suggest efforts in this area may be quite worthwhile.

The nonresponse-adjusted weights used in the present investigation did not include the later stages of weighting that are conducted as part of the normal monthly estimation process in CPS. These adjustments, because they use demographic data, are likely to have larger effects on reducing these kinds of differences observed between respondents and nonrespondents. The effect of these weighting adjustments on reducing the observed differences between respondents and nonrespondents is an important next step in this research. In addition, although nonresponse adjustment procedures have typically not taken into account the kind of nonresponse, it may be important to distinguish between refusals, noncontacts, and other noninterviews because the underlying causes of each type of nonresponse may be quite different. Recently, Groves and Couper (1995, 1998) have shown that contact with a household and refusal to participate have different correlates and have suggested that models based on these could be used to improve nonresponse adjustment procedures. This approach represents a promising area for further research.

REFERENCES

Goyder, (1987). *The Silent Minority: Nonrespondents in Sample Surveys.* Boulder, CO: Westview Press.

Groves, R. M. and Couper, M. P. (1995). Theoretical motivation for post-survey nonresponse adjustment in household surveys. *Journal of Official Statistics, 11,* 93-106.

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology, 12,* 1-16.

Madow, W. G., Nisselson, H., and Olkin, I. (1983). Incomplete Data in Sample Surveys. New York: Academic Press.

Oh, and Scheuren, F. (1983). Weighting adjustment for unit nonresponse. In Madow et al. (Eds.) *Incomplete data in sample surveys,* vol. 2, pp. 143-184. New York: Academic press.

Stasny, E. A. and Fienberg, S. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the conference on gross flows in labor force statistics.* Washington, DC.

Tucker, C., and Harris-Kojetin, B. A. (1997). *The Impact of nonresponse on the Unemployment Rate in the Current Population Survey.* Paper presented at the 8th International Workshop on Household Survey Nonresponse, Mannheim, Germany.

Westat. (1997). *A Users Guide to WesVarPC.* Rockville, MD: Westat.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

CALCULATION OF CHANGE FOR ANNUAL BUSINESS SURVEYS

Pascal Rivière[1]

ABSTRACT

The disseminated results of annual business surveys inevitably contain statistics that are changing. Since the economic sphere is increasingly dynamic, a simple difference of aggregates between n-1 and n is no longer sufficient to provide an overall description of what has happened. The change calculation module in the new generation of annual business surveys divides overall change into various components (births, deaths, inter-industry migration) and calculates change on the basis of a constant field, assigning special importance to restructurings. The main difficulties lie in establishing subsamples, reweighting, calibrating according to calculable changes, and taking account of restructurings.

KEY WORDS: Change; sampling; business demographics; restructuring.

1. INTRODUCTION

Annual business surveys are conducted in France by six survey sections, respectively covering manufacturing, agricultural and food industries, commerce, services, construction and transportation. In the new generation of business surveys, a number of statistical methods have been adjusted and refined in a way that is common to the various survey sections.[2] The calculation of change is one of the topics being investigated. The entire problem lies in determining relevant changes, that is, in taking account of an breaking down the group of factors that contribute to the change in an economic aggregate.

The change calculation module described here, which is common to the various annual business surveys, applies to some fifteen core variables used in all those surveys (number of employees, sales, value added, etc.), and additional variables considered useful by individual survey sections.

2. BASIC CONCEPTS

Consider a quantitative variable X for which we want to calculate the change from one year to another. Let T_n denote the total of this variable for the industry considered in year n. In itself, this calculation poses no particular problem. The change in the variable between n-1 and n is written as follows:

$$\Delta_{m / n-1} = \frac{T_n - T_{n-1}}{T_{n-1}} \qquad (1)$$

But this expression, while not false, yields only an "apparent change", and is not sufficient to truly represent the change, since from one year to the other the field for which the aggregates are calculated has evolved, owing, for example, to the demographics of the units. If we want to represent the change over the

[1] Pascal Rivière, INSEE, Timbre E210, 18 Bd A. Pinard, 75675 PARIS CEDEX 14

[2] See Rivière, 96b.

period realistically, it is therefore necessary to provide other information: to break this change down among different subpopulations, to attempt to determine changes on the basis of a constant field, etc.
Thus, in practical terms, the change calculation module provides two indexes: **apparent change**, described above, and "**pure**" **economic change** (in which inter-industry shifts and restructurings are dealt with); added to this, for each change, is the breakdown of these indexes into different components: ongoing firms, deaths, births, firms migrating from one industry to another, and restructured firms.

It should be noted that the "births" category includes units that were inactive the previous year; by the same token, businesses that become inactive are included in the "deaths" category. The "entries to the industry" category includes units that were outside the industry in n-1, and, the "exits from the industry" category includes *for purposes of apparent change* firms that have left the industry, but these are not taken into account in determining economic change.

Calculations were carried out by activity, in the entire field, in NAF (nomenclature d'activités fançaise - French industrial classification) or in NES (nomenclature économique de synthèse - general economic classification). The level of detail (class, group or division in NAF) is a parameter of the program.

For the part of the change calculation that concerns restructurings, we have *restructuring envelopes*, which in particular include the list of input units (reflecting the situation before restructuring) and the list of output units (representing the situation after restructuring): this enables us to calculate changes on a "constant field" basis.

3. FORMALIZATION[3]

The population is divided into six categories, numbered as follows:

(1) Ongoing firms: firms that exist in *n* and *n-1* and remain in the industry
(2) Firms created in n
(3) Firms that ceased to exist in n
(4) Firms that entered the industry in *n* and were not in the industry in *n-1*
(5) Firms that exited the industry in *n*
(6) Firms that have taken part in a restructuring.

Let T_n^k denote the total of variable *X* for all firms in category *k*, belonging to the reference universe in year *n*, for the industry considered; T_{n-1}^k will also be the total for category *k*, but for year *n-1*; T_n and T_{n-1} will be the respective totals of the variable for the industry, regardless of the category, in *n* and *n-1*.

3.1 Apparent change

We therefore have: $T_n = \sum_{k=1}^{6} T_n^k \,; T_{n-1} = \sum_{k=1}^{6} T_{n-1}^k$. The apparent change $\Delta_{n/n-1}$ is rewritten as:

$$\Delta_{n/n-1} = \frac{T_n^1 + T_n^2 + T_n^4 + T_n^6 - T_{n-1}^1 - T_{n-1}^3 - T_{n-1}^5 - T_{n-1}^6}{T_{n-1}} \qquad (2)$$

Suppose that $\alpha_t^k = \dfrac{T_1^k}{T_t}$, and $\beta_{n-1}^k = \dfrac{T_n^k}{T_{n-1}} = \dfrac{T_n^k}{T_n} \cdot \dfrac{T_n}{T_{n-1}} = \alpha_n^k . \left(1 + \Delta_{n/n-1}\right)$. After simplifying, we obtain:

[3] This section is largely based on Christine, 95.

$$\Delta_{n/n-1} = \alpha_{n-1}^{1}.\Delta^{1} + \alpha_{n}^{2}\left(1+\Delta_{n/n-1}\right) + \alpha_{n}^{4}\left(1+\Delta_{n/n-1}\right) - \alpha_{n-1}^{3} - \alpha_{n-1}^{5} + \alpha_{n-1}^{6}.\Delta^{6} \tag{3}$$

$$\Delta_{n/n-1} = \frac{\alpha_{n-1}^{1}.\Delta^{1} + \alpha_{n}^{2} + \alpha_{n}^{4} - \alpha_{n-1}^{3} - \alpha_{n-1}^{5} + \alpha_{n-1}^{6}.\Delta^{6}}{1-\alpha_{n}^{2} - \alpha_{n}^{4}} \tag{4}$$

In this expression, the α_{t}^{k} elements represent the "weight" of firms in category *k*, for variable *X*, in relation to all firms at time *t*; Δ^{1} and Δ^{6} represent *category-specific* changes, associated with categories 1 (ongoing firms) and 6 (restructured firms) respectively:

$$\Delta^{1} = \frac{T_{n}^{1} - T_{n-1}^{1}}{T_{n-1}^{1}}\,;\, \Delta^{6} = \frac{T_{n}^{6} - T_{n-1}^{6}}{T_{n-1}^{6}} \tag{5}$$

Equation (3) gives the breakdown of the apparent change into its constituent parts. Here it is necessary to distinguish between contribution, change and weight: thus, the *weight* of ongoing firms is α_{n-1}^{1}, their *change* is Δ^{1}, and their *contribution* is $\alpha_{n-1}^{1}\,\Delta^{1}$; the contribution of categories 2 and 4 is of the form $\alpha_{n}^{k}\left(1+\Delta_{n/n-1}\right)$, and that of categories 3 and 5 equals $-\alpha_{n-1}^{k}$. Formula (4) shows the advantage of providing an equation solved in $\Delta_{n/n-1}$.

The objective of the change calculation model is therefore to estimate:

- weights $\alpha_{n-1}^{1}, \alpha_{n-1}^{3}, \alpha_{n-1}^{5}, \alpha_{n-1}^{6}, \alpha_{n}^{2}$ and $\alpha_{n}^{4}; \beta_{n-1}^{2}$ and β_{n-1}^{4};
- changes Δ^{1} and Δ^{6};
- the 6 contributions: $\alpha_{n-1}^{1}.\Delta^{1}, \alpha_{n}^{2}\left(1+\Delta_{n/n-1}\right), -\alpha_{n-1}^{3}, \alpha_{n}^{4}\left(1+\Delta_{n/n-1}\right), -\alpha_{n-1}^{5}, \alpha_{n-1}^{6}.\Delta^{6}$

3.2 Pure economic change

Our objective here is to calculate a "constant field" change. For this purpose, we will take as our basis the list of units in industry s in year n (plus the units of s that ceased to exist in n). Thus, for the units entering industry s in n, we will recover their data for n-1, even though they were not in the industry at that time. Conversely, units leaving the industry between n-1 and n will not be taken into account in the calculation.

The constant field principle also calls for a special treatment of restructurings: in economic change, a unit that is part of a restructuring will be assigned to the industry to which its restructuring envelope belongs, and not to its own industry.

Lastly, we will systematically take account of the employment size group in the calculations: we always place ourselves in a given industry s and a given employment size group *t*.

Following a breakdown similar to the one for apparent change, we thus recalculate an overall change. We again break this change down into 6 factors, but they are no longer entirely the same. This time the factors are: ongoing firms remaining in the same employment size group; births; deaths; entries to the industry; firms that have participated in a restructuring; and entries to the employment size group (ongoing units which did not migrate from another industry, but which belonged to another employment size group in n-1). The economic change is written in the following form:

$$\Delta^{e}_{n/n-1} = \alpha'^{11}_{n-1}.\Delta'^{11} + \alpha'^{12}_{n-1}.\Delta'^{12} + \beta'^{2}_{n-1} + \alpha'^{4}_{n-1}.\Delta'^{4} - \alpha'^{3}_{n-1} + \alpha'^{6}_{n-1}.\Delta'^{6} \qquad (6)$$

If we compare this to the apparent change, we observe that logically, the term for firms exiting the industry disappears by construction, and that the category of ongoing firms has split into two (category 11, consisting of those that remain in the same employment size group; and category 12, comprising those that have changed from one group to another). The specific changes are also more numerous, once again by construction.

Note: while the apparent change may be calculated by any user of statistics, this is not the case with the economic change. Thus, it will not be necessary to calibrate according to some total, since it is in fact the sum of the components that yields the economic change.

4. PARTITIONING UP THE SAMPLE

One of the difficulties of calculating change lies in determining the subsamples that will be used as the basis for the calculation, in n and n-1. This determination is closely linked to the tool used to draw the samples. This is described in detail in Rivière, 96a.

4.1 Characteristics of samples

For drawing samples, annual business surveys use the OCEAN sample selection and co-ordination system [OCEAN: outil de coordination des enquêtes annuelles - annual surveys co-ordination tool], used in France since 1990 and described in Cotton, 1989. The draws are stratified simple random samples. The stratification variables remain the same from one year to the other: main activities at the finest level, employment size group, region. The strata do not change, except for the recent introduction of a new employment size group. There is a take-all portion to the sample, roughly corresponding to units with at least 20 employees. Clearly, then, many elements remain unchanged, but the sampling rate per stratum can vary in relation to the previous years.

The OCEAN sampling frame is divided into two parts, of equal size in expectation. From one year to another, one of these two parts remains the same: exactly the same units are retained; this is what is known as the retained part of the sample. In the second half-frame, a new sample selection, independent of the preceding one, is carried out: this is the renewed part of the sample. The following year, we switch: the renewed part will be from the first half-frame, and the retained part from the second.

4.2 Determination of subsamples

The theoretical definition of subpopulations does not pose any major problem. On December 31 of year n, in the annual business surveys database, the unit has a certain number of characteristics: inclusion in the field or not, status (active or not), main activity, employment size group, presence of absence of restructuring. This information, in n-1 and n, is sufficient to determine the change category to which the unit belongs.

Putting this into practice is not as simple a matter, since in fact we do not have this information for the entire sampling frame. In particular, when a unit is new to the sample, it is not possible to know whether or not it has migrated from another industry or become in scope between n-1 and n.

There are two cases in which we have all the information: firstly, births, identified in the inventory; and secondly, restructurings. Since we are confining ourselves to known restructurings from the CITRUS system, they are identified, in effect by construction.

In the other cases, it is essential to focus on a subsample for which the information is available. For this, we confine ourselves to units from the sample for year n that were already in the n-1 sample. All that is necessary, then, is to take the retained part of the sample, plus what is called the "renewed-retained" part: units in the renewed part that were sampled in n-1.

5. ESTIMATION

The subsamples for the different change categories are now defined, in n-1 and n. Combining them yields an overall subsample, to which our estimates will then apply. Now we must first determine the new weights to be applied to the units for the purpose of calculating change, then specify the estimators.

5.1 Principles

The principle for determining weights is, in the initial approximation, fairly straightforward:

- the weight would be equal to 1 in the take-all portion;
- it would be equal to the sampling weight where we have the entire sample for the calculation of change;
- it would be equal to twice the sampling weight when we are working on the part retained between *n-1* and *n*.

However, such a method cannot be applied, owing to four supplementary effects:

- the presence of a "renewed - retained part", of variable size;
- the existence of tabulatable units that we do not want to incorporate into the calculation of change;
- the possibility that a sample will change in size;
- the fact that reweighting must depend on the change category considered.

In what follows, we will deal solely with the reweighting due to the renewed-retained part, as well as the matter of changes in the sampling rates.[4] Note that in the case of units participating in restructuring, the weight to be applied is always rate *j* fraction and the problems described below do not arise.

5.2 The problem of the renewed - retained part

We will place ourselves in a given survey stratum *s* in year *n*. This stratum may be divided in two: into a retained part and a renewed part. By construction, the sum of the weights of the units in the retained part must be equal to the number *N* of units in the stratum divided by 2. Our goal will be to multiply each weight by a factor λ, so that the sum of the weights will be equal to *N*.

We can easily find:

$$\lambda = 2\frac{\underset{i\in retained-part}{\Sigma p_i}}{\underset{i\in retained-part-extended}{\Sigma p_i}} \tag{7}$$

5.3 Change in sampling ratio between *n-1* and *n*

If the sample size varies between *n-1* and *n*, the question that arises is no longer that of the multiplicative factor, but rather of the sampling weight by which that factor will be multiplied - which of the two? Of course, the question does not arise for births, firm resuming activity, firms returning from out of scope and deaths. For the others, the idea, insofar as possible, [5] is to calculate changes by taking account of the same units in both periods: it is in fact the overlap between the two samples that is taken, if we refer to the definition of the subpopulations. Thus, if one sample contains the other, it is the smaller one that will serve as our reference, meaning that we will use the higher weights.

[4] For further details on the rest, see Rivière, 98.

[5] That is, except in one specific case, namely where take-all thresholds are crossed.

Consequently, in all calculations, the sampling weight q_i to be used will systematically be: $q_i = max(p_i(n), p_i(n-1))$ instead of p_i, except if p_i equals 1.

5.4 Calibration

The aggregates T_n and T_{n-1} are actually estimated, independently of any calculation of change. Applying the Horvitz-Thompson formula, let $\hat{T}_n$ and $\hat{T}_{n-1}$ denote the corresponding estimates. Then

$$\hat{T}_n = \sum_i p_{ni} Y_{ni} \,;\, \hat{T}_{n-1} = \sum_i p_{n-1,i} Y_{n-1,i} \qquad (8)$$

Since these aggregates are published, the apparent change is accessible to any user of statistics. But is happen that if we calculate the six components of the change in the "natural" way, there is every change that in adding them up, we will not obtain the apparent change, because the samples are different.

To solve this, we calculate the Horvitz-Thompson estimators of: $T_n^1, T_{n-1}^1, T_n^2, T_{n-1}^3, T_n^4, T_{n-1}^5, T_{n-1}^6, T_n^6$, with the appropriate weight for each unit. In each case, we merely apply the formula, but we apply it to the subsample associated with each subcategory. We thus obtain estimators $\tilde{T}_n^i$ and $\tilde{T}_{n-1}^i$.

The aggregates *estimated by summing the contributions* are therefore:

$$\tilde{T}_n = \tilde{T}_n^1, \tilde{T}_n^2, \tilde{T}_n^4, \tilde{T}_n^6 \,;\, \tilde{T}_{n-1} = \tilde{T}_{n-1}^1, \tilde{T}_{n-1}^3, \tilde{T}_{n-1}^5, \tilde{T}_{n-1}^6 \qquad (9)$$

And it is clear that $\tilde{\Delta}_{n/n-1} = \dfrac{\tilde{T}_n - \tilde{T}_{n-1}}{\tilde{T}_{n-1}} \neq \hat{\Delta}_{n/n-1} = \dfrac{\hat{T}_n - \hat{T}_{n-1}}{\hat{T}_{n-1}}$.

To estimate the components of the change, note firstly that the components relating to newly created firms and restructurings can be estimated as such: $\hat{T}_n^2 = \tilde{T}_n^2, \hat{T}_n^6 = \tilde{T}_n^6, \hat{T}_{n-1}^6 = \tilde{T}_{n-1}^6$.

The other $\hat{T}_k^i$ components will be estimated on the basis that their share of the total, truncated (since we remove Part 6 and possibly Part 2) is the same for $\tilde{T}_k$ and $\hat{T}_k$. This gives us:

$$\frac{\hat{T}_n^i}{\tilde{T}_n^i} = \frac{\hat{T}_n - \tilde{T}_n^2 - \tilde{T}_n^6}{\tilde{T}_n - \tilde{T}_n^2 - \tilde{T}_n^6} (i = 1,4); \quad \frac{\hat{T}_{n-1}^i}{\tilde{T}_{n-1}^i} = \frac{\hat{T}_{n-1} - \tilde{T}_{n-1}^6}{\tilde{T}_{n-1} - \tilde{T}_{n-1}^6} (i = 1,3,5) \qquad (10)$$

In short, this method of calculation, from which other parameters are deducted, ensures consistency with the apparent change.

6. TAKING RESTRUCTURINGS INTO ACCOUNT

Since the 1996 fiscal year, restructurings have been incorporated into the annual business surveys database by means of the CITRUS information system, which is described in Corbel, 1996. Restructurings especially complicate the statistical analysis, particularly since they often involve very large units and thus carry a heavy weight in the statistics. Thus a business enterprise absorbed by an industrial enterprise simply disappears from the commercial sector. It will contribute negatively to economic changes in commerce and positively to those in industry. Large restructurings may therefore introduce breaks in structural statistics, which it is important to be able to explain. This is why the CITRUS system is of interest.

To be able to calculate constant-field change, CITRUS creates restructuring envelopes. Each envelope contains the *list of firms prior to restructuring* and the *list of firms after restructuring*. For each envelope, a *main activity for the envelope* is also calculated. Thus, for economic change, the units involved in a restructuring will be classified in the industry to which their envelope belongs, which may be different from the industry to which the firm belongs. Finally, for each envelope, a multiplier known as the *aggregation coefficient* is determined. The later is used for dealing with non-additive variables, such as the sales figure: in the case of a merger of two units A and B, for example, the sales that took place between A and B in year n-1 obviously no longer appear in n. To the extent possible, they must be eliminated from the calculation. Hence the value of applying, when calculating the change in sales, a multiplier in n-1 to eliminate the effect of "intra" exchanges.

Theoretically, all this lends itself to generating constant-field changes. In practice, things are not so simple. For the 1996 fiscal year, we counted 816 restructuring operations in CITRUS, with the total amount of assets transferred in all these operations exceeding 1 trillion francs. The 698 restructuring envelopes that result for the annual business surveys contain 1,757 firms. For 716 of them, or nearly half, we have no individual business survey data, especially in n-1 (units out-of-sample , outside the field, non-responding, etc.). Calculating the aggregation coefficient is therefore pointless in many cases; but above all, this lack of data has a negative effect on the calculation of economic change as such.

While the idea of constant-field change for restructurings is not easy to put into practice, the fact remains that the calculation of change provides the user with highly informative results. The problem, however, is to sort through the mass of indicators. The challenge at this point, then, is no longer to refine the calculation methodology, but rather to develop another methodology - one that will enable us to make good use of the results of this new program.

REFERENCES

Christine, M. (95) Note sur les calculs d'évolution, *note INSEE n° 360/F230*

Corbel, P. (96) CITRUS, un nouvel outil de la statistique d'entreprise, *Le courrier des statistiques n° 78*

Cotton, F. (89), OCEAN, outil de coordination des enquêtes annuelles, *Le courrier des statistiques n° 52*

Rivière, P. (96a) Les calculs d'évolution sectorielle dans l'EAE4G, *note INSEE n° 182/E210*

Rivière, P. (96b) Enquêtes annuelles d'entreprise: à la rencontre du 4e type, *Le courrier des statistiques n° 78*

Rivière, P. (98), Les calculs d'évolution dans les enquêtes annuelles d'entreprise, *document INSEE E9801*

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

VARIANCE ESTIMATION IN LONGITUDINAL STUDIES OF INCOME DYNAMICS

Susana R. Bleuer[1] and Milorad S. Kovačević[2]

ABSTRACT

We address the problem of estimation for the income dynamics statistics calculated from complex longitudinal surveys. In addition, we compare two design-based estimators of longitudinal proportions and transition rates in terms of variability under large attrition rates. One estimator is based on the cross-sectional samples for the estimation of the income class boundaries at each time period and on the longitudinal sample for the estimation of the longitudinal counts; the other estimator is entirely based on the longitudinal sample, both for the estimation of the class boundaries and the longitudinal counts. We develop Taylor linearization-type variance estimators for both the longitudinal and the mixed estimator under the assumption of no change in the population, and for the mixed estimator when there is change.

Key Words and Phrases: Gross Flows, Longitudinal Proportions, Taylor Linearization.

1. INTRODUCTION

Estimates of gross flows and transition rates between different income classes from longitudinal surveys are required in studies of income dynamics. The boundaries of these income categories are often defined by linear functions of income quantiles which have to be estimated from the survey as well. For example, one measure of income inequality that is frequently used, the low income line (Low Income Measure), is defined as half of the median income, where income is adjusted for family size. Thus, estimators of counts of transitions to and from "low income" require the estimation of the income medians at the time period of interest.

In this study, the income class boundaries refer to the respective cross-sectional populations. We estimate these from the cross-sectional samples to obtain design-consistent estimators. If the change in population from one wave to the other is negligible, a longitudinal sample, defined as the intersection of the cross-sectional samples, may represent the population at both time points, and we could estimate the income categories from the longitudinal sample. Otherwise, estimation of income categories from the longitudinal sample may yield biased estimates.

Two design-based approaches are considered for the estimation of gross flows and transition rates, under the assumption of no births between the two waves. One approach is based on the cross-sectional samples for the estimation of the class boundaries at each time period and on the longitudinal sample for the estimation of counts of units in the longitudinal population (longitudinal counts). This results in an estimator that we term *the mixed estimator*. The other approach uses an estimator based on the longitudinal sample for both the class boundaries and the longitudinal counts, and we call it *the longitudinal estimator*.

On the other hand, if the number of births between the two waves considered is relatively large, a new panel is usually added to the sample to represent better the population in the second wave. In this situation the

[1] Susana R. Bleuer, Business Survey Methods Division, Statistics Canada, Ottawa, K1A0T6, Canada, rubisus@statcan.ca

[2] Milorad Kovačević, Social Survey Methods Division, Statistics Canada, Ottawa, K1A0T6, Canada. kovamil@statcan.ca

longitudinal estimator would no longer have "good " properties.

The main objective of this study is to develop Taylor-linearization variance estimators for these statistics. Variance estimation has to take into account the variability of the income class boundaries, as well as changes in the population over time and attrition. In section 2 we define the parameters of interest. In section 3 we define the longitudinal and the mixed estimators, and we state the conditions for their consistency. We also show that the Taylor linearization method yields different formulas for change vs. no-change in the population and attrition vs. no-attrition. For the calculation of the variance we assume that the model of compensation for the missing data is correct and the resulting estimates are unbiased under the model.

2. CONCEPTS, ASSUMPTIONS, DEFINITIONS

Let U_t represent the population at time t=0,1. We define the longitudinal population at times 0 and 1 by $U_{0,1} = U_0 \cap U_1$. "Deaths" and "births" from one time period to the next are considered as change in the population. By "deaths" we mean real deaths and/or emigration; similarly, "births" means real births and/or immigration. Let U_D denote the set of individuals who belong to the population U_0 at time $t=0$ and do not belong to the population U_1 at $t=1$ due to "death". Let U_B denote the set of "births" from time 0 to 1. Hence, the population at $t=1$ can be expressed as $U_1 = (U_0 \setminus U_D) \cup U_B$, and the longitudinal population can be then expressed as $U_{0,1} = U_0 \setminus U_D = U_1 \setminus U_B$. Let N_0, N_1, N_L, N_D and N_B denote the sizes of the populations U_0, U_1, $U_{0,1}$, U_D and U_B respectively.

Let s_0 and s_1 denote the samples representing the population U_0 and U_1 respectively. Let s_d be the sub sample of individuals in s_0 who "died" between $t=0$ and $t=1$, and let s_b be the sub sample of individuals in s_1 "born" between $t=0$ and $t=1$, such that s_d is a representative sample of U_D, and s_b is a representative sample of U_B. In this study the longitudinal sample is defined in terms of two waves by $s_L = s_0 \cap s_1$.

Non-respondents to the initial wave at $t=0$ exist but they are relatively few compared to non-respondents in later waves. When there is attrition, non-respondents cumulate over time, and the longer the study lasts, the greater is the non-response. For the sake of simplicity, assume that s_0 is a sample with the initial non-response removed and with the associated weights already adjusted for it.

We assume a stratified two-stage design with a large number H of strata and relatively few clusters or primary sampling units (PSU) sampled within each stratum. Although the number of strata does not change from one time period to the next, the number of clusters within a stratum and/or the size of the clusters may change to account for "births" and "deaths" in the population.

Without loss of generality we consider only the longitudinal low income proportions. The results can be extended easily to other parameters of income dynamics. Let x_{hij} and y_{hij} be the family income adjusted for family size for the j-th ultimate unit in the i-th PSU of stratum h at times t=0 and t=1 respectively. For the sake of simplicity we denote by j the individual j in cluster i of stratum h. Then the proportion of individuals with income less than or equal to x at time 0 and income less than or equal to y at time 1 is given by

$$F(x,y) = \frac{1}{N_L} \sum_{j=1}^{N_L} I(x_j \leq x)\ I(y_j \leq y) \qquad (1)$$

where N_L is the total number of the units in the longitudinal population $U_{0,1}$. I is the indicator function. $F(x,y)$ is the bivariate distribution function of incomes at times 0 and 1. Note that the sum in (1) is over the units in U_0 that did not "die" by $t=1$.

The Low Income Proportion at times t=0 and t=1 (the two-wave LIP) is given by $\theta_0 = F(l_0,l_1)$ where $l_0 = M_0/2$ is half the median income at time t=0 and $l_1 = M_1/2$ is half the median income at $t=1$. The following situations arise:

1. ***No change in population***: there are no births nor deaths from time t=0 to time t=1. The size of the

longitudinal population coincides with the sizes of the marginal populations. Then the marginal proportions of individuals with income less or equal than x and y respectively, can be expressed by

$$F_0(x) = F\ (x, +\infty) = \sum_{j=1}^{N_0} I(x_j \leq x) / N_0 \ ,$$

and

$$F_1(y) = F\ (+\infty, y) = \sum_{j=1}^{N_1} I(y_j \leq y) / N_1 \ .$$

2. ***Change in population***: there are either births or deaths or both, from time t=0 to time t=1. Then the marginal proportions can be expressed by

$$F_0(x) = c_0\ F\ (x, +\infty) + c_D\ F_D(x) \ ,$$

where $c_0 = N_L / (N_L + N_D)$, $c_D = N_D / (N_L + N_D)$ and $F_D(x)$ is the proportion of individuals in U_D with income less than or equal to x, and

$$F_1(y) = c_1\ F\ (+\infty, y) + c_B\ F_B(y) \ ,$$

where $c_1 = N_L / (N_L + N_B)$, $c_B = N_B / (N_L + N_B)$ and $F_B(y)$ is the proportion of individuals in U_B with income less than or equal to y.

3. ***Attrition***: the model for non-response in the second wave is that of the observations missing at random within response classes (M.A.R.): $p_m(I_j(r) = 1) = p_a$ if individual j belongs to response class "a", where $I_j(r)$ is the response indicator for individual j. The corresponding weight adjustments are given by the inverse of the probabilities of response: $w_j(r) = 1 / p_a$ if individual j belongs to response class "a".

3. ESTIMATION

A ratio estimator of $F(x,y)$ is given by

$$\hat{F}(x,y) = \sum w_j\ I(x_j \leq x)\ I(y_j \leq y) / \sum w_j, \qquad (2)$$

where the sum is over the sample elements and w_j is the survey weight attached to the sample element j. We set $\hat{N}_0 = \Sigma w_j$. The survey weights w_j are chosen so that $\hat{N}_0 \hat{F}(x,y) / N_0$ is design-unbiased for $F(x,y)$. We denote by $\hat{F}_0(x)$, $\hat{F}_1(y)$, $\hat{l}_0$, and $\hat{l}_1$ the estimators of $F_0(x)$, $F_1(y)$, l_0, and l_1 based on the complete sample at time t = 0,1, defined by $\hat{F}_0(x) = \hat{F}_0(x, +\infty)$, $\hat{F}_1(y) = \hat{F}_0(+\infty, y)$, and $2\hat{l}_0 = \hat{F}_0^{-1}(1/2)$.. Then, under the regularity conditions, $\hat{l}_0$, $\hat{l}_1$, $\hat{F}$ and $\hat{F}(\hat{l}_0, \hat{l}_1)$ are design-consistent estimators of l_0, l_1, F and $F(l_0, l_1)$ respectively. This result is a direct generalization of Shao and Rao's (1993). We denote by $\tilde{F}$, $\tilde{l}_0$ and $\tilde{l}_1$ the estimators of F, l_0 and l_1 based on the longitudinal sample and adjusted for non-response. The adjusted weights are of the form $w_j(L) = w_j \cdot w_j(r)$.

If there is no change in the population from time t=0 to time t=1, we consider the two design-based estimators of $\theta_0 = F(l_0, l_1)$, namely,

$$\hat{\theta}_{Mixed} = \tilde{F}\ (\ \hat{l}_0, \hat{l}_1\ ) \ \text{ with } \ 2\hat{l}_t = \hat{F}_t^{-1}\ (\ 1/2\ ),\ t = 1,\ 2. \qquad (3)$$

and

$$\hat{\theta}_{Long} = \tilde{F}\ (\ \tilde{l}_0, \tilde{l}_1\ ) \ \text{ with } \ 2\tilde{l}_t = \tilde{F}_t^{-1}\ (\ 1/2\ ),\ \ t = 1,\ 2 \qquad (4)$$

If there is change in the population, and the sample at time 1 contains a sub sample representative of the births occurring from time t=0 to time t=1, then we estimate the marginal proportions by

$$\hat{F}_0(x) = c_0 \, \hat{F}\,(x, +\infty) + c_D \, \hat{F}_D(x)$$

where $\hat{F}_D(x)$ is a domain estimator of $F_D(x)$, and

$$\hat{F}_1(y) = c_1 \, \hat{F}\,(+\infty, y) + c_B \, \hat{F}_B(y) \ ,$$

where $\hat{F}_B(y)$ is a domain estimator of $F_B(y)$, based on the birth component of the sample at time t=1. We also define $2\hat{l}_t = \hat{F}_t^{-1}(1/2)$, t = 0,1. We then consider the estimator of θ_0 given by

$$\hat{\theta}_0 = \hat{F}\,(\hat{l}_0, \hat{l}_1) \tag{5}$$

In order to show "consistency" of estimators (3), (4), and (5) and to provide linearization variance estimators, we need to assume a model for non-response and certain limiting conditions about the population and the sampling design. By "consistency" here, we mean that $\hat{\theta} - \theta \rightarrow 0$ in probability, where the probability is defined in a product space determined by the model and the design (there exists a product space and a probability measure defined on it, whose projection onto the design and the model probability spaces coincides with the respective design and model probability measures (Bleuer, 1998)).The assumed conditions are similar to those made by Shao and Rao (1993) for the linearization variance of the cross-sectional Low Income Proportion. In addition, we also assume the attrition model described in the previous section. Therefore, we assume the following: the design is stratified, multistage on a sequence of increasing finite populations, such that, as the overall number of clusters in the sample increases $(n \rightarrow \infty)$, no survey weight is disproportionate (basically, the sizes of the clusters are bounded); the sequence of finite population Low Income Measures are bounded; the finite population distribution functions converge to a distribution function which is differentiable; the first derivatives of the limit distribution function are bounded from below and from above.

Then, under the assumption of no change in population, and under the model for non-response in the second wave, we have the following results:

i. Both $\hat{\theta}_{Mixed}$ and $\hat{\theta}_{Long}$ are "consistent" estimators of $\hat{\theta}_0$;

ii. the longitudinal estimator of the two-wave Low Income Proportion is approximated as $n \rightarrow \infty$, by:

$$\hat{\theta}_{Long} - \theta_0 \sim \left\{\bar{F}(l_0,l_1) - F(l_0,l_1)\right\} - a_x\left\{\bar{F}(\,2l_0,\infty) - ½\right\} - a_y\left\{\bar{F}(\infty,\,2l_1) - ½\right\} \tag{6}$$

where

$$a_x = F_x^1\,(\,l_0,\,l_1\,)\;/\;2F_x^1(\,2l_0,\,\infty),$$

and

$$a_y = F_y^1(\,l_0,\;l_1)\;/\;2F_y^1\,(\infty,\;2l_1)\ ,$$

where F_x^1, F_y^1 are the first derivates of the limiting distribution function. Thus $\hat{\theta}_{Long}$ is approximated by the sum of three terms: the first term accounts for the variability of the longitudinal proportion $\hat{F}$, and the second and third terms account for the variability originated by the estimation of the Low Income Measures (LIM) l_0 and l_1, respectively.

The three terms in (6) are weighted sums which together can be expressed as the estimator of a single total based on the longitudinal sample:

$$\hat{\theta}_{Long} - \theta_0 \sim \Sigma_{s_L} w_j(L)\,Z_j = \Sigma_{s_0} w_j(L)\,I_j(r)\,Z_j,$$

where the Z_j are finite population values which are functions of income indicators $I(x_j \leq l_0)$, $I(y_j \leq l_1)$, $I(x_j \leq 2l_0)$, and $I(y_j \leq 2l_1)$. Hence a Taylor-linearization variance estimator can be obtained in the form $\hat{V}(\hat{\theta}_{long}) = \hat{V}\left[\Sigma_{s_L} w_j(L)\, Z_j\right]$; note that the variance is taken with respect to the model for non-response and the design.

iii. the mixed estimator can be written as:

$$\hat{\theta}_{Mixed} - \theta_0 = (\hat{\theta}_{Long} - \theta_0) + (\hat{\theta}_{Mixed} - \hat{\theta}_{Long}) \tag{7}$$

where

$$\hat{\theta}_{Mixed} - \hat{\theta}_{Long} \sim \left[\, F_x^1(l_0,l_1) \;/\; 2F_x^1(2l_0,\infty)\right] \cdot \left[\tilde{F}(2l_0,\infty) - \hat{F}(\,2l_0,\infty)\right]. \tag{8}$$

Expression (8), the second term in (7), accounts for the difference in estimating the LIM at time 0, based on the initial sample s_0 before attrition occurred, and estimating the LIM at time 1, based on the longitudinal sample, already affected by non-response. The difference is, as $n \rightarrow \infty$, a linear combination of indicators of sample selection and response, and it can be approximated by

$$\hat{\theta}_{Mixed} - \hat{\theta}_{Long} \sim \Sigma_{s_0}(w_j(L)\, I_j(r) \;-\; 1)\, Z_j^*,$$

where the Z_j^* are finite population values. Hence a Taylor-linearization variance estimator can be obtained in the form

$$\hat{V}(\hat{\theta}_{Mixed}) = \hat{V}\left[\Sigma_{s_0} w_j(L)\, I_j(r) + \Sigma_{s_0}(w_j(L)\, I_j(r) \;-\; 1)\, Z_j^*\right],$$

where the variance is calculated with respect to the model for non-response and the design.

iv. If q_r is the minimum non-response rate among all the response classes, then the variance of the difference between the mixed and the longitudinal estimators increases with q_r and

$$Var(\hat{\theta}_{Mixed} - \hat{\theta}_{Long}) \;/\; V_p \geq q_r \;, \tag{9}$$

where V_p is the variance under a complete response model.

v. Under the assumption of change in the population, where the sample at time 1 contains a new panel of births, and complete response, $\hat{\theta}_0$ is a design-consistent estimator of θ_0 and is approximated as $n \rightarrow \infty$, by:

$$\begin{aligned} \hat{\theta}_0 - \theta_0 \sim & \left\{\hat{F}(l_0,l_1) - F(l_0,l_1)\right\} - a_x\left\{\hat{F}(2l_0,\infty) - 1/2\right\} - a_y\left\{\hat{F}(\infty,2l_1) \;-1/2\right\} \\ & - a_x^1(\hat{F}_D(2l_0) - 1/2) - a_y^1(\,\hat{F}_B\,(\,2l_1\,) - 1/2\,), \end{aligned} \tag{10}$$

where

$$a_x = c_0\, F_x^1(l_0,l_1) \;/\; 2\left[c_0\, F_x^1(2l_0,\infty) + c_D\, F_D^1(2l_0)\right],$$

$$a_x^1 = c_D\, F_x^1(l_0,l_1) \;/\; 2\left[c_0\, F_x^1(2l_0,\infty) + c_D\, F_D^1(2l_0)\right].$$

$$a_y = c_1 F_y^1(l_0,l_1) \;/\; 2 \left[c_1 F_y^1(\infty,2l_1) + c_B F_B^1(2l_1)\right] ,$$

and

$$a_y^1 = c_B F_y^1(l_0,l_1) \;/\; 2 \left[c_1 F_y^1(\infty,2l_1) + c_B F_B^1(2l_1)\right].$$

$\hat{F}_D(x)$ and $\hat{F}_B(y)$ are the corresponding domain estimators of $F_D(x)$ and $F_B(y)$ respectively. Thus $\hat{\theta}_0$is approximated by the sum of five terms: the first three terms account for the variability originated in the estimation of F, and the LIM's at times 0 and 1 respectively, and the other terms account for the variability of the sample of individuals who "died", s_D, and the sample of individuals who were "born", s_B.

The first four terms are linear combinations of sample indicators of the first panel and the fifth term carry sample indicators of the new panel introduced at time 1, which was selected independently of the first. Hence, the right side of (10) can be expressed as the sum of two independent estimators of single totals and a Taylor-linearization variance estimator can be obtained in the same way as in ii. and iii. above.

The proof of the consistency and the approximations follows similar arguments of Serfling (p.75, 1980) for the quantile of the "superpopulation" and Shao and Rao (1993) for the quantile of a finite population. Our estimators differ from Shao and Rao's in that they contain response indicators and thus we have to work in a product space where the model and the sampling design "live" together (Bleuer (1998)). The complete version of the proof will appear in a later paper.

REFERENCES

Binder, D. and Kovacevic, M.S (1995). Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equation Approach. *Survey Methodology*, 21, 151-159

Bleuer, S. and Kovacevic, M. (1998). Some Issues in the Estimation of Income Dynamics, Statistics Canada *Methodology Branch Working Paper Series*, No. SSMD-98-006E, Statistics Canada.

Kovacevic, M.S. and Yung, W. (1997). Estimating the Sampling Variances of Measures of Income Inequality and Polarization - An Empirical Study. *Survey Methodology*, 23, 41-52.

Lavallee, P. and Hunter, L. (1992). Weighting for the Survey of Labour and Income Dynamics. In *Proceedings from Design and Analysis of Longitudinal Surveys*, Symposium 92, Statistics Canada.

Bleuer, S. R. (1998). Research Sabbatical Report, Part I. Inference for parameters of the superpopulation, *Statistics Canada Series no.*, Statistics Canada.

Shao, J. and Rao, J.N.K. (1993). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhya*, B, 55, 393-414.

SESSION 6

MODELLING I

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

MARGINAL MODELS FOR REPEATED OBSERVATIONS: INFERENCE WITH SURVEY DATA

J.N.K. Rao[1]

ABSTRACT

In longitudinal surveys, sample subjects are observed over several time points. This feature typically leads to dependent observations on the same subject, in addition to the customary correlations across subjects induced by the sample design. Much research in the literature has focussed on modeling the marginal mean of a response as a function of covariates. Liang and Zeger (1986) used generalized estimating equations (GEE), requiring only correct specification of the marginal mean, and obtained standard errors of regression parameter estimates and associated Wald tests, assuming a "working" correlation structure for the repeated measurements on a sample subject. Rotnitzky and Jewell (1990) developed quasi-score tests and Rao-Scott adjustments to "working" quasi-score tests under marginal models. These methods are asymptotically robust to misspecification of the within-subject correlation structure, but assume independence of sample subjects which is not satisfied for complex longitudinal survey data based on stratified multi-stage sampling. We propose asymptotically valid Wald and quasi-score tests for longitudinal survey data, using the Taylor Linearization and jackknife methods. Alternative tests, based on Rao-Scott adjustments to naive tests that ignore survey design features and on Bonferroni-t, are also developed. These tests are particularly useful when the effective degrees of freedom, usually taken as the total number of sample primary units (clusters) minus the number of strata, is small.

1. INTRODUCTION

In a longitudinal survey sample subjects are observed over two or more time points. Such surveys are suited to study individual changes over time, unlike cross-sectional surveys. Applications of longitudinal surveys include: (a) Gross flows: estimation of transition counts between a finite number of states for individuals in a population from one point in time to the next. Such flow estimates are important to researchers and policy analysts for understanding labor market dynamics. (b) Event history modelling; for example unemployment spells. (c) Elimination of the effect of latent variables in regression models using individual changes in the response and explanatory variables between two consecutive time points. (d) Modelling the marginal means of responses as functions of covariates. (e) Conditional modelling of the response at a given time point as a function of past responses and present and past covariates. Such models can provide better understanding of the underlying dynamics than the marginal models (d). Binder (1998) gave an excellent account of the issues related to longitudinal surveys.

Longitudinal surveys typically lead to dependent observations on the same subject, in addition to the customary cross-sectional correlations induced by the clustering in the sample design. In this paper we focus on marginal modelling and analysis of such longitudinal survey data. The case of a simple random sample of individuals has been studied extensively in the literature, especially in the analysis of data occuring in biomedical and health sciences. Liang and Zeger (1986) used generalized estimating equations, requiring only correct specification of the marginal mean. They obtained standard errors of regression parameter estimates and

[1] J.N.K. Rao, Carleton University, Ottawa, Canada.

associated "Wald" tests, assuming a "working" correlation structure for the repeated measurements on a sample subject. Rotnitzky and Jewell (1990) developed "quasi-score" tests and "Rao-Scott" adjustments to working quasi-score tests, under marginal models. These methods are asymptotically valid regardless of the true within-subject correlation structure, but assume independence of sample subjects which is not satisfied for complex longitudinal survey data based on stratified cluster samples.

In this paper, Wald and quasi-score tests for longitudinal survey data are proposed, using the Taylor linearization and jackknife methods. These methods take account of the survey design features (clustering, stratification, unequal sampling weights etc.) as well as the longitudinal feature and thus are asymptotically valid.

2. INDEPENDENCE ESTIMATING EQUATIONS

Suppose the survey population U consists of M individuals and a sample, s, of individuals is selected using stratified multistage sampling. Let denote the basic design weight attached to the k-th sample individual in the i-th sample cluster ($i = 1, \ldots, n_h$) from the h-th stratum ($h = 1, \ldots, L$). In a longitudinal survey, the sample s is observed over a specified number of time points, say T, but in practice some of the sample individuals may not respond. The sample weights of respondents, , on the first occasion are first adjusted for unit nonresponse, and then subjected to post-stratification adjustment to ensure consistency with known benchmark totals, e.g., age-sex counts obtained from external sources. We denote the final weights as w^{*}_{hik}, often called longitudinal weights.

Suppose that the i-th respondent is observed for T_i occasions ($1 \le T_i \le T$ and $i \in s_r$); here "i" refers to some "hik". We assume that the responses are missing completely at random (MCAR), i.e., the response probabilities for an individual do not depend on the missing responses and the observed responses, following Liang and Zeger (1986). The data for the i-th sample individual ($i \in s_r$) consists of $\{(y_{it}, \boldsymbol{x}_{it}), t = 1, \ldots, T_i\}$ where y_{it} is response on occasion t and $\boldsymbol{x}_{it}$ is a $p \times 1$vector of associated covariates. In the case of binary response, $y_{it} = 1$ if subject i has the attribute at time t, and 0 otherwise.

The marginal model assumes that the mean response $\mu_{it} = E_m(y_{it})$ is a specified function of $\boldsymbol{x}_{it}$ and regression parameters $\boldsymbol{\beta}$; in particular $g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$ where $g(\cdot)$ is called the link function. With binary responses, the logit link function $g(\mu) = \log\{\mu/(1-\mu)\}$ is a natural choice, leading to a logistic regression model. In this section, we assume "working" independence so that

$$\operatorname{cov}(y_i) = V_{0i} = \operatorname*{diag}_{1 \le t \le T_i} (V_{0it}),$$

where $V_{0it} = V_0(\mu_{it}) = \operatorname{var}(y_{it})$ is the working variance. For example, in the binary response case, $V_0(\mu_{it}) = \mu_{it}(1 - \mu_{it})$.

The above formulation permits time varying regression coefficients. For example if $T = 2$ and $g(\mu_{it}) = \alpha_t + \beta_t z_{it}, t = 1,2,$ then we can define $\boldsymbol{x}_{i1} = (1\, z_{i1}\, 0\, 0)^T$, $\boldsymbol{x}_{i2} = (0\, 0\, 1\, z_{i2})^T$ and $\boldsymbol{\beta} = (\alpha_1\, \beta_1\, \alpha_2\, \beta_2)^T$. In this case, it might be of interest to test the constancy of the slope coefficient over time, i.e., $H_0: \beta_1 = \beta_2$ which is of the form $H_2: \boldsymbol{c}^T \boldsymbol{\beta} = 0$ with $\boldsymbol{c} = (0\, 1\, 0\, -1)^T$.

We assume that the marginal model holds for the whole population of M subjects so that we get the "census" model

$$g(\mu_{it}) = \boldsymbol{x}'_{it} \boldsymbol{\beta}, \quad t = 1, \ldots T_i; i = 1, \ldots, M \tag{2.1}$$

where T_i now refers to the number of consecutive occasions the i-th population subject would respond if contacted. We further assume that the population of M subjects is a self-weighting sample from a superpopulation obeying the marginal model. It is not necessary to regard the population as a random sample from the superpopulation. The census generalized estimating equations (GEE) are then given by

$$S_\ell(\beta) = \sum_{i=1}^{M} u_{i\ell}(\beta) = 0, \quad \ell = 1,\ldots,p \tag{2.2}$$

where

$$u_{i\ell}(\beta) = \sum_{t=1}^{T_i} \frac{\partial \mu_{it}}{\partial \beta_\ell} \frac{\left(y_{it} - \mu_{it}\right)}{V_{0it}} \tag{2.3}$$

Under general conditions, the solution of (2.2), β_M, is a consistent estimator of β. We denote β_M as the census regression parameter and make statistical inferences on β_M, following Binder (1983). Such inference are also valid for β under certain conditions. For simplicity, we do not distinguish between β_M and β in this paper.

Noting that the left hand side of (2.2) is the population total of $u_{i\ell}(\beta)$, a design-consistent estimator of $S(\beta) = (S_1(\beta),\ldots,S_p(\beta))^T$, called the sample GEE, is given by

$$\hat{S}(\beta) = \sum_{hik\in s_r} w^{*}_{hik}\, \boldsymbol{u}_{hik}(\beta) = \boldsymbol{0} \tag{2.4}$$

where $\boldsymbol{u}_{hik}(\beta) = [u_{hikl}(\beta),\ldots,u_{hikp}(\beta)]^T$ with $u_{hik\ell}(\beta)$ obtained from (2.3) by changing "i" to "*hik*". The solution of (2.4), $\hat\beta$, is a design-consistent estimator of β_M.

3. WALD TESTS UNDER WORKING INDEPENDENCE

It is a common practice among social scientists and others to use normalized weights $\tilde{w}_{hik} = m w^{*}_{hik} / \Sigma s_r\, w^{*}_{hik}$, where m is the size of s_r, and then apply standard methods, using SAS or some other standard software package. Using the normalized weights in the standard "sandwich" covariance estimator of $\hat\beta$, we get the following naive covariance estimator:

$$v_N\left(\hat\beta\right) = \left[\tilde{\boldsymbol{I}}(\hat\beta)\right]^{-1} \left(\sum_{s_r} \tilde{w}_{hik}\, \boldsymbol{u}_{hik}(\hat\beta)\, \boldsymbol{u}_{hik}(\hat\beta)^T\right) \left[\tilde{\boldsymbol{I}}(\hat\beta)\right]^{-1} \tag{3.1}$$

where $\tilde{\boldsymbol{I}}(\hat\beta)$ is the estimated information matrix with

$$\tilde{\boldsymbol{I}}(\beta) = -\sum_{s_r} \tilde{w}_{hik}\, E_m\left[\partial \boldsymbol{u}_{hik}(\beta) / \partial \beta^T\right] \tag{3.2}$$

This follows by applying the Liang-Zeger sandwich covariance estimator formula to the census parameter:

$$v\left(\beta_M\right) = \left[\boldsymbol{I}\left(\beta_M\right)\right]^{-1} \left[\sum_{hik\in U} \boldsymbol{u}_{hik}\left(\beta_M\right) \boldsymbol{u}_{hik}\left(\beta_M\right)^T\right] \left[\boldsymbol{I}\left(\beta_M\right)\right]^{-1} \tag{3.3}$$

and then replacing each term in (3.3) by its estimator based on the normalized weights, $\tilde{w}_{hik}$, where $\boldsymbol{I}(\beta_M)$ is the census information matrix with

$$I(\beta) = - \sum_{hik \in U} E_m \left[\partial \boldsymbol{u}_{hik}(\beta) / \partial \beta^T \right]. \tag{3.4}$$

In the case of a simple random sample and no post-stratification adjustment, we have $\tilde{w}_i = 1$ for all the sample subjects i and (3.1) reduces to the Liang-Zeger formula.

Suppose we are interested in testing a hypothesis of the form H_0: $\beta_2 = \beta_{20}$, using the sample data $\{(y_{hikt}, x_{hikt}); hik \in s_r, t = 1,\dots,T_{hik}\}$, where β is partitioned as $\beta = (\beta_1^T, \beta_2^T)^T$ with β_2 a $r \times 1$ vector and β_1 a $q \times 1$ vector $(q + r = p)$. For example, β_2 could represent interaction terms and we are interested in testing for the absence of interactions, i.e., $\beta_{20} = \mathbf{0}$. A "naive" Wald test of H_0 treats

$$W_N = \left(\hat{\beta}_2 - \beta_{20}\right)^T \left[v_{N22}(\hat{\beta})\right]^{-1} \left(\hat{\beta}_2 - \beta_{20}\right) \tag{3.5}$$

as a χ^2 variable with r degrees of freedom (d.f.), where $v_{N22}(\hat{\beta})$ is the submatrix of $v_N(\hat{\beta})$ corresponding to β_2. This test, however, is asymptotically incorrect under stratified multistage sampling or any other complex sampling design. In fact, W_N is asymptotically distributed as a weighted sum of independent χ_1^2 variables, where the weights are the eigenvalues of a "design effects" matrix. As a result, the naive test could lead to inflated significance levels relative to the nominal level, say 0.05.

We assume that the sampling design provides consistent, asymptotically normal estimators of totals. Following Binder (1983), under certain regularity condition, $\hat{\beta}$ is then asymptotically normal with mean β_M and its covariance matrix, cov $(\hat{\beta})$, can be consistently estimated by

$$\upsilon_L(\hat{\beta}) = \left[\hat{J}(\hat{\beta})\right]^{-1} v(\hat{S}) \left[\hat{J}(\hat{\beta})\right]^{-1} \tag{3.6}$$

Here

$$\hat{J}(\beta) = - \partial \hat{S}(\beta) / \partial \beta^T = - \sum_{hik \in s_r} w_{hik}^{\bullet} \, \partial \boldsymbol{u}_{hik} / \partial \beta^T \tag{3.7}$$

and $v(\hat{S})$ is the estimated covariance matrix of $\hat{S}(\beta)$ under the specified sampling design evaluated at $\beta = \hat{\beta}$. Note that $v(\hat{S})$ is obtained from a standard survey variance estimator, noting that $\hat{S}(\beta)$ is the vector of estimated totals of $u_{hik\ell}(\beta)$, $\ell = 1,\dots,p$. However, the variance estimator used should account for post-stratification and nonresponse adjustment. For example, if the post-stratification indicator variables are denoted by z_{hik}, $hik \in s$, and nonresponse is absent, then $v(\hat{S})$ is the estimated covariance matrix of $\hat{E}(\beta) = \sum_{hik \in s} w_{hik} \, e_{hik}(\beta)$, where $e_{hik\ell}(\beta) = u_{hik\ell}(\beta) - z_{hik}^T \hat{B}_\ell$ with

$$\hat{B}_\ell = \left(\sum_s w_{hik} \, z_{hik} \, z_{hik}^T \right)^{-1} \left(\sum_s w_{hik} \, z_{hik} \, u_{hik\ell}(\beta) \right), \ \ell = 1,\dots,p$$

Letting $\boldsymbol{e}_{hi}^{\bullet} = n_h \sum_k w_{hik} \, \boldsymbol{e}_{hik}(\hat{\beta})$, we have

$$v(\hat{S}) = \sum_h \frac{1}{n_h(n_h - 1)} \sum_i \left(\boldsymbol{e}_{hi}^{\bullet} - \boldsymbol{e}_{h.}^{\bullet}\right)\left(\boldsymbol{e}_{hi}^{\bullet} - \boldsymbol{e}_{h.}^{\bullet}\right)^T \tag{3.8}$$

where $\boldsymbol{e}_{h.}^{\bullet} = \sum \boldsymbol{e}_{hi}^{\bullet} / n_h$. The formula (3.8) assumes that the first stage clusters are either drawn with replacement in each stratum or the first stage sampling fractions are negligible.

In the case of nonresponse with weighting classes cutting across post-strata, the formula for $v(\hat{S})$ becomes more complicated (see Yung, 1996, Chapter 4).

If post-stratification is not employed, then $w^{*}_{hik} = w_{hik}$ and we replace $e_{hik}(\hat{\beta})$ by $u_{hik}(\hat{\beta})$ in (3.8) to get $v(\hat{S})$.

An alternative version of $v_L(\hat{\beta})$ is obtained by changing $\hat{J}(\hat{\beta})$ to $\hat{I}(\hat{\beta})$, where $\hat{I}(\beta) = E_m \hat{J}(\beta)$ in (3.6). We suspect that $\hat{I}(\hat{\beta})$ is more stable than $\hat{J}(\hat{\beta})$. Note that $\hat{J}(\beta) = E_m \hat{J}(\beta)$ for logistic regression with binary response.

It may be noted that post-stratification may not lead to increased efficiency because the model residuals $u_{hikl}(\hat{\beta})$ may be unrelated to the post-stratifiers z_{hik}, particularly when the model fits the data well.

The jackknife method can be used in a straightforward manner to estimate the covariance matrix of $\hat{\beta}$. An advantage of the jackknife is that post-stratification and unit-nonresponse adjustment are automatically taken into account, unlike the linearization method.

Using the estimated cov $(\hat{\beta})$, Wald test of hypothesis of the form $H_0: \psi = C\beta = 0$ are readily obtained, where **C** is a $r \times p$ full rank matrix of known constants and β is $p \times 1$ vector ($r <$ p). Under H_0,

$$X_W^2 = \hat{\psi}^T \left(C v_L(\hat{\beta}) C^T\right)^{-1} \hat{\psi} \tag{3.9}$$

is distributed asymptotically as χ_r^2, a χ^2 variable with r d.f., where $\hat{\psi} = C\hat{\beta}$. Therefore, the p-value associated with H_0 is computed as $P[\chi_r^2 > X_W^2(\text{obs})]$, where $X_W^2(\text{obs})$ is the observed value of the Wald statistics X_W^2. More general hypotheses of the form $H_0: \psi = h(\beta) = 0$ can also be tested using the Wald method, where $h(\beta)$ is a $r \times 1$ vector. Under H_0, we have

$$X_W^2 = \hat{\psi}^T \hat{\Sigma}_{\hat{\psi}}^{-1} \hat{\psi} \tag{3.10}$$

is asymptotically χ_r^2, where $\hat{\psi} = h(\hat{\beta})$ and $\hat{\Sigma}_{\hat{\psi}} = H(\hat{\beta})\, v_L(\hat{\beta})\, H(\hat{\beta})^T$ with $H(\beta) = \partial h(\beta) / \partial \beta^T$, a $r \times p$ full rank matrix.

4. QUASI-SCORE TESTS UNDER WORKING INDEPENDENCE

For the Wald tests, we have to fit the full model $g(\mu_{it}) = x_{it}^T \beta$ which could lead to unstable estimates if the full model contains a large number of terms. For example, with a factorial structure of explanatory variables containing a large number of interactions we may be interested in testing the significance of interaction effects, denoted as $H_0: \beta_2 = \beta_{20} = 0$. On the other hand, for the quasi-score tests we need only to fit the simple null model, $g(\mu_{it}) = x_{1it}^T \beta_1$, where $x_{it} = (x_{1it}^T, x_{2it}^T)^T$. Moreover, the quasi-score tests are invariant to nonlinear transformations of β, unlike the Wald tests (Boos, 1992). Rao and Scott (1996) studied quasi-score tests in the context of cross-sectional survey data.

Let $\tilde{\beta} = (\tilde{\beta}_1^T, \beta_{20}^T)^T$ be the solution of $\hat{S}_1(\beta_1^T, \beta_{20}^T) = 0$, where $\hat{S} = (\hat{S}_1^T, \hat{S}_2^T)^T$ is partitioned in the same way as β. The analogue of the score test, called quasi-score test, is given by

$$X_S^2 = \tilde{S}_2^T\, v\left(\tilde{S}_2\right)^{-1} \tilde{S}_2 \tag{4.1}$$

where $\tilde{S}_2 = \hat{S}_2(\tilde{\beta})$ and $v(\tilde{S}_2)$ is a design-consistent estimator of cov $(\tilde{S}_2)$. We now sketch a proof to show that X_S^2 is asymptotically χ_r^2 under H_0.

Expanding $\hat{S}_1(\tilde{\beta})$ and $\hat{S}_2(\tilde{\beta})$ around the true value $\beta^* = (\beta_1^T, \beta_{20}^T)^T$ gives

$$\mathbf{0} = \hat{S}_1\left(\tilde{\beta}\right) \approx \hat{S}_1\left(\beta^*\right) - J_{11}^*\left(\tilde{\beta}_1 - \beta_1^*\right) \tag{4.2}$$

and

$$\hat{S}_2 \approx \hat{S}_2\left(\beta^*\right) - J_{21}^*\left(\tilde{\beta}_1 - \beta_1^*\right) \tag{4.3}$$

where $J^* = \hat{J}(\beta^*)$ is the value of $\hat{J}(\beta) = -\partial\hat{S}(\beta)/\partial\beta^T$ at $\beta = \beta^*$ and J^* is partitioned as

$$J^* = \begin{bmatrix} J_{11}^* & J_{12}^* \\ J_{21}^* & J_{22}^* \end{bmatrix}$$

Now replacing J^* by its expected value I^* and substituting for $\tilde{\beta}_1 - \beta_1^*$ from (4.2) into (4.3), we get

$$\begin{aligned}\tilde{S}_2 = \hat{S}_2\left(\tilde{\beta}\right) &\approx \hat{S}_2\left(\beta^*\right) - I_{21}^* I_{11}^{*-1} \hat{S}_1\left(\beta^*\right) \\ &= \sum_{hik \in s_r} w_{hik}^* \tilde{u}_{2hik}\left(\beta^*\right),\end{aligned} \tag{4.4}$$

where $\tilde{u}_{2hik}(\beta^*) = u_{2hik}(\beta^*) - A^* u_{1hik}(\beta^*)$ with $A^* = I_{21}^* I_{11}^{*-1}$ and $u_{hik} = (u_{1hik}^T, u_{2hik}^T)^T$. It follows from (4.4) that $\tilde{S}_2$ is approximately equal to a vector of estimated totals so that $\tilde{S}_2$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\text{cov}(\tilde{S}_2)$. Thus X_S^2 is asymptotically χ_r^2 under H_0. Note that $E(\tilde{S}_2) \approx \mathbf{0}$ under H_0.

Calculation of the quasi-score test X_S^2 requires an estimator of $\text{cov}(\tilde{S}_2)$. A jackknife estimator, $v_J(\tilde{S}_2)$ is obtained in a straight foward manner. The jackknife final weights, $w_{hik(gj)}^*$, when the (gj)-th sample cluster is deleted are obtained in the same manner as w_{hik}^*, using the jackknife basic weights $w_{hik(gj)} = w_{hik} b_{gj}$ where $b_{gj} = 0$ if $(hi) = (gj)$; $n_g/(n_g - 1)$ if $h = g$ and $i \neq j$; $=1$ if $h \neq g$. Replacing w_{hik}^* by $w_{hik(gj)}$, we get $\hat{S}_{(gj)}(\beta)$, $\tilde{\beta}_{(gj)}$ and $\tilde{S}_{2(gj)} = \hat{S}_{2(gj)}(\tilde{\beta}_{(gj)})$. Using $\tilde{S}_{2(gj)}$ we get

$$v_J\left(\tilde{S}_2\right) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \left(\tilde{S}_{2(gj)} - \tilde{S}_2\right)\left(\tilde{S}_{2(gj)} - \tilde{S}_2\right)^T \tag{4.5}$$

Computation of $\tilde{\beta}_{(gj)} = (\tilde{\beta}_{1(gj)}^T, \beta_{20}^T)^T$ can be simplified by performing only a single Newton-Raphson iteration

for the solution of $\hat{S}_{1(gj)}(\beta_1^T, \beta_{20}^T) = \mathbf{0}$, using $\tilde{\beta}$ as the starting value. The jackknife quasi-score test (4.1) is invariant to a one-to-one reparametrization of β with non-singular Jacobian, unlike the Wald test X_W^2.

A Taylor linearization estimator of $\mathbf{cov}(\tilde{S}_2)$, denoted as $v_L(\tilde{S}_2)$, can also be obtained using the asymptotic representation (4.4) of $\tilde{S}_2$ as a vector of estimated totals. We replace $\tilde{u}_{2hik}(\beta^*)$ by $\tilde{u}_{2hik}(\tilde{\beta}) = u_{2hik}(\tilde{\beta}) - \tilde{A} u_{1hik}(\tilde{\beta})$, where $\tilde{A}$ is an estimator of A^*, and then use (3.8) with $u_{hik}(\hat{\beta})$ changed to $\tilde{u}_{2hik}(\tilde{\beta})$. There are several possible choices for $\tilde{A}$. It might seem natural to use $\hat{J}(\tilde{\beta})$ in place of I^*, where $\hat{J}(\beta)$ is given by (3.7). For the special case of scalar β_2 (i.e., $r = 1$) and one time point (i.e., $T_i = 1$), Binder and Patak (1994) used this form of quasi-score test to construct confidence intervals for β_2, although their approach is different from that given here. This choice, however, does not have the desired invariance property in general. We can get an invariant quasi-score test by taking the expectation of $\hat{J}(\beta)$ under the mean specification defined by (2.1), i.e. by using

$$\hat{I}(\tilde{\beta}) = \sum_{hik \in s_r} \sum_{t} w^{*}_{hik} D_{hikt}(\tilde{\beta}) \; D_{hikt}(\tilde{\beta})^{T} \, / \, V_0(\tilde{\mu}_{hikt}) \tag{4.6}$$

where $D_{hikt}(\beta) = \partial \mu_{hikt} / \partial \beta$ with $\tilde{\mu}_{hikt}$ denoting the value of the mean μ_{hikt} at $\beta = \tilde{\beta}$. Moreover, the resulting test is likely to be more stable. Of course, for the binary response logistic regression case $\hat{I}(\beta) = \hat{J}(\beta)$.

Under the stratified multistage sampling set-up it can be shown that $v_J(\tilde{S}_2) \approx v_L(\tilde{S}_2)$, so that the jackknife and Taylor linearization quasi-score tests are asymptotically equivalent.

More general hypotheses of the form $H_0 : \psi = h(\beta) = 0$ can also be tested using the quasi-score method. The estimate $\tilde{\beta}$ under H_0 is obtained by solving

$$\hat{S}(\beta) - H(\beta)^{T} \lambda = 0; \quad h(\beta) = 0 \tag{4.7}$$

for β and λ, where λ is the $r \times 1$ vector of Lagrange multipliers. Let $\tilde{S}_h = H(\tilde{\beta}) \, [\hat{I}(\tilde{\beta})]^{-1} \, \hat{S}(\tilde{\beta})$, then the jackknife quasi-score is given by

$$X_S^2 = \tilde{S}_h^{T} [\, v_J(\tilde{S}_h) \,]^{-1} \, \tilde{S}_h , \tag{4.8}$$

where $v_J(\tilde{S}_h)$ is the jackknife estimator of cov $(\tilde{S}_h)$ which is obtained in a straightforward manner from (4.5) by changing $\tilde{S}_2$ to $\tilde{S}_h$ and $\tilde{S}_{2(gj)}$ to $\tilde{S}_{h(gj)}$.

A Taylor linearization estimator of cov $(\tilde{S}_h)$, denoted as $v_L(\tilde{S}_h)$, can also be obtained using the following asymptotic representation of $\tilde{S}_h$:

$$\begin{aligned} \tilde{S}_h &\approx H^{*} I^{*-1} \, \hat{S}(\beta^{*}) - H^{*}(\tilde{\beta} - \beta) \\ &\approx H^{*} I^{*-1} \, \hat{S}(\beta^{*}), \end{aligned} \tag{4.9}$$

noting that $0 = h(\tilde{\beta}) \approx h(\beta^{*}) + H^{*}(\tilde{\beta} - \beta)$ so that $H^{*}(\tilde{\beta} - \beta^{*}) \approx 0$ under H_0, where $H^{*} = H(\beta^{*})$. Now letting $\tilde{u}_{hik}(\tilde{\beta}) = H(\tilde{\beta}) \hat{I}(\tilde{\beta})^{-1} \, u_{hik}(\tilde{\beta})$, it follows from (4.9) that $v_L(\tilde{S}_h)$ is given by (3.8) with $u_{hik}(\hat{\beta})$ changed to $\tilde{u}_{hik}(\tilde{\beta})$, assuming complete response.

5. WORKING CORRELATION STRUCTURE

In this section we generalize the previous results on quasi-score tests to the case of a "working" correlation matrix of y_i, assuming $T_i = T$. The working covariance matrix of $y_{hik} = (y_{hik1}, \ldots, y_{hikT})^{T}$ is assumed to be the form $V_{0hik} = A_{hik}^{1/2} R A_{hik}^{1/2}$ with common correlation structure across units (*hik*), i.e., $R_{hik} = R$, where $A_{hik} = \text{diag}(V_{0hik1}, \ldots, V_{0hikT})$ and $V_{0hikt} = \text{var}(y_{hikt})$.

We use $\tilde{\beta}$, obtained under working independence and H_0, to get an estimator of **R**:

$$\hat{R}(\tilde{\beta}) = \sum_{s_r} w^{*}_{hik} \tilde{R}_{hik} / \sum_{s_r} w^{*}_{hik} , \tag{5.1}$$

where $\tilde{R}_{hik} = R_{hik}(\tilde{\beta})$ with

$$R_{hik}(\beta) = A_{hik}^{-1/2} (y_{hik} - \mu_{hik}(\beta)) (y_{hik} - \mu_{hik}(\beta))^{T} A_{hik}^{-1/2} . \tag{5.2}$$

Note that $\hat{R}(\tilde{\beta})$ is a design consistent estimator of the census parameter $R_M = \Sigma R_{hik}(\beta)/M$.

Now using $\hat{R}(\tilde{\beta})$, we get

$$u^*_{hik}(\beta) = (\partial \mu^T_{hik} / \partial \beta)\, \tilde{V}^{-1}_{0hik}\, (y_{hik} - \mu_{hik}), \tag{5.3}$$

where $\tilde{V}_{0hik} = \tilde{A}^{1/2}_{hik}\, \hat{R}(\tilde{\beta})\, \tilde{A}^{1/2}_{hik}$. The results of Section 3 under working independence can be extended by changing $u_{hik}(\beta)$ to $u^*_{hik}(\beta)$ given by (5.3). The information matrix (4.6) now changes to

$$\hat{I}(\tilde{\beta}) = \sum_{hik \in s_r} w^*_{hik}\, D_{hik}(\tilde{\beta})\, \tilde{V}^{-1}_{0hik}\, D_{hik}(\tilde{\beta})^T \tag{5.4}$$

where $D_{hik}(\beta) = \partial \mu^T_{hik} / \partial \beta$. Properties of the resulting score tests are under investigation.

Liang and Zeger (1986) consider the case of general T_i, assuming working exchangeable correlation structure, moving average process (MA-1) or autoregression process (AR-1). However, this approach can lead to inefficient estimators of β under misspecification of the correlation structure, as demonstrated by Sutradhar and Das (1999).

6. CONCLUDING REMARKS

The Wald and quasi-score tests become unstable if the effective degrees of freedom is small. In the context of stratified multistage sampling, effective degrees of freedom, f, is usually taken as the total number of sample primary units minus the number of strata. For a subgroup (or domain), f can be much less if the subgroup is not uniformly distributed across the sampled primary units. If f is not large, F-version of the Wald or quasi-score tests might perform better in controlling the size of the test. An F-version of the quasi-score test treats

$$F_S = [(f - r + 1)/(fr)]\, X^2_S \tag{6.1}$$

as an F-variable with r and $f - r + 1$ degrees of freedom.

Alternative Rao-Scott (1984) corrected tests or Bonferroni-t tests (Korn and Graubard, 1990) might perform better than X^2_S or F_S when f is small. Rotnitzky and Jewell (1990) proposed Rao-Scott corrected score tests in the case of a simple random sample of subjects. We plan to study the properties of these alternative tests.

REFERENCES

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279-292.

Binder , D.A. (1998). Longitudinal Surveys: Why are these surveys different from all other surveys? *Survey Methodology*, **24**, 101-108.

Binder, D.A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *J. Amer. Statist. Assoc.*, **89**, 1035-1043.

Boos, D.D. (1992). On generalized score tests. *Amer. Statist.*, **46**, 327-333.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t statistics. *Amer. Statist.*, **44**, 270-276.

Liang, K.Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contigency tables with cell proportions estimated from survey data. *Ann. Statist.*, **12**, 46-60.

Rao, J.N.K. and Scott, A.J. (1996). Quasi-score tests with survey data. *Proc. Survey Sec.*, Statistical Society of Canada Annual Meetings, 1996, 33-38.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing in semi-parametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485-497.

Sutradhar, B.C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86**, (in press).

Yung, W. (1996). Contributions to poststratification in stratified multi-stage samples. Unpublished Ph.D. thesis, Carleton University, Ottawa, Canada.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

ESTIMATING THE INCIDENCE OF DEMENTIA FROM LONGITUDINAL TWO-PHASE SAMPLING WITH NONIGNORABLE MISSING DATA

Sujuan Gao[1], Siu L. Hui[1,3], Kathleen S. Hall[2], Hugh C. Hendrie[2]

ABSTRACT

Two-phase sampling designs have been conducted in waves to estimate the incidence of a rare disease such as dementia. Estimation of disease incidence from longitudinal dementia study has to appropriately adjust for data missing by death as well as the sampling design used at each study wave. In this paper we adopt a selection model approach to model the missing data by death and use a likelihood approach to derive incidence estimates. A modified EM algorithm is used to deal with data missing by sampling selection. The non-parametric jackknife variance estimator is used to derive variance estimates for the model parameters and the incidence estimates. The proposed approaches are applied to data from the Indianapolis-Ibadan Dementia Study.

KEY WORDS: nonignorable missing data, selection model, EM algorithm, jackknife.

1. INTRODUCTION

Two-phase sampling designs are often used in studies where a disease is rare and diagnosis of the disease is expensive or difficult. In the first phase of the study a large random sample from the targeted population is screened, in the case of dementia, with questionnaires administered by trained interviewers. Based on the results of the screening, study subjects are stratified and randomly selected within strata to receive extensive clinical evaluations to determine disease status.

In a longitudinal study on dementia and Alzheimer's disease, the two-phase sampling design is first used at the study baseline to estimate disease prevalence, and then repeated several years apart to estimate disease incidence and to identify potential risk factors of the disease. Therefore, the two-phase sampling is conducted in waves to identify new disease case. The estimation of disease incidence requires appropriate adjustment for the complex samplings used at each of the study waves. In a recent article Clayton *et al* (1998) give an excellent review on statistical techniques in longitudinal dementia studies with two-phase sampling designs. They proposed the method of inverse probability weighting, a full maximum likelihood approach, and a mean score imputation approach and compared these methods on simulated data sets.

Most longitudinal dementia studies, as acknowledged by Clayton *et al* (1998), are also complicated by the presence of missing data by "happenstance". For example, some study subjects die before the next study wave, some refuse further participation in the study and some moved out of the study areas. For studies on elderly subjects, death is an inevitable source for missing data. For these subjects disease status prior to death cannot be ascertained and remains missing in the data. Under the assumption of missing completely at random (MCAR) or missing at random for the missing data mechanism, using the terminology of Little and Rubin (1987), valid incidence estimates can still be derived provided appropriate likelihood or Bayesian approaches are taken and all covariates contributing to missing data process are included in the model. However, medical studies following demented subjects in cohort studies have found that demented subjects are more likely to die than non-demented subjects. Hence there are reasons to

[1] Division of Biostatistics, Department of Medicine

[2] Department of Psychiatry

[3] Regenstrief Institute for Health Care, Indiana University

suggest the data missing by death are probably nonignorable, meaning that the missingness may depend on the unobserved disease status. Estimation procedures without adjusting for this type of missing data may lead to underestimation of disease incidence.

Statistical inference with nonignorable missing data has mostly relied on making additional assumptions about the missing data mechanism, in situations where information external to the data cannot be used to determine the missing data process. The approaches have been broadly classified into a selection model approach, as in Diggle and Kenward (1994), and the pattern mixture model approach, as reviewed by Little (1995).

In this paper we attempt to estimate the incidence of dementia from the longitudinal two-phase sampling setting under a nonignorable missing data assumption for the deceased subjects by using the selection model approach of Diggle and Kenward (1994). Diggle and Kenward demonstrated their approach on longitudinal continuous outcomes. Here for the dementia data we consider longitudinal binary disease outcomes with complex sampling designs. We assume a logistic regression model for the mortality outcome and use maximum likelihood approach to derive parameter estimates. The EM algorithm approach is used to deal with data that are missing by design, i.e. by sampling selection. The incidence estimates are then obtained as the predicted means from the disease models. Since it is difficult to derive analytic variance estimators for the incidence estimates with the nonignorable missing data and the complex sampling design, we propose to use the jackknife variance estimator to estimate the variances of the model parameters and the incidence estimates.

2. PROPOSED METHODOLOGY

We consider the situation of just one follow up wave after baseline wave in this paper. Let Y_1 and Y_2 be two binary variables for disease status at study baseline and follow up wave, respectively. Let X_1 and X_2 be two *N x p* matrices of covariates measured at baseline and follow up wave, respectively. *N* is the total number of subjects in the study cohort. Assume that *D* is the dichotomous variable for survival status with $D = 1$ indicating a subject has died between baseline and follow up and $D = 0$ for the survivors. We suppress the use of subscript *i* for the *i*th subject in the study for simplicity in notations.

For the joint modeling of Y_1 and Y_2, it is reasonable to model the baseline dementia probability and the incidence dementia probability because these quantities are usually of interest themselves. Furthermore, the baseline probability of dementia is often modeled using a logistic regression model to achieve robust variance estimates in prevalence estimates because of the small number of cases in some sampling strata (Roberts *et al*, 1987, Beckett *et al*, 1992). Clayton *et al* (1998) have also demonstrated that the modeling approach is more efficient than the inverse probability weighting approach in the longitudinal setting. Therefore, we use the following logistic regression model for the baseline disease probability:

$$\operatorname{Prob}\left(Y_1=1 \mid X_1\right)=\operatorname{logit} X_1 \alpha \tag{1}$$

We now assume a logistic model for the conditional probability of dementia given a subject was dementia free at study baseline:

$$\operatorname{Prob}\left(Y_2=1 \mid Y_1, X_2\right)=\begin{cases}\operatorname{logit} X_2 \beta & \text{if } Y_1=0, \\ 1 & \text{if } \mathrm{Y}_1=1.\end{cases} \tag{2}$$

Note the above conditional model enforces the irreversible disease constraint on the conditional probability. Since the dementia data consists data selected by complex sampling as well as data missing by death, we develop our proposed method in two steps. The first step deals with nonignorable missing data by death, and the second considers that data are obtained by sampling.

2.1 Selection model without sampling

For the time being we assume that every subject had baseline clinical evaluation and all survivors had follow up diagnosis. Let Y_2^* be the latent disease status for the deceased subjects who are missing disease status at the follow up wave.

Following Diggle and Kenward (1994) we assume the probability of dying between the two study waves is nonignorable and follows the model:

$$\Pr\text{ob}\left(D=1\middle|Y_1,Y_2,Z\right)=\log\text{it}\,(Z\theta_0+\theta_1 Y_1+\theta_2 Y_2) \tag{3}$$

The likelihood function for the observed data can be written as:

$$\begin{aligned} L &= \prod_{i=1}^{N} \int f\left(Y_1,Y_2,Y_2^*\middle|X_1,X_2;\alpha,\beta\right) f\left(D\middle|Y_1,Y_2,Y_2^*,Z;\theta\right) dY_2^* \\ &= \prod_{i=1}^{N} f\left(Y_1\middle|X_1\right) \prod_{(D=0)} f\left(Y_2\middle|Y_1,X_2\right)\left(1-p_d\left(Z,Y_1,Y_2\right)\right) \prod_{(D=1)} \int f\left(Y_2^*\middle|Y_1,X_2\right) p_d\left(Z,Y_1,Y_2^*\right) dY_2^* \end{aligned}$$

where $p_d\,(Z,Y_1,Y_2)=\Pr ob\left(D=1\middle|Z,Y_1,Y_2\right)$

Adopting the notation that $p_1=\Pr ob\left(Y_1=1\middle|X_1\right)$, and $p_2=\Pr ob\left(Y_2=1\middle|Y_1,X_2\right)$, then the log likelihood function for the observed data can be decomposed into three parts:

$$l_1=\sum_{i=1}^{N} Y_1\left\{\log p_1+(1-Y_1)\log(1-p_1)\right\}, \tag{4}$$

$$l_2=\sum_{\{D=0\}}\left\{(1-Y_1)\left[Y_2\log p_2+(1-Y_2)\log(1-p_2)\right]+\log(1-p_d)\right\}, \tag{5}$$

$$l_3=\sum_{\{D=1\}}\log\left[p_2^{(1-Y_1)}\frac{1}{1+e^{-z\theta_0-\theta_1 Y_1-\theta_2}}+(1-p_2)^{(1-Y_1)}\frac{1}{1+e^{-z\theta_0-\theta_1 Y_1}}\right], \tag{6}$$

Note to obtain the maximum likelihood estimate of α we only need to maximize l_1. To obtain maximum likelihood estimates of β and θ's we need to maximize l_2+l_3.

The incidence of disease is then estimated by the predicted probability $\Pr\text{ob}\left(Y_1=0,\,Y_2=1\right)$:

$$\hat{r}=\frac{1}{N}\sum_{i=1}^{N}(1-\hat{p}_1)\hat{p}_2 \tag{7}$$

2.2 Selection model with sampled data

In dementia studies, there are usually a large number of subjects who were not clinically evaluated at either the baseline or the follow up wave. Therefore, the above maximum likelihood method cannot be applied directly. We propose to use a modified EM algorithm (Dempster, 1977) to obtain the maximim likelihood estimates for the model parameters and disease incidence.

First we propose to use the baseline data to obtain consistent estimates of α. This can be achieved by performing a weighted logistic regression on the baseline dementia outcome on the covariates

X_1 with sampling weights. Strictly speaking, the estimates $\hat{\alpha}$ derived by weighted logistic regression are not necessarily maximum likelihood estimates. Instead, they are the so called pseudo-maximum likelihood estimates. Under the assumption that the model $f\left(Y_1 | X_1\right)$ is correctly specified and the sampling selection does not depend on Y_1, the pseudo-maximum likelihood estimates are the same as the maximum likelihood estimates (skinner *et al*, 1989). In dementia study sampling selection is independent of disease status given screening results. Therefore, it is important that screening groups be included in the covariates X_1 so that $\hat{\alpha}$ obtained by weighted logistic regression is also approximately maximum likelihood estimates.

Conditional on the estimates of α we then obtain estimates of β and θ by maximizing $l_2 + l_3$ using the EM algorithm.

For the E-step of the EM algorithm conditional on the observed data, $\hat{p}_1$ and the parameter estimates from the *t*th iteration we have:

$$E(l_2) = \sum_{\{D=0\}} \{[E(Y_2) - E(Y_1Y_2)] \log p_2^{(t)} + [1 - E(Y_1) - E(Y_2) + E(Y_1Y_2)] \log(1 - p_2^{(t)})$$

$$+ \log(1 - p_{d00}^{(t)})(1 - \hat{p}_1)(1 - p_2) + \log(1 - p_{d01}^{(t)})\, p_2^{(t)}\, (1 - \hat{p}_1) + \log(1 - p_{d11}^{(t)})\, \hat{p}_1 \},$$

where

$$p_{d00}^{(t)} = \frac{1}{1 + e^{-Z\theta_0^{(t)}}},\ p_{d01}^{(t)} = \frac{1}{1 + e^{-Z\theta_0^{(t)} - \theta_2^{(t)} Y_2}},\ p_{d11}^{(t)} = \frac{1}{1 + e^{-Z\theta_0^{(t)} - \theta_1^{(t)} Y_1 - \theta_2^{(t)} Y_2}}.$$

$$E(l_3) = \sum_{\{D=1\}} \{\log[p_2^{(t)} \frac{1}{1 + e^{-Z\theta_0^{(t)} - \theta_2^{(t)}}}, + (1 - p_2^{(t)}) \frac{1}{1 + e^{-Z\theta_0^{(t)}}}](1 - \hat{p}_1)$$

$$+ \log[\frac{1}{1 + e^{-Z\theta_0^{(t)} - \theta_1^{(t)} - \theta_2^{(t)}}} + \frac{1}{1 + e^{-Z\theta_0^{(t)} - \theta_1^{(t)}}}]\, \hat{p}_1 \}.$$

The expectations in E(l_2) are:

$$E(Y_1) = \begin{cases} Y_1 & \text{if } Y_1 \text{ observed,} \\ \hat{p}_1 & \text{if } Y_1 \text{ unobserved.} \end{cases}$$

$$E(Y_2) = \begin{cases} Y_2 & \text{if } Y_2 \text{ observed,} \\ \hat{p}_1 + p_2^{(t)} - \hat{p}_1 p_2^{(t)} & \text{if } Y_2 \text{ unobserved.} \end{cases}$$

$$E(Y_1Y_2) = \begin{cases} Y_1Y_2 & \text{if } Y_1 \text{ and } Y_2 \text{ observed,} \\ Y_1E(Y_2) & \text{if } Y_1 \text{ observed } Y_2 \text{ unobserved,} \\ Y_2E(Y_1) & \text{if } Y_2 \text{ observed } Y_1 \text{ unobserved,} \\ \hat{p}_1 & \text{if both } Y_1 \text{ and } Y_2 \text{ unobserved.} \end{cases}$$

For the M-step of the algorithm we derive estimates by maximizing the expected likelihood function $E(l_2)+E(l_3)$. A Newton-Raphson algorithm can be used to derive the maximum for $E(l_2)+E(l_3)$ at each parameter iteration. The process iterates between the E-step and the M-step until the parameter estimates converge. Incidence estimates can again be obtained using equation (7).

2.3 Variance estimation

Variance estimation for the model parameters and for the incidence estimation is difficult to derive analytically because the use of the EM algorithm and the fact that $\hat{\beta}$ and $\hat{\theta}$ are obtained conditional on $\hat{\alpha}$. Therefore we propose to use a non-parametric variance estimator, the jackknife estimator.

Suppose we are to obtain variance estimate for a parameter ϕ, we then carry out the following steps with the jackknife approach: (1). At the *i*th step, delete the *i*th observation from the data. (2). Carry out the analysis procedure on the new data set to obtain an estimate of ϕ, denoted by $\hat{\phi}_{(i)}$. (3). Repeat previous two steps for *I*=1, ..., *N*, generating a series of estimates $\hat{\phi}_{(i)}, \ldots, \hat{\phi}_{(N)}$. The jackknife variance estimates is given by:

$$\hat{\mathrm{var}}\left(\hat{\phi}\right)=\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\phi}_{(i)}-\bar{\phi}_{(.)}\right)^2, \text{ where } \bar{\phi}_{(.)}=\frac{1}{N}\sum_{i=1}^{N}\hat{\phi}_{(i)}.$$

Diggle and Kenward (1994) in their illustration on continuous longitudinal outcomes proposed to use the likelihood ratio test for testing that $\theta=0$. The test is especially important on θ_2 which determines if the missing data mechanism is nonignorable under the assumed missing data model. The likelihood ratio test cannot be used here because our estimates are obtained by maximizing a conditional expected log-likelihood, instead of the original full likelihood estimates. If the estimate α derived under the pseudo-likelihood approach is also maximum likelihood estimates, and the EM algorithm converges to a global maximum, then under some regularity conditions, the estimates $\hat{\beta}$ and $\hat{\theta}$ are asymptotically normally distributed with mean β and θ, respectively. Due to the missing data by sampling selection, the information matrices are difficult to derive analytically. The jackknife variance estimator has been used for data collected from complex sampling and is proven to be a consistent variance estimator (Krewski and Rao, 1981). Rao and Wu (1988) also demonstrated empirically that the jackknife variance estimators is more stable than the bootstrap estimator. Therefore we use the parameter estimates divided by its jackknife standard error estimates as the test statistic to test the null hypothesis of $\theta=0$. The test statistic follows a standard normal distribution approximately.

3. THE INDIANAPOLIS-IBADAN DEMENTIA STUDY

The Indianapolis-Ibadan (I-I) Dementia Study is an on-going comparative epidemiological study on dementia and Alzheimer's disease supported by NIH. The study populations are African Americans age 65 and older living in Indianapolis, USA and Nigerian Africans age 65 and older living in Ibadan, Nigeria. The primary goal of the study is to estimate and compare the prevalence and incidence of dementia and Alzheimer's disease in the two populations and identify potential risk factors that contribute to the difference in disease rates. The I-I study is designed to have a baseline wave followed by three follow up waves two years apart. The two-phase sampling design is used at each wave of the study. Based on the screening results from the first phase the subjects are divided into three groups: good performance, intermediate performance and poor performance. The second phase sampling uses stratified random sampling to select individuals for extensive clinical evaluation to determine disease status. In this paper we illustrate our methods using data from Indianapolis and focus on estimating the incidence of dementia. We will restrict our attention to data missing by death, therefore, ignore data missing by other causes. We therefore use a study cohort with 1997 subjects. 188 of them died before the first follow up wave.

Comparison of baseline characteristics suggests that the deceased subjects are significantly older, had poorer performances at baseline than the survivors. There are significantly higher proportions of demented subjects in the deceased group than in the surviving group. Male subjects have a marginally significant higher proportion of dying than female subjects. In dementia studies it has been reported that subjects in the poor performance groups are more likely to develop incident dementia than those in the good performance group. Therefore, a complete case analysis in which all deceased subjects are excluded may lead to the underestimation of disease incidence rates.

The set of covariates used to model baseline dementia probability, $f(Y_1 \mid X_1)$, are age at baseline, male sex, and baseline performance groups. The covariates, Z, for modeling the nonignorable missing data by death are age at death for the deceased subjects, age at follow up screening interview for survivors, male sex and baseline performance groups. The covariate, X_2, for modeling the conditional probability of dementia at follow up wave includes age at follow up, male sex, performance group at follow up wave.

Note that incidence estimation from (7) requires that all subjects have the covariate X_2 for predicting incident dementia. Age at death for the deceased subjects is used. However, the deceased subjects do not have the follow up performance groups. For illustration purpose we assume that the deceased subjects are in the same performance groups as at baseline. In other words, this assumption does not allow the deceased subjects to decline drastically to a lower performance group. Incidence estimates derived under this assumption may be conservative. Further research and efforts are under way to also deal with missing covariates in our proposed approach.

Details on the numerical methods and results for the model parameters are not presented here and are available from the first author. Incidence estimates using various approaches are presented in Table 1.

Table 1. Estimates annual incidence of dementia using the complete case analysis and the selection model approach. Data from the Indianapolis site of the Indianapolis-Ibadan Dementia Study. The annual incidence rates are derived by dividing each incidence estimates by the mean follow up time 1.74 years. The standard error estimates by jackknife method are in the parentheses.

Age Group	Complete Case Analysis	Selection Model
65-74	0.5405(0.1473)	0.7784(0.1577)
75-84	2.2054(0.4739)	3.1164(0.5709)
85+	5.4731(1.3596)	7.1174(1.1442)

A complete case analysis is used in which we use the disease models (1) and (2) excluding the missing data by death. This approach assumes that the data missing by death is either MCAR or missing dependent on the covariate X_1 and X_2 . Parameter estimates are derived using EM algorithm to deal with the data missing by sampling at baseline and at follow up. The selection model approach uses the disease models (1) and (2) with the missing data model (3). Parameter estimates are derived using the conditional EM algorithm outlined in Section 2.2. The selection model approach yields higher incidence estimates than the complete case analysis approach by adjusting for the missing data by death.

4. DISCUSSION

The approach presented here is only the first attempt to adjust incidence estimates for missing data in longitudinal dementia studies. It should be pointed out that the results obtained by assuming a model for the missing data mechanism can be very sensitive to the model assumption. Therefore, alternative approaches should also be taken to explore the sensitivity of results to the various assumptions in adjusting for the missing data. Alternative estimators may also be used for estimating the variance of the incidence estimates. For example, the stratified bootstrap sampling approach used by Clayton *et al* [2] can be adopted and compared with the performance of the jackknife estimator proposed here. The methods presented in this paper have largely depended on large sample properties. For example, the consistency of the jackknife estimator and the test statistic for $\theta = 0$. It will be interesting to assess the biases and efficiencies of the parameter estimators and the variance estimators in a well planned simulation study with reasonable sample sizes.

ACKNOWLEDGEMENTS

This research was supported in part by NH grant R01 09956, AG04518 and P30 AG10133.

REFERENCES

Beckett, L.A., Scherr, P.A. and Evans, D.A. (1992). Population prevalence estimates from complex samples. *J clin Epidemiol*, 45, 393-402.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling. *J.R. statisti. Soc*, B, 60, 71-87.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J Roy Statisti Soc*, 39, 1-38.

Diggle, P., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49-93.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and Balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Laird, N.M. (1988). Missing data in longitudinal studies. *Stat Med.* 1988; 7, 305-315.

Little, R.J.A. and Rubin, D.B. *Statistical Analysis with Missing Data*, John Wiley, New York, 1987.

Little, J.A. (1995). Modelling the drop-out mechanism in repeated-measure studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Rao, J.N.K. and Wu, C. J. (1988). Resampling inference with complex survey data. *Journal of American Statistical Association*, 83, 231-241.

Roberts, G., Rao, J.N.K., Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

Skinner, C.J., Holt, D. and Smith, T.M.F. *Analysis of Complex Surveys*, John Wiley, New York, 1989.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

WEIGHTING AND VARIANCE ESTIMATION FOR THE EXPLORATION OF POSSIBLE TIME TRENDS IN DATA FROM THE U.S. THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

J.L. Eltinge, V.L. Parsons and M.D. Carroll[1]

ABSTRACT

The U.S. Third National Health and Nutrition Examination Survey (NHANES III) was carried out from 1988 to 1994. This survey was intended primarily to provide estimates of cross-sectional parameters believed to be approximately constant over the six-year data collection period. However, for some variable (e.g., serum lead, body mass index and smoking behavior), substantive considerations suggest the possible presence of nontrivial changes in level between 1988 and 1994. For these variables, NHANES III is potentially a valuable source of time-change information, compared to other studies involving more restricted populations and samples. Exploration of possible change over time is complicated by two issues. First, there was of practical concern because some variables displayed substantial regional differences in level. This was of practical concern because some variables displayed substantial regional differences in level. Second, nontrivial changes in level over time can lead to nontrivial biases in some customary NHANES III variance estimators. This paper considers these two problems and discusses some related implications for statistical policy.

KEYWORDS: Change estimators; Collapse effect; Diagnostic; Interaction; Informative design; NHANES III; Stratified multistage sample survey; Timeliness; Variance estimator stability.

1. INTRODUCTION

1.1 The Third National Health and Nutrition Examination Survey (NHANES III)

The U.S. Third National Health and Nutrition Examination Survey (NHANES III) was carried out to assess the health and nutritional status of the U.S. noninstitutionalized civilian population. NHANES III was carried out over six years (1988-1994), due to the expense and logistical constraints associated with the medical equipment used in this study.

For most variables measured in NHANES III, population parameters were believed to be essentially constant over time. However, for at least three health phenomena, substantive considerations suggested the serious possibility of nontrivial changes in parameters over time. These phenomena were the following.

1. Serum lead. Over the past quarter century, U.S. legislation has imposed increasing restrictions on the use of lead in paint, gasoline and other industrial products. Also, in some

[1] J.L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 U.S.A.; V.L. Parsons and M.D. Carroll, National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782 U.S.A.

U.S. areas there have been serious public health efforts on lead abatement in older housing stock and on increasing public awareness of risks associated with lead. Consequently, there is substantive interest in exploring possible decreases in blood lead levels in the U.S. population. NHANES III collected serum lead data on over 10,000 adults; see Brody et al. (1994) for some general background. Due to the population-based nature of NHANES III, estimation of changes in population lead levels from NHANES III are potentially more informative than estimates from more limited clinical or community studies.

2. Overweight. Some recent clinical and community studies have suggested an increase in the proportion of the U.S. population that is overweight. NHANES III collected information on body mass index (BMI), defined as the ratio of a person's weight (in kilograms) divided by the square of the person's height (in meters). Persons generally are classified as overweight if their BMI values are above the mid 20s.

3. Smoking. Legal and medical controversies associated with smoking often are influenced by perceptions of whether rates and intensity of smoking have increased or decreased in recent years. However, many population-based surveys that collect data on smoking behavior rely on self-reported interview responses; two U.S. examples are the National Health Interview Survey and the Behavioral Risk Factor Surveys. Due to the increased stigma associated with smoking in recent years, there is a potential for concern that self-reported smoking data may have been subject to increased levels of nonreporting or underreporting. NHANES III collected similar self-reported interview data, but also recorded blood-sample measurements of cotinine, a byproduct of the metabolic breakdown of nicotine. Consequently, exploration of possible changes in cotinine levels can potentially offer insight into possible changes in the prevalence and intensity of true smoking behavior. In addition, comparison of cotinine and self-reported smoking for individuals may offer insight into the abovementioned possible changes in nonreporting or underreporting. For discussion of some relevant substantive issues, see, e.g., Pirkle et al. (1996) and references cited therein.

1.2 Limitations of NHANES III Data for Assessment of Change

As noted above, NHANES III was carried out primarily to estimate cross-sectional parameters believed to be approximately constant over 1988-1994. For this reason, as well as practical constraints involving respondent burden, NHANES III did not use a repeated-measurement or rotation-sample design commonly used for surveys intended specifically for estimation of change over time.

Instead, NHANES III used the following cross-sectional design. The U.S. was partitioned into 49 strata. Two primary sample units (PSUs, roughly equivalent to individual counties) were selected from each stratum. Additional subsampling was carried out at the secondary unit (segment), household and individual person levels.

Within a given stratum, one of the two selected PSUs was randomly assigned to Phase I (interview and examination work scheduled for 1988-1991) and the other selected PSU was assigned to Phase 2 (work scheduled for 1991-1994). We emphasize here that the term "phase" here is standard usage for NHANES III, but conceptually is somewhat different from the use of "phase" in customary discussion of multiphase sampling in the survey literature (e.g., Cochran, 1977, Chapter 12).

Phase 1 PSUs were then assigned to a specific year (1988-1989, 1989-1990 or 1990-1991) in a somewhat ad hoc manner intended to give some balance in coverage of specified regions and demographic groups in each year; and similarly for Phase 2 PSUs. Also, to reduce weather-related logistical problems, NHANES III fieldwork followed a cyclical pattern within a given year, with the south and southwest covered in the late fall, winter and early spring, and the northern half of the U.S. covered in the balance of the year. These design features lead to the following complications in the use of NHANES III data to assess possible changes over time.

a. Possible interaction of year with region. Despite attempts to give some coverage of each region in each year, regions did not receive balanced coverage in each year. For example, out of 16 PSUs covered by NHANES III in its final year (1993-1994), only one was from the Northeast region. There was not direct adjustment of weights to account for these one-year imbalances. Consequently, if a time change pattern differed substantially across regions, direct use of customary NHANES III weights could lead to biased estimation of the the true change in level over time.

b. Possible confounding of region with season. Substantive considerations indicate that some measured variables may display nontrivial seasonal and regional variation. Due to the abovementioned pattern of fieldwork in the north during warmer months and in the south during colder months, the seasonal and regional effects are potentially confounded. Thus, it is important to attempt to account explicitly for this combined seasonal/regional effect in the exploration of possible changes over time.

c. Limitations on effective sample size. For each of cases (1) through (3) in Section 1.1, there is some qualitative interest in the full population, but perhaps greater interest in time changes within specified demographic groups. This in turn entails limitations on effective sample size and on power to detect changes over time.

d. Bias in variance estimators due to implicit collapse. Finally, the customary NHANES III variance estimator is based on a sum of squared differences between residual terms associated, respectively, with the Phase 1 and Phase 2 PSUs within a given stratum. For population characteristics that are constant over time, the observed (Phase 1 - Phase 2) differences in a given stratum are attributable to sampling variability and thus lead to an approximately unbiased variance estimator. However, standard stratum-collapse arguments indicate that a nontrivial time trend will inflate the expected squares of the observed (Phase 1 - Phase 2) difference in a given stratum. This in turn implies that time changes can induce a nontrivial positive bias in the customary NHANES III variance estimator.

The remainder of this paper addresses issues (a) through (d) in the following way. Section 2 explores definitions of finite-population time changes, with special reference to comparison of the magnitude of change to standard errors or to specified differences Δ_o considered to be a practical importance. Section 3.1 describes some possible variantes and extensions of the ideas in Section 2. Finally, Section 3.2 discusses some related possible implications for statistical policy. Due to limitations of space, discussions here is limited to relatively general material; technical details and empirical results will be discussed further elsewhere.

2. POPULATION, PARAMETERS AND SAMPLES

2.1 Definition of Parameters Over Time

Following the motivating example introduced in Section 1, consider a nominally cross-sectional survey carried out over an extended period of time, $\left[t_1, t_2\right]$, say, for a given $t \in \left[t_1, t_2\right]$, let U_t be the population of interest at time t, e.g., all persons over age 20 living in the U.S. at time t. Also, let N_t be the number of elements in U_t, and let Y_{ti} be a vector of attributes for element $i \in U_t$ at time t. In addition, let

$$Y_t = \sum_{i \in U_t} Y_{ti}$$

be the associated vector of population totals; and let $\theta_t = f(Y_t)$ be our parameter of principal interest, where $f(\cdot)$ is a well-defined function with continuous first and second derivatives.

In general, we anticipate that U_t, N_t, Y_t and θ_t will change as t increases from t_1 to t_2. For instance, in the abovementioned U.S. example, additional persons reach age 20 or immigrate over $t \in \left[t_1, t_2\right]$, while other persons who were U.S. adults at time t_1 will have died or emigrated by time t_2. In addition, as t increases, a person i who is continuously present in $U_t, t \in \left[t_1, t_2\right]$ may have changes in Y_{ti}, e.g., due to changes in health status.

To avoid notational complications, let U_{12} be the union of all distinct population elements in $U_t, t \in \left[t_1, t_2\right]$. Also, for $i \in U_{12}$, if $i \notin U_t$, then define $Y_{ti} = 0$. Within this framework, consider the definition of a target parameter averaged over $\left[t_1, t_2\right]$. One possible definition is the simple time average of θ_t:

$$\bar{\theta}_{12} = \left(t_2 - t_1\right)^{-1} \int_{t_1}^{t_2} \theta_t dt \ .$$

A second definition (in some cases more attractive due to consistency issues) involves similar averaging of the vector totals Y_t:

$$\theta_{12}^* = f\left(Y_{12}^*\right)$$

where

$$Y_{12}^* = \left(t_2 - t_1\right)^{-1} \int_{t_1}^{t_2} Y_t dt \ .$$

There are two related cases in which the averaged parameters $\bar{\theta}_{12}$ or θ_{12}^* are of substantive interest. First, let Δ_o be the smallest change in θ_t over time that would be considered of practical importance. If

(C.1) the differences $\theta_t - \theta_{t_1}, t \in \left(t_1, t_2\right]$, are uniformly small relative to Δ_o

then $\bar{\theta}_{12}$ may be considered a satisfactory approximation for θ_t across $t \in \left[t_1, t_2\right]$. Second, in some cases the instantaneous tru parameter θ_t is subject to seasonal variability or other short-term effects that are not of primary substantive interest. For these cases, $\hat{\theta}_{12}$ or a similarly averaged parameter may be of greater substantive interest than the instantaneous parameter θ_t.

A third, potentially more problematic case arises if: (a) the instantaneous parameter θ_t is of principal substantive interest; (b) there is no articulated scientific consensus regarding a practical-difference quantity Δ_o; and (c) for the customary design-based estimator θ_{D12}, say, of $\bar{\theta}_{12}$:

(C.2) the differences $\theta_t - \bar{\theta}_{12}, t \in \left[t_1, t_2\right]$, are uniformly small relative to the design-based standard error of $\hat{\theta}_{D12}$.

For this third case, estimation of $\bar{\theta}_{12}$ may be potentially of interest. However, if a subsequently identified Δ_o is small relative to the standard error of $\hat{\theta}_{D12}$ and if condition (C.1) does not hold, then the information conveyed by $\hat{\theta}_{D12}$ may be of somewhat limited practical value. This arises, for example, when $\bar{\theta}_{12}$ is a parameter of a relatively small subpopulation.

2.2 Approximately Unbiased Estimation of θ_{12}^*

Let *D* be a nominally cross-sectional design implemented to sample from U_{12}. Note that a realization of *D* effectively specifies *both* the elements of U_{12} that are selected *and* when each selected element *i* is measured. Also, note that in many cases *D* will be based on an incomplete frame, e.g., due to problems with birth and death units, new construction, and errors in time-related sub-population membership identification. In addition, partition the interval $\left(t_1, t_2\right]$ into *J* subintervals $\left(t_{j1}, t_{j2}\right]$ and define

$$\bar{\theta}_{(j)} = \left(t_{2j} - t_{1j}\right)^{-1} \int_{t_{1j}}^{t_{2j}} \theta_t \, dt \ .$$

For our NHANES III example, the subintervals $\left(t_{j1}, t_{j2}\right]$ of principal interest are six one-year segments or two three-year segments. In addition, let $\hat{\theta}_{Dj}$ be a design-based estimator of $\bar{\theta}_{(j)}$ based on a direct application of customary cross-sectional weights originally developed for the design *D* and the full time period $\left(t_1, t_2\right]$.

Work with estimation of the parameters $\bar{\theta}_{12}, \theta^*_{12}$ or $\bar{\theta}_{(j)}$ also depends on the informativeness of the design *D*. For example, consider the following assumption.

(C.3) The sample design *D* is uninformative for the differences $\theta_t - \theta_{t_1}, t \in \left(t_1, t_2\right]$

Note that under conditions (C.1) through (C.3), direct implementation of the design *D* with the estimator $\hat{\theta}_D$ will lead to approximately unbiased estimation of θ^*_{12} , which in turn is a satisfactory proxy parameter for $\theta_t, t \in \left(t_1, t_2\right]$. This appears to be the case for many cross-sectional surveys carried out over a relatively short period of time. However, if either condition (C.1) or (C.3), or both, are not satisfied, then time effects may cause serious problems with the customary parameter estimator $\hat{\theta}_D$; some simple cases are considered in Section 3.

Finally, note that the preceding work focused on the instantaneous parameters θ_t and Y_t , and on time averages thereof. In some cases, one could consider changes that are approximately linear in $t \in \left[t_1, t_2\right]$. For such cases, one could also consider estimation of a slope coefficient, defined as,

$$\beta_1 = \left(t_2 - t_1\right)^{-1} \left(\theta_{t_2} - \theta_{t_1}\right) .$$

3. DISCUSSION

3.1 Variants and Extensions

Sections 1 and 2 considered issues associated with analysis of possible time changes in data from NHANES III. These ideas can be extended in several ways. First, recall that NHANES III had a relatively small number of primary sample units assigned to a given year (approximately 16, depending on the year). For surveys with large numbers of primary units included per year, one could model inclusion probabilities for a given year as a function of region and other PSU-level explanatory variables. Following, e.g., Little (1986) and Czajka et al. (1992), one could then use the resulting estimated probabilities for within-year weighting adjustment.

Second, we restricted attention to estimation of functions of the parameters $\theta_t, t \in \left[t_1, t_2\right]$. However, in many practical cases, the time-averaged parameters $\bar{\theta}_{12}$ or θ^*_{12} are sometimes used as proxies for the parameters $\theta_t, t > t_2$. For example, estimates based on the 1988-1994 NHANES III data are used in current (1998) health-policy discussions. Justification for this approach would involve the assumption that the differences $\theta_t - \bar{\theta}_{12}$ are uniformly small for *t* in some interval $\left[t_2, t_3\right]$, say; and by information from related empirical studies that support this assumption. Related issues arise when two independent cross-sectional studies are carried out in the time intervals $\left[t_1, t_2\right]$ and $\left[t_3, t_4\right]$, say, where $t_3 > t_2$. Then for parameters θ_t for which the differences $\theta_t - \theta_{t_1}$ are uniformly small over $t \in \left[t_1, t_4\right]$, one could consider combined use of data from these two cross-sectional surveys to estimate a new averaged parameter,

$$\bar{\theta}_{14} = \left\{\left(t_2 - t_1\right) + \left(t_4 - t_3\right)\right\}^{-1} \left(\int_{t_1}^{t_2} \theta_t \, dt + \int_{t_3}^{t_4} \theta_t \, dt \right) ,$$

say. This is of special interest for small subpopulations for which the standard error of $\hat{\theta}_{D12}$, say, is relatively large. A practical example of this issue will arise in 1999 and subsequent years, when the next U.S.

National Health and Nutrition Examination Survey is expected to begin production of data.

3.2 Policy Issues

To complement the preceding discussion, we close with some brief comments on statistical policy. Specifically, when a survey is conducted over time, the following issues can be relevant to statistical agency policy regarding publication of official estimates and research results.

1. Publication of "most recent estimates" and related issues of timeliness.

   a. If a given parameter displays sufficiently strong changes over time, published estimates of, e.g., a six-year average level may be relatively uninformative. Consequently, in some cases in is worthwhile to consider modified methods intended to estimate parameters for a more limited time period, e.g., the last one to three years covered by the survey.

   b. Point (1.a) has implications for the broader issue of *timeliness*. For a cross-sectional survey with fieldwork carried out in a relatively brief period, timeliness generally involves two considerations: (i) rapid dissemination of results after completion of data collection; and (ii) relatively frequent repetition of the survey. See e.g., Bradburn (1997) for discussion of (i) and (ii). Implicit in point (ii) is the assumption that more recent data can lead to better indications regarding the parameter θ_t at the current time *t*. This may be widely accepted for relatively volatile parameters associated with, e.g., economic statistics, opinion polling, or tracking a winter influenza epidemic. This may also be widely accepted, e.g., for health parameter estimates based on measurement methods that have recently undergone substantial improvements. However, timeliness issues become considerably more complex when substantive considerations indicate that the parameters in question are essentially constant over time. In some such cases, the best indication of the current parameter θ_t may be provided by combining data collected over an extended period of time, supplemented by appropriate checks for possible change over that extended period; of the comments in Section 3.1 on the combined parameter $\bar{\theta}_{14}$.

   c. In addition, issues (1.a) and (1.b) can be complicted by: (i) the bias issues mentioned in Section 1 and 2 above; (ii) increases in sampling errors due to reduction of effective sample sizes; and (iii) the need to avoid possible public misinterpretation of moderate changes in mean estimates over time.

2. Guidelines for formal identification of nontrivial change over time. In keeping with point (1.a.iii) above, it is also useful for an agency to establish guidelines for identification and interpretation of nontrivial changes over time. Some relevant issues are as follows.

   a. Risks of inflated Type I error rates, especially if possible trends are explored for a large number of variables and subpopulations.

   b. Public health implications of: (i) reporting a change that subsequently turns out to be strictly an artefact of sampling error; or (ii) delays in reporting a true nontrivial change over time.

   c. Evaluation of time changes observed in a given dataset, within the context of: (i) substantive considerations consistent with the presence or absence of a nontrivial change over time; and (ii) applicable evidence provided by related surveys or limited-population case studies.

ACKNOWLEDGEMENTS

The authors thank Deborah Brody, Vicki Burt, Cliff Johnson, Anne Looker and Kurt Maurer for helpful comments on the NHANES III data. Research by J. Eltinge was supported in part by the U.S. National Center for Health Statistics. The views expressed here are those of the authors and do not necessarily represent the policies of the U.S. National Center for Health Statistics.

REFERENCES

Bradburn, N.M. (1997), "The Future of Federal Statistics in the Information Age." Morris Hansen Memorial Lecture presented to the Washington Statistical Society, October 22, 1997.

Brody, D.J., Pirkle, J.L., Kramer, R.A., Flegal, K.M., Matte, T.D., Gunter, E.M. and Paschal, D.C. (1994). "Blood Lead Levels in the U.S. Population: Phase I of the Third National Health and Nutrition Examination Survey." *Journal of the American Medical Association* **272**, 277-283.

Cochran, W.G. (1977), *Sampling Techniques* (3rd. ed.), New York: John Wiley.

Czajka, J.L., Hirabayashi, S.M. Little, R.J.A. and Rubin, D.B. (1992), "Projecting from Advance Data Using Propensity Modelings: An Application to Income and Tax Statistics." *Journal of Business and Economic Statistics* **10**, 117-131.

Little, R.J.A. (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review* **54**, 139-157.

National Center for Health Statistics (1996). NHANES III Reference Manual and Reports. CD-ROM, Data Dissemination Branch, U.S. National Center for Health Statistics, Hyattsville, MD 20782.

Pirkle, J.L., Flegal, K.M., Bernert, J.T., Brody, D.J., Etzel, R.A. and Maurer, K.R. (1996), "Exposure of the U.S. Population to Environmental Tobacco Smoke: The Third National Health and Nutrition Examination Survey, 1988 to 1991." *Journal of the American Medical Association* **275**, 1233-1240.

SESSION 7

APPLICATIONS

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

THE LONG-TERM CONSEQUENCES OF GROWING UP WITH A SINGLE PARENT

Miles Corak and Andrew Heisz,[1]

ABSTRACT

The objective of this research project is to examine the long-term consequences of being raised in a single parent household. We examine the impact of parental separation or divorce on the adult labour market behaviour of children ten to fifteen years after the event. In particular, we relate the family income and household characteristics of a cohort of individuals who are 16 to 19 years of age in 1982 to their labour market earnings, reliance on social transfers (UI and Income Assistance), and marital/fertility outcomes during the early 1990s, when they are in their late 20s and early 30s. Our data is based upon the linked income tax records developed by us at Statistics Canada, the Survey of Labour and Income Dynamics, and the National Longitudinal Survey of Children and Youth.

Existing studies of the consequences of single parenthood suffer from the fact that it is difficult to discern the impact of parental separation from the impact of low income. Single-parenthood and low income are inter-related events that both have an impact on the ultimate labour market outcomes of children. We propose to devise a "natural" experiment to isolate the separate effects of these variables by using the group of individuals who suffered a loss of a parent through death as a control group. We also devise an experiment based upon the fact that there were significant changes in the legislation governing divorces in Canada.

We feel that our research will be relevant to federal government policy toward child poverty. Current policy discussions implicitly assume that there are long-term consequences to child poverty, and that greater financial assistance will be of long-term benefit to the children. Isolating the "true" effect of income and of family structure will inform this discussion.

We also feel that our proposal best fits into the theme of the conference described as "Causal Analysis of Panel Data" as it illustrates the use of "natural" experiments as a means of developing causal models with longitudinal data.

[1] Miles Corak and Andrew Heisz, Statistics Canada.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

PROBABILITY OF VICTIMIZATION OVER TIME: RESULTS FROM THE U.S. NATIONAL CRIME VICTIMIZATION SURVEY

Sharon Lohr and Siyi Sun[1]

ABSTRACT

Victimizations are not randomly scattered through the population, but tend to be concentrated in relatively few victims. Data from the U.S. National Crime Victimization Survey (NCVS), a multistage rotating panel survey, are employed to estimate the conditional probabilities of being a crime victim at time t given the victimization status in earlier interviews. Models are presented and fit to allow use of partial information from households that move in or out of the housing unit during the study period. The estimated probability of being a crime victim at interview t given the status at interview (t-1) is found to decrease with t. Possible implications for estimating cross-sectional victimization rates are discussed.

KEY WORDS: Gross flow estimation; Longitudinal data; Nonresponse; Panel attrition

1. INTRODUCTION

The U.S. National Crime Victimization Survey (NCVS) is one of the richest sources of information available for gaining knowledge of nationwide victimization rates and experiences of crime victims. The NCVS design prescribes interviewing the residents of each housing unit in the sample every six months for a total of seven interviews. But much of the longitudinal information in the data has gone uninvestigated, and the data used primarily for cross-sectional estimation of victimization rates.

Lehnen and Reiss (1978) studied early NCVS (called the *National Crime Survey*, NCS for data prior to 1992) data for effects of the panel design of the survey on estimates. They found that households that were in the sample longer tended to have lower victimization rates, and that respondents were more likely to report a victimization to the survey if they had reported a victimization on a previous interview. Saphire (1984) found similar results—that victimization rates were related to time-in-sample. Fienberg (1980) fit Markov models to estimate transition probabilities for households that were victimized at least twice. All of these analyses used the longitudinal files constructed from the earliest NCS data, 1972-75, by Reiss (1980).

Lehnen and Reiss suggested that respondent fatigue may be a problem; respondents familiar with the survey may be less likely to report victimizations because they know that reporting an incident on the screening questionnaire will lead to further follow-up questions. The higher victimization rate for households that were previously victimized could be due to familiarity with the survey—repeat respondents know more about what is considered a victimization, and might have better recall—or could be related to victim proneness.

Stasny (1990) and Conaway (1993) fit models to NCVS data that incorporated explicit modeling of nonresponse. Stasny and Conaway both used a symmetry model for the gross flows; both found that incorporating partial information into the model increases estimated victimization rates.

[1] Department of Mathematics, Arizona State University, Tempe, AZ USA 85287-1804

Although modelling households' experiences with crime over time is not one of the primary objectives of the NCVS, the panel design allows such longitudinal analyses. In this paper, we consider models based on Stasny (1986) for the panel data that fit the transition probabilities of victimization. We employ a subset of the data consisting of those households that are present for the first interview.

2. NCS LONGITUDINAL DATA STRUCTURE

The NCS and NCVS are both stratified multistage cluster surveys with a rotating panel design. Occupants of housing units (HUs) selected to be in the sample are interviewed every six months for 3 1/2 years, for a total of seven interviews. To avoid an abrupt change of all HUs at the end of 3 1/2 years, entry of new HUs into the sample is staggered; a new rotation group enters the sample every six months, replacing a rotation group that has completed all seven interviews. The first interview of HU residents is used for bounding purposes only, and not for calculation of annual victimization rates.

The NCS Longitudinal File (U.S. Department of Justice, 1993), abbreviated NCSL, contains records for the 33,272 HUs in rotation groups 4-6 of sample J14. A substantial number of those HUs have no associated data, because the HU is vacant or does not exist, leaving 24,499 households that participate in the first interview.

The NCS and NCVS follow HUs, not households (HHs) over time; consequently, if HH *x* moves out of the HU at interview 3 and HH *y* moves into the HU at interview 4, then the victimization record for the HU is that of HH *x* for interviews 1-3 and HH *y* for interviews 4-7. Thus, the record for a HU may contain responses from several different HHs. To construct longitudinal models for HHs in the NCVS, then, we must be careful to include only information for one HH per HU. Of the original 24,499 households that provide data for the first interview, 16,305 (67%) complete all seven interviews. The persistence of the original households is illustrated in Table 1. Table 1 presents a different picture of longitudinal nonresponse than found in earlier studies of NCS data. In the data set from Stasny (1990), only about 70% of households participating in Interview 1 in a given year also participated in Interview 2.

The reason that a HH drops out of the sample may be associated with its previous crime victimization experience or some of the demographic characteristics. Table 2 provides some information about the differences (victimization experience and demographic characteristics) between the HHs that drop out and those that stay in the sample. Table 3 shows the difference in victimization experience for the respondents and longitudinal nonrespondents. HHs that stay in the sample longer are more likely to be crime-free than those that drop out earlier. HHs that complete the seven interviews have the fewest victimizations. Also, the percentage of the HHs that report at least one victimization in Interview 1 (the first row) is higher than the percentage of HHs reporting victimizations in Interview 2 through Interview 7, for each column. This might be caused by telescoping—HHs in Interview 1 may recall incidents that occurred more than six months ago, and have no previous interview information to be used for editing out duplicate incidents.

3. CONDITIONAL PROBABILITIES OF VICTIMIZATION

In all of the following, we consider only two states for a HH's victimization status at an interview: nonvictim (no victimization incidents during the 6-month period preceding the interview) or victim (at least one victimization incidents during the 6-month period).

Stasny (1986) considered the observed data to be the result of a two-stage process. The first stage, unobserved, pretends there is no nonresponse. Thus P_{ij} is the probability that an observation is in state *i* at time 1 and state *j* at time 2, for $i,j \in \{N,V\}$. For the second stage, an observation in the (i,j) cell of the victimization matrix for times $(t - 1)$ and t has probability $\varphi_{ij}(t)$ of missing the information at time $(t - 1)$, and probability $\psi_{ij}(t)$ of missing the information at time *t*. Using Stasny's two-stage model, the underlying probabilities for observed victimization data are:

			Interview *t*	
		Nonvictim	Victim	Missing
Interview	Nonvictim	$(1-\varphi_{NN}-\psi_{NN})\ P_{NN}$	$(1-\varphi_{NV}-\psi_{NV})\ P_{NV}$	$\Sigma\ \psi_{Nj}\ P_{Nj}$
t – 1	Victim	$(1-\varphi_{VN}-\psi_{VN})\ P_{VN}$	$(1-\varphi_{VV}-\psi_{VV})\ P_{VV}$	$\Sigma\ \psi_{Vj}\ P_{Vj}$
	Missing	$\Sigma\ \varphi_{iN}\ P_{iN}$	$\Sigma\ \varphi_{iV}\ P_{iV}$	

We fit five models for the φ's and ψ's:

	Model				
	A	B	C	D	E
$\varphi_{ij}(t)$	$\lambda_{t-1(j)}$	λ_{t-1}	$\lambda_{(j)}$	$\lambda_{t-1(j)}$	λ_{t-1}
$\psi_{ij}(t)$	$\lambda_{t(i)}$	λ_t	$\lambda_{(i)}$	λ_t	$\lambda_{t(i)}$

In Model A, the saturated model, the probability of missing one of the two consecutive interviews depends on the interview period as well as on the victimization classification when observed. Under Model B, the probability of missing either of the two consecutive interviews (Interview t-1 and Interview t) depends only on the interview period, not the victimization classification. Under Model C, the probability of missing either of the two consecutive interviews depends only on the victimization classification. Models D and E are hybrid models not considered by Stasny (1986), but thought to be contenders here because of the possibility that HHs moving in to the survey may differ from those moving out after a period in the survey. Under the models considered, $\varphi_{ij}(t)$,the probability of missing the first interview, is a function only of *j* and *t*, and $\psi_{ij}(t)$ is a function of *i* and *t*. Thus the likelihood is a product of one factor involving the P_{ij} and a second factor involving the φ's and ψ's.

Model A, of course, provided a perfect fit to the data. A surprising result was that Model D also fit well, especially for the later periods. The estimated expected cell counts are in Table 4 and the estimates of the parameters for Model D in Table 5.

Unfortunately, clustering and stratification information is not available on the NCS public use data sets for these years, so we were not able to incorporate the clustering information into our estimates of the standard errors. However, we note that the overall design effect of the NCS is about two, so dividing the Pearson and likelihood-ratio chi-square statistics in Table 5 by two should give p-values that are close to the true values (Rao and Scott, 1992). After this adjustment, Model D fits the transition periods 4-5, 5-6, and 6-7 well. It fits the data in the earlier interviews less well.

Incorporating the partial information led to increased estimated probabilities in all victimization categories. Model D, which fits the last three time periods well, suggests that for HHs that participate in at least four interviews, the probability of nonresponse on the next interview is unrelated to victimization status. HHs that report at least one victimization at time *t*, however, are 1.5-2 times more likely to have been nonrespondents at time (t-1) than are HHs reporting no victimizations at time *t*. This may be because they move as a result of previous victimization, or may be because HHs on their first interview always report more victimizations because of telescoping. We note that between 22 and 24% of the "drop-in" HHs report at least one victimization in the interview for which they are present, a figure consistent with the percentages in the first row of Table 3. This lends support to the theory that a large part of the difference between drop-ins and drop-outs is due to the lack of a bounding interview for the drop-ins.

4. INCORPORATING THE COMPLEX SURVEY DESIGN

In the models reported in this paper, we did not use the survey weights or account for the complex design in variance estimation. We were unable to do the latter because clustering information was not available in the data set. However, we note that pseudostratum information has very recently been added to the 1992-95 NCVS public-use data sets (U.S. Department of Justice, 1998). This will permit replication methods to be used in estimating variances for longitudinal as well as cross-sectional analyses. Results from this research, with additional models incorporating the entire victimization history of HHs, are forthcoming.

Weights for the NCS and NCVS are calculated to give cross-sectional estimates of victimizations, to estimate number of victimizations for the entire population. The survey is designed so that initially, every HH has equal probability of selection. Weights are then calculated to adjust for subsampling in some areas, to adjust for nonresponse, and to conform survey population estimates to independent estimates based on the 1990 Census adjusted for the undercount. Thus, the same HH will have different cross-sectional weights at different times, reflecting changes in nonresponse and population over time.

We considered the weight provided for the first time in sample as the weight for each HH. This presents some difficulties, as the three rotation groups in NCSL have different start dates and these weights depend on the composition of the sample at the different dates; however, it gave us a basis for exploring the effects of incorporating weights into the analysis. We found that the probabilities in Table 3 and in the gross flow models of Section 3 were little changed when weights were used. In other surveys, though, weights may make more of a difference in longitudinal analyses. The models in Section 3 may be extended for non-self-weighted surveys by adopting a pseudo-MLE approach as in Binder (1983). Latouche and Michaud (1997) studied the effects of different weighting structures in the Canadian Survey of Labour and Income Dynamics.

Incorporating partial information from respondents present at only one of the two interviews increases the estimated conditional and unconditional probabilities of victimization. The models fitted in this paper, however, suggest that much of the increase is due to the incoming HHs that do not have a bounding interview. Further investigation is needed of the effect of including these unbounded "drop-in" HHs when estimating cross-sectional victimization rates.

REFERENCES

Binder, D. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-294.

Conaway, M.R. (1993). Non-ignorable Non-response Models for Time-ordered Categorical Variables. *Applied Statistics*, 42, 105-115.

Fienberg, S.E. (1980). The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. *The Statistician*, 29, 313-350.

Latouche, M. and Michaud, S. (1997). Cross-sectional Weighting of a Longitudinal Survey and its Impact on Analysis. Paper presented at Joint Statistical Meetings, Anaheim.

Lehnen, R.G. and Reiss, A.J. (1978). Response Effects in the National Crime Survey. *Victimology*, 3, 110-124.

Rao, J.N.K. and Scott, A.J. (1992). A Simple Method for the Analysis of Clustered Binary Data. *Biometrics*, 48, 577-585.

Reiss, A.J. (1980). Victim Proneness by Type of Crime in Repeat Victimization. Indicators of Crime and Criminal Justice: Quantitative Studies. Washington, D.C.: U.S. Government Printing Office, pp. 41-53.

Saphire, D.G. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey.* New York: Springer-Verlag.

Stasny, E.A. (1986). Estimating Gross Flows Using Panel Data with Nonresponse: An Example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

Stasny, E.A. (1990). Symmetry in Flows Among Reported Victimization Classifications with Nonrandom Nonresponse. *Survey Methodology*, 16, 305-330.

U.S. Department of Justice, Bureau of Justice Statistics (1993). National Crime Surveys: National Sample, 1986-1991 (Computer file). Conducted by U.S. Dept. of Commerce, Bureau of the Census. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (producer and distributor).

U.S. Department of Justice, Bureau of Justice Statistics (1998). National Crime Victimization Survey, 1992-1995 [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. 4th ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].

ACKNOWLEDGEMENTS

This research was partly supported by a grant from the U.S. National Institute of Justice. The original collector of the data, ICPSR, and the Department of Justice bear no responsibility for uses of this collection or for interpretations or inferences based upon such uses.

TABLES

Table 1. Households that stay in or drop out of NCSL. The horizontal arrows indicate the households that stay in the survey; the vertical arrows indicate households that drop out. Of the 24499 households that participated in the first interview, 22782 also participated in the second interview and 1717 did not participate in the second interview.

Interview Number

1		2		3		4		5		6		7
24499	→	22782	→	21475	→	19996	→	18683	→	17288	→	16305
↓		↓		↓		↓		↓		↓		
1717		1307		1479		1313		1395		983		

Table 2. Differences in households with longitudinal nonresponse. All percentages, and the average age, are for interview 1. The reference person (RP) is identified as owning or renting the living quarters.

	Number of Interviews Completed						
	1	2	3	4	5	6	7
Number of HHs	1717	1307	1479	1313	1395	983	16305
% who own HU	34.8	37.4	49.4	52.7	57.0	53.7	74.5
% 1-person record	39.1	36.9	32.9	30.9	29.1	30.6	22.5
% married RP	43.6	43.5	49.6	50.9	53.5	52.2	64.3
% crime-free	76.0	76.5	78.0	78.8	79.4	78.1	80.9
% white RP	84.7	83.5	85.0	84.2	85.7	85.0	88.5
% never-married RP	25.7	26.9	18.8	17.9	17.1	19.2	10.4
% income < $15K	42.9	41.7	35.4	36.6	33.2	35.1	28.8
Average age, RP	41.0	41.6	44.1	45.6	46.2	45.1	49.4

Table 3. Percent of HHs with at least one victimization experience in the six months prior to Interview *n*.

	Number of Interviews Completed						
n	1	2	3	4	5	6	7
1	34.00	23.49	22.04	21.25	20.65	21.87	19.07
2	-----	18.82	15.75	16.37	15.77	18.11	14.32
3	-----	-----	17.58	15.92	14.92	14.65	13.56
4	-----	-----	-----	14.78	15.27	13.63	13.26
5	-----	-----	-----	-----	14.77	13.84	12.80
6	-----	-----	-----	-----	-----	15.36	12.27
7	-----	-----	-----	-----	-----	-----	12.03

Table 4. Estimated expected cell counts from gross flow models. Estimated standard deviations of the counts range from 138 for the first entry in the table (18517) to 27 for the last entry in the table (611). Column 5 estimates p(V|i), the conditional probability that a HH is in class V at time *t* given victimization status *i* at time (*t* − 1), from records that have complete data at both times. The second estimate of p(V|i), in Column 6, is from the model fit in Section 3.

		Time t		p(V|i) (complete data)	p(V|i) (model)
Time t-1		Nonvictim	Victim		
1-2	Nonvictim	18517	2705	.121	.127
	Victim	3889	1507	.268	.279
2-3	Nonvictim	18291	2643	.119	.126
	Victim	2692	1048	.267	.280
3-4	Nonvictim	17649	2470	.115	.123
	Victim	2431	909	.258	.272
4-5	Nonvictim	16521	2232	.111	.119
	Victim	2170	792	.252	.267
5-6	Nonvictim	15721	2018	.106	.114
	Victim	1942	707	.249	.264
6-7	Nonvictim	14678	1828	.103	.111
	Victim	1760	611	.243	.258

Table 5. Estimates of parameters for Model D. Standard deviations (assuming independent observations) are in parentheses behind the estimates. X^2 and G^2 are the Pearson and likelihood-ratio test statistics for goodness-of-fit, under the assumption that HHs are independent.

	$\varphi_{iN}(t)$	$\varphi_{iV}(t)$	$\psi_{ii}(t)$	X^2	G^2
Interviews	$\lambda_{t-1(N)}$	$\lambda_{t-1(N)}$	λ_t		
1-2	.0721 (.0017)	.1199 (.0050)	.0645 (.0015)	15.8	17.2
2-3	.0685 (.0017)	.1235 (.0054)	.0530 (.0014)	14.5	13.7
3-4	.0761 (.0019)	.1353 (.0059)	.0630 (.0016)	14.5	15.7
4-5	.0700 (.0019)	.1357 (.0062)	.0605 (.0016)	1.5	1.5
5-6	.0763 (.0020)	.1415 (.0066)	.0683 (.0018)	3.6	5.5
6-7	.0754 (.0021)	.1425 (.0071)	.0521 (.0016)	7.4	7.0

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

AN INCREMENT-DECREMENT MODEL OF SECONDARY SCHOOL PROGRESSION FOR CANADIAN PROVINCES

Geoff Rowe and Edward Chen[1]

ABSTRACT

A model of secondary school progression has been estimated using data from the 1991 School Leavers Survey conducted by Statistics Canada. The data on which the school progression model was based comprised current educational status and responses to retrospective questions on the timing of schooling events. These data were sufficient for approximate reconstruction of educational event histories of each respondent. The school progression model was designed to be included in a larger, continuous time micro-simulation model. Its main features involve estimation -- by age, month of birth and season for both sexes in each province -- of rates of graduation, of dropout, of return and of dropout graduation. Estimation was reinforced with auxiliary 1991 Census and administrative data.

KEY WORDS: multistate life table, education, microsimulation, hazard function

1. INTRODUCTION

The secondary school progression model was developed as a component of a larger micro-simulation model (LifePathsSL) whose immediate purpose was to facilitate analysis of post-secondary education funding policies, student needs and resources as well as post-education outcomes. It is particularly important that such a model accurately reflect demand for post-secondary education in terms of the numbers of secondary graduates available and that it meshes with other components of the larger model. In addition, it is important that such a model faithfully reflect the structural differences among the education systems in place among the provinces. Thus, the aim was to provide an accurate description of the dynamics of secondary school progression rather than an explanatory model. For example, the model does not include socio-economic or other parental influences, which would be obvious candidates for inclusion in a more comprehensive model. LifePathsSL was developed, in part, under contract to Human Resources Development Canada.

The typical dynamics of elementary/secondary school progression are simple -- they comprise lock step grade advancement from the time of (say) 1st grade entry to graduation 10-12 grades later depending on province. Complexity only arises if students skip or are held back a grade, if there are dropout (DO) spells of appreciable length or both. The implications of such factors could be studied directly only by analyzing linked administrative micro-data and/or prospective or retrospective survey data. Currently, the dynamics are not reflected well in administrative data of the form available -- as is indicated in a comparison of 'High School Non-completion Rates' Table 1 (Gilbert et.al. 1993):

[1] Geoff Rowe and Edward Chen, Statistics Canada , Socio-Economic Modeling Group, 24th Floor R.H.Coats Building, Ottawa, Ontario, Canada, K1A 0T6

Table 1: Selected High School Non-completion Rates

	Administrative Data		Survey Data		
	Complement of age cumulative graduation rate	Apparent cohort dropout rate	Census Age 20	Labour Force Survey Age 20	School Leavers Survey Age 20
Canada	31%	32%	21%	20%	18%

2. 1991 SCHOOL LEAVERS SURVEY

The 1991 School Leavers Survey made use of 1986-90 Family Allowance files as a sampling frame to select a stratified random sample of 18,000 18-20 year olds. Sampled individuals were contacted by telephone from April to June of 1991. The survey achieved a contact rate of 60% traced and of those 88% were interviewed producing about 9,500 useable responses. A wide range of questions were included in the interview – dealing with demographic background, school experience, as well as post-school labour market and other outcome measures. We focus only on those questions that can aid in reconstructing each respondent's experience of the secondary school system.

1) Basic Demographics - Sex, Year & Month of Birth, Month of Interview and Province of Study
2) Current Status - Graduate / Leaver (Dropout) / Continuer (Still in School)
3) Has respondent ever dropped out of high school?
4) What was the last month and year you went to high school, junior high or elementary school?
5) About how many months passed between the first and last time you dropped out?

Table 2: Survey Response Categories, Modeled Events and Representation of Event Histories

Modeled Event **Response Category**	Graduation Never DO GND	1st Dropout DO1	Return to School RTS	Graduation Ever DO GDO	2nd Dropout DO2
Continuer Never DO	RC	RC	RC	RC	RC
In 1st DO Spell	--	UC	IC or RC	RC	RC
Graduate Never DO	UC	--	--	--	--
Continuer Ever DO	--	Jointly IC		RC	RC
Graduate Ever DO	--	Jointly IC		UC	--
In Last DO Spell	--	--	--	--	UC
Unobserved DO Graduation	--	UC	Jointly IC		--

RC - right censored UC – uncensored IC - interval censored

Using these questions, we are able to identify precisely or to bracket the age at which dropout, return to school or graduation events occurred. The reported intervals between dropout events tended to be short (i.e., <6 months), so we have assumed that at most two dropout events can occur in our model. In addition, we have assumed a minimum age of 9 for the earliest possible dropout event (i.e., the youngest such age reported in any province). Table 2 displays the relations among seven categories of response and five events classified by their contribution to the representation of event histories.

Not all questions were asked of all respondents; for instance, the timing of dropout events for continuers was not reported precisely (e.g., 'minimum age < age(DO) & Age(Return) < interview' -- dropout and return events were jointly interval censored). The response category 'Unobserved DO Graduation' was identified from responses combining a reported dropout spell, a successful graduation and an age at leaving below the usual provincial age. Since, the ages at leaving were often below 15, we could reasonably assume that the reported age referred to a dropout event. Evidently, these responses represent completion of secondary school requirements outside of conventional secondary schools (e.g., in 'storefront schools'). Since our model has to do with the timing of labour force entry and/or eligibility for post-secondary education, we do not need to distinguish among means of satisfying secondary school requirements (i.e., these graduations are not treated as a distinct event type).

3. MODEL STRUCTURE

3.1 Event Hazards and Monthly Transitions

The model's initial state is 'in school' (i.e., Continuer Never DO - CNDO), it is assumed that the whole population is in this state as of their 9th birthday. Table 2 specifies the categories of event that form the basis of the model. Graduation and DO2 represent terminal states that most of the population is expected to reach by their 22nd birthday (some respondents had attained age 21 by the interview). Modeled transitions correspond to the hazards specified in Table 3. It is an important feature of the model that return to school is permitted, but thereafter special graduation and DO hazards are applied (i.e., experience dependent hazards).

Table 3: Model Hazard Matrix ($\mathbf{h} = [{}_e\mathbf{h}]$)

	CNDO	GND	DO1	RTS	GDO	DO2
CNDO	0	${}_{GND}h$	${}_{DO1}h$	0	0	0
GND	0	0	0	0	0	0
DO1	0	0	0	${}_{RTS}h$	0	0
RTS	0	0	0	0	${}_{GDO}h$	${}_{DO2}h$
GDO	0	0	0	0	0	0
DO2	0	0	0	0	0	0

Given the initial state, the proportions of the population in each of the six states at any given age (**P(a)**) are determined by the product of monthly transition matrices. Each transition matrix is calculated using the current age-specific hazard matrix (i.e., by a matrix exponential of **h** minus a diagonal matrix containing the row sums of **h** – expressed, with unit column vector **u**, as **diag (h*u)** – see, for example, Hoem and Funck-Jenson (1982)).

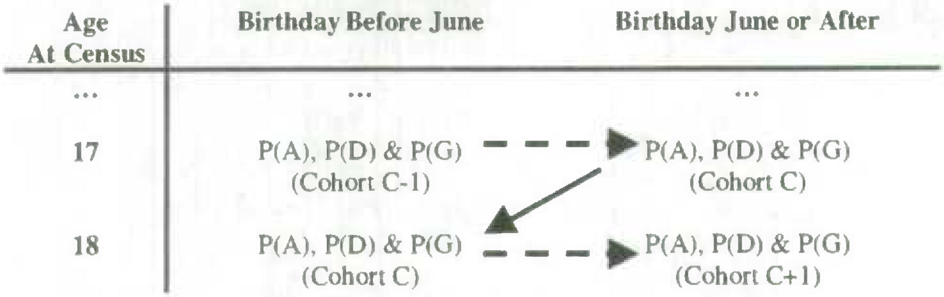
$$P(a'') = P(a') \prod_{a}^{a''} \exp\left[h(a) - \mathrm{diag}\left[h(a) * u \right] \right] \qquad (1)$$

Given the hazards, probabilities of occupying a given state or of making a transition between any pair of states can be calculated for any month in the interval between the 9th and 22nd birthdays.

3.2 Age Categories – Auxiliary 1991 Census Data

In the absence of grade progression data, age serves as a proxy. 1991 Census tabulations provide evidence that grouping by age at December 31 forms appropriate age categories (i.e., providing a good proxy for a 'school entry cohort').

Figure 1: Paired Comparisons - Between and Within School Entry Cohorts from Census: Proportions (within each cell) of Attenders (A), Dropouts (D) and Graduates (G)



Age At Census	Birthday Before June	Birthday June or After
...	...	...
17	P(A), P(D) & P(G) (Cohort C-1)	P(A), P(D) & P(G) (Cohort C)
18	P(A), P(D) & P(G) (Cohort C)	P(A), P(D) & P(G) (Cohort C+1)

The census tabulations consist of the proportions of secondary school graduates, secondary school attenders (i.e. not graduates) and dropouts (i.e., not graduates and not attending in the previous nine months) classified by age and month of birth. The proportions sum to 1.0 within each cell; so, for example, P(G) corresponds roughly to a cumulative graduation rate. Paired comparisons Between and Within school entry cohorts may be formed in the manner indicated schematically in Figure 1.

The proportions of graduates progressively increase with age, with a corresponding progressive decline in the proportions of attenders. Thus, successive older-younger differences in graduate proportions give an indication of dissimilarity in cumulative graduation rates Between Cohorts (horizontally) and Within Cohorts (diagonally). An example of these differences (Ontario Males) is displayed in Figure 2. The disparity between Within Cohort and Between Cohort curves is sufficient to motivate forming hazard age groups based on age at year-end. Evidently, the groups compared Between Cohort tend to differ in school grade, as well as age. Smaller, but significant Within Cohort differences remain, which might be explained by earlier school entry for some in the older group and/or by higher failure rates within the younger group.

Figure 2: Differences (Older-Younger) in Proportions of Graduates -- Ontario Males with 95% Confidence Intervals

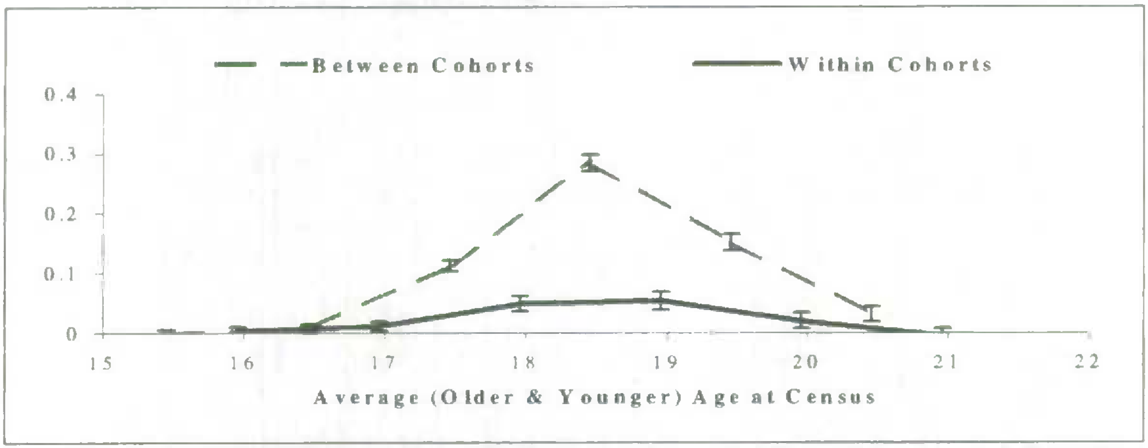
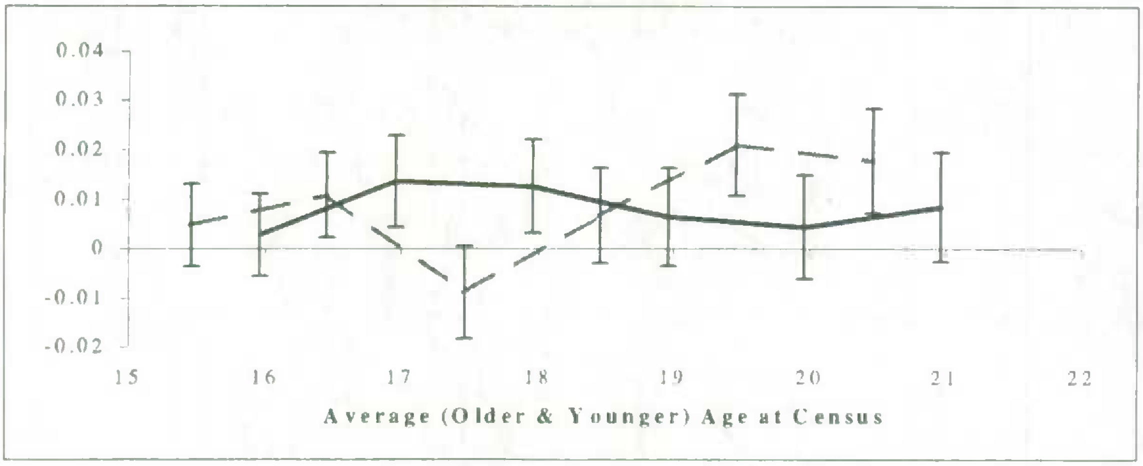


Figure 3 displays corresponding results for Dropout proportions. In this case however, the effects of age differences Within Cohorts appear to be as strong as the effects of age differences Between Cohorts. This seems to indicate that -- all else being equal -- the older members of a grade/class are more likely to dropout than younger members. The magnitude of these effects (1-2%) understates the differential that might be expected in dropout rates, because the census data represents net flows of only relatively long-term dropouts (9 or more months).

The Ontario Male results are qualitatively consistent with findings for males and females in all provinces. However, the magnitudes of Between and Within cohort differentials are smaller for Quebec than in other provinces, perhaps because of the lower graduating grade.

Figure 3: Differences (Older-Younger) in Proportions of Dropouts -- Ontario Males with 95% Confidence Intervals

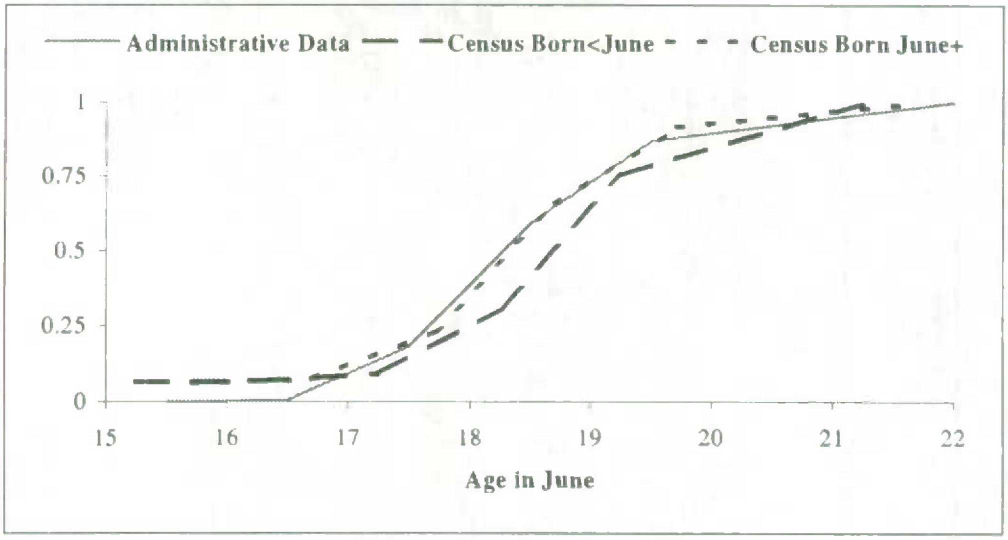


3.3 Seasonality – Auxiliary Administrative Data

The timing of graduation is critical to the larger microsimulation model given the September start of many post-secondary courses. The relative importance of June graduation is highlighted in Figure 4. We compare a cumulative distribution from administrative data on ages at graduation to corresponding distributions derived from census data. It might be expected that the administrative distribution would coincide with the average of the two 'school entry cohorts' derived from census data. Instead, there is a discrepancy (equivalent to about one school grade), which we assume is accounted for by differences in the reporting period for the two data sources. The census reporting period is effectively May, while the administrative data represents certification of graduates as of the end of June.

Results similar to those in Figure 4 may be obtained for males and females in all provinces. Note that there appears to be a positive bias in the census data at ages 15-16, perhaps due to self reported 'graduate equivalent' standing. Removing the bias would increase the discrepancy between the data sources, however the magnitude of the bias at ages 17+ is unknown.

Figure 4: Cumulative Distribution of Age at Graduation – Ontario Males
1990-91 Administrative Data and 1991 Census Data



3.4 Hazard Specifications

The model hazards by event type (${}_e\mathbf{h}$) have been configured, using the results from sections 3.2 and 3.3, as outlined below:

Dimension	Classification
s - Sex	**Female, Male**
p – Province	**Newfoundland, ... , British Columbia**
t – Season	**June, Summer, School Year**
m – month of birth	**January=1, ..., December=12**
a – age at year end	**<15, 15, ..., 19, 20+**

Month of birth effects are specified in terms of relative risks that are linear in month as follows:

$$e^{h}_{December} \Big/ e^{h}_{January} = 2 \left(e^{h}_{June} \Big/ e^{h}_{January} \right) \qquad (2)$$

4. MODEL ESTIMATION

4.1 Survey Weights

Survey weights have been post-stratified using 1991 Census data taking into account age, sex, province of residence and school attendance status (secondary or post-secondary). We have not been able to resolve the problems that migration poses. Our principal concern is with patterns of progression through each province's school system (i.e., province of study). Thus, post-dropout/graduation migration will introduce biases of unknown magnitude.

4.2 Penalized Likelihood

We have made use of a complicated, but conservative, fitting criterion that was selected to compensate for problems anticipated with samples as small as that of Prince Edward Island. Using (1), we are able to compute an unconditional probability of occupying any of the states (${}_s\mathbf{P(a)}$) and unconditional probabilities of a transition between pairs of states ($\mathbf{f_{ij}(a)}$ • ${}_e\mathbf{f(a)}$: event probability). By replicating these calculations 12 times, defining each replicate series as starting in a different month of the 9th birthday, we are able to account for all possible seasonal patterns. Throughout, months of birth are assigned equal weight when averaged.

Following Green (1987), the fitting criterion combines terms representing lack-of-fit (LOF) and a roughness penalty for the fitted values (PEN). A generalized cross-validation criterion is defined by a weighted (•) combination of these terms - **[LOF + • PEN]/RDF**^{2} (RDF is residual degrees of freedom). Increasing • will result in increases to both LOF and RDF and decreases in PEN. The optimal combination balances lack-of-fit and smoothness.

The roughness penalty has five components, each representing the sums of squares of 2nd differences (•^{2}) of probabilities for an event (logarithms of probabilities aggregated to integer ages):

$$\text{PEN} = \sum_{e} \lambda_{e} \left\| \Delta^{2} \ln\Big(\sum_{a \in \text{int}(a)} {}_{e}\hat{f}(a) \Big) \right\|^{2} \qquad (3)$$

The principal function of the roughness penalty is to prevent the estimation algorithm from being trapped by a local singularity in the likelihood. Penalized estimates correspond approximately to a smoothing spline. The penalty was assessed over integer age to prevent conflict between smoothness and seasonal effects.

Lack-of-fit has three components (**LOF = -2ln(L) + D(C, •) + D(A,Â)**) – outlined as follows:

i. –2*Log Likelihood of Survey Response (**-2*ln(L)**) computed for each event and accounting for each type of censoring.

$$L = \prod_{i \in UC} {}_e\hat{f}^{\,w_i^*} \prod_{j \in IC} \left(\sum_a {}_e\hat{f}\right)^{w_j^*} \prod_{k \in RC} {}_S\hat{P}^{\,w_k^*} \qquad (4)$$

where W_I^* represents the survey weight associated with the 1^{th} event divided by the average weight.

ii. Distance from the 1991 Census tabulations which underlie Figures 2 & 3: fitted proportions are grouped by age and state type (**s•**) and averaged over month of birth within the September-May interval -- to match census concepts (i.e., in terms of person-months by category). ${}_C N$ represents the approximate sample size underlying the census tabulation.

$$D(C,\ \bar{C}) = 4 \sum_a {}_C N_a \sum_{s'} \left(\sqrt{{}_{s'}P(a)} - \sqrt{{}_{s'}\hat{P}(a)}\right)^2 \qquad (5)$$

iii. Distance from the administrative data which underlies Figure 4: graduation probabilities in June by age (aggregated over types of graduate and normalized - ${}_G f^*$). ${}_A N$ represents the number of graduates underlying the administrative data.

$$D(A,\ \hat{A}) = 4\ {}_A N \sum_a \left(\sqrt{{}_G f^*(a)} - \sqrt{{}_G \hat{f}^*(a)}\right)^2 \qquad (6)$$

Simpson (1987) shows that the form of distance employed in (5) & (6) is, approximately, commensurate with **-2*ln(L)** (4), and has robustness properties that may limit the influence of (apparently) biased observations, such as the 15 year old census graduates.

Finally, estimation was subject to a number of explicit constraints:

(a) Hazards are less than 1.0	**$0 \bullet {}_e h < 1$**
(b) Graduation at specified months	**${}_G h(a) > 0$ at June, August & December**
(c) No Return at Graduation months	**${}_R h(a) = 0$ at June, August & December**
(d) No graduation occurs before age 15	**${}_{GND} h(a)$ & ${}_{GDO} h(a) = 0$, at age < 15**
(e) Graduate age distributions	**${}_{GND} F^* \bullet {}_{GDO} F^*$**
(f) Relative Risks	**$\bullet \bullet 0$**

Optimization employed a modified quasi-newton algorithm implemented in GAUSS 3.2. Starting values were obtained from a few thousand steps of the Metropolis-Hastings algorithm.

5. RESULTS

5.1 Validation

As may be seen in the charts below, the model for Ontario Males reproduces the timing of graduation represented in the administrative data (Figure 5), but does not appear to do as well with the timing implied by census data (Figure 6). The difficulty seems to involve conflict between the administrative data and the census, rather than lack-of-fit, per se. Lack-of-fit with the census would be improved by a bias adjustment to the census data at ages <20. Nevertheless, the fitted values agree with census data in terms of both the month of birth differences and the eventual proportion graduating. Figure 7 indicates a good fit of census data on dropouts, except perhaps at ages 16 and below. The fit for census data on attenders is essentially the complement of that for graduates.

5.2 Provincial Comparisons

Having fit a complete set of models, it is possible to derive comparisons among provinces that would not otherwise be direct. Tables 4 and 5 illustrate such comparisons, focusing on the eventual graduation of dropouts. It appears that, as of age 22, most dropouts have not returned to complete secondary school. Note that the percentages in Table 4 are considerably higher than corresponding percentages for 20-year-olds from the 1991 School Leavers Survey. They are, however, more in line with the 25% (nationally) of respondents who had been dropouts as of the 1991 survey and who reported having graduated by the time of the 1995 School Leavers Follow-up Survey (ages 22-24).

Table 4: **Model Percent of Ever Dropouts Graduating by Age 22**

	NFLD	PEI	NS	NB	QUE	ONT	MAN	SASK	ALTA	BC
Males	16.9	40.1	36.9	45.8	38.9	33.6	33.6	16.4	14.5	72.8
Females	20.7	40.4	21.2	25.0	54.9	55.9	21.8	18.2	34.5	37.5

It may not be surprising that dropouts who eventually graduate appear to have returned to school and resumed their education promptly. Gilbert et.al. (1993) report that dropouts are more likely to have failed a grade than are others. That being the case, if dropout graduates complete their requirements about one year later than others (Table 5), then much of the additional time might be accounted for by an extra year spent repeating a grade. Dropout graduates may have spent little time out of school.

Table 5: **Model Average Difference in Age at Graduation**
Ever Dropout – Never Dropout

	NFLD	PEI	NS	NB	QUE	ONT	MAN	SASK	ALTA	BC
Males	0.9	1.8	2.1	0.9	1.8	2.1	1.9	2.1	2.8	1.3
Females	3.0	2.3	2.1	2.3	1.6	1.8	1.4	1.8	2.5	1.2

6. CONCLUSIONS

The model provides estimates that represent a compromise between 'reconstructed event histories' and cross-sectional data. Implemented as a component of the larger microsimulation model, it generates realistic secondary school progression patterns. It further serves to highlight some features of secondary school progression that may often be overlooked – notably: the ages of school entry cohorts, seasonality and differences in experience by month of birth. Nevertheless, the deficiencies of the model and gaps revealed in the data serve to emphasize the value that prospective data on a school entry cohort would have. Failing that, more detailed retrospective event histories might be collected.

REFERENCES

Gilbert, Sid, Barr, Lynn, Clark, Warren, Blue, Matthew, Sunter, Deborah and Devereaux, Mary Sue (1993), *Leaving School: Results from a national survey comparing schoolleavers and high school graduates 18 to 20 years of age,* LM-294-07-93E, Statistics Canada, Ottawa.

Green, Peter J. (1987), Penalized Likelihood for General Semi-parametric Regression Models, International Statistical Review, 55(3), 245-259.

Hoem, Jan M. and Funck-Jensen, Ulla (1982), Multistate Life Table Methodology: A Probabilist Critique in *Multidimensional Mathematical Demography,* Kenneth C. Land and Andrei Rogers (eds), New York: Academic Press.

Simpson, Douglas G. (1987), Minimum Hellinger Distance Estimation for the Analysis of Count Data. Journal of the American Statistical Association, 82, 802-807.

Figure 5: Ontario Males – Fitted Percent of Graduates by Age at Graduation

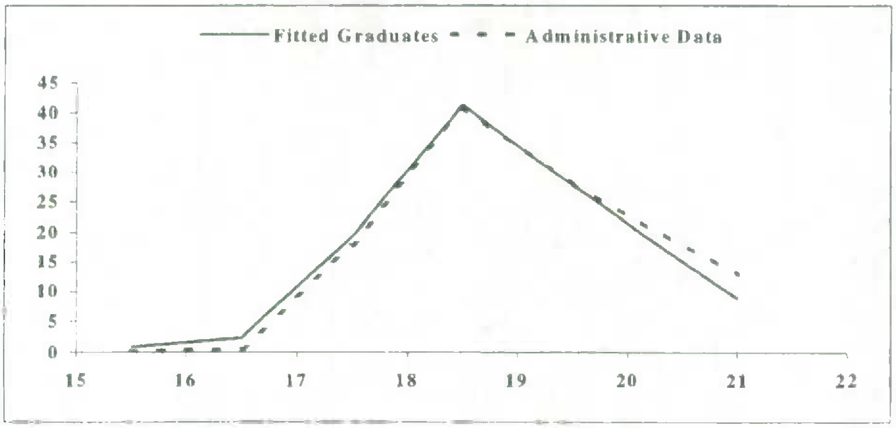


Figure 6: Ontario Males – Fitted Percent Graduates PYs by Age & Month of Birth (<June, June+)

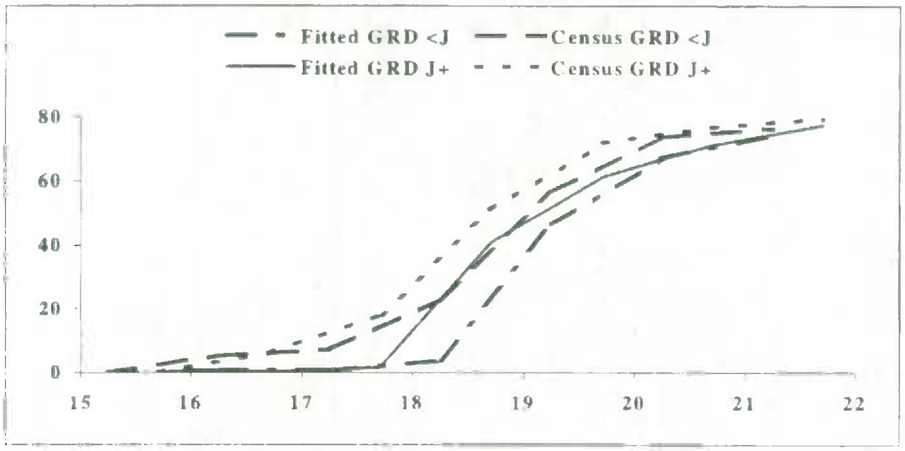
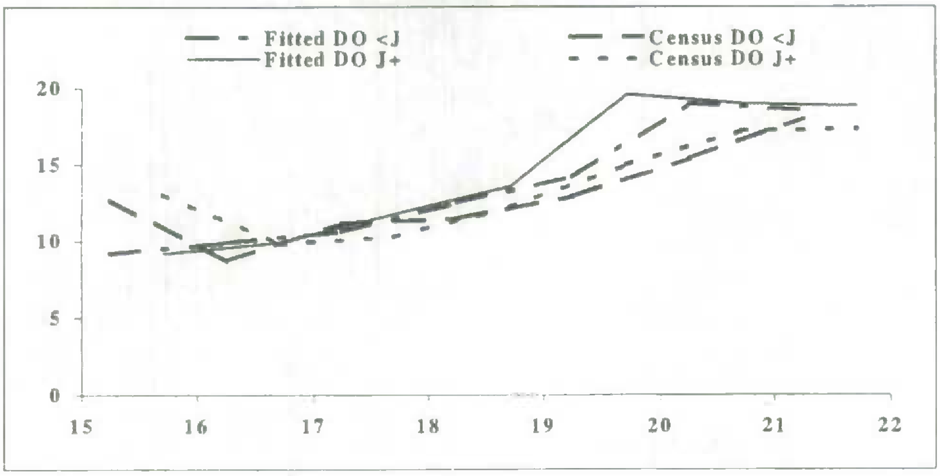


Figure 7: Ontario Males – Fitted Percent Dropouts PYs by Age & Month of Birth (<June, June+)



SESSION 8

MODELLING II

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

ESTIMATION WITH PARTIAL OVERLAP LONGITUDINAL SAMPLES

Wayne A. Fuller[1]

ABSTRACT

In a longitudinal survey conducted for *k* periods some units may be observed for less than *k* of the periods. Examples include, surveys designed with partially overlapping subsamples, a pure panel survey with nonresponse, and a panel survey supplemented with additional samples for some of the time periods. Estimators of the regression type are exhibited for such surveys. An application to special studies associated with the National Resources Inventory is discussed.

KEY WORDS: Survey sampling; Regression estimation; Panel surveys.

1. INTRODUCTION

Study of the dynamics of populations requires observations made at multiple time points on units in the population. We define a pure panel survey to be a survey in which the same units are observed at each time point of a survey conducted at more than one time point. A longitudinal survey is a survey conducted at more than one time point with some units observed at more than one time point, but the term is generally used for surveys conducted at more than two points in time with multiple observations on some units planned as part of the survey design. A rotation survey is one in which a unit is observed for a partial set of time points and is not observed for the remaining set of time points in the study. The Canadian Labor Force Survey and the United States Current Population Survey are examples of surveys designed to run continuously in which units rotate into the sample for a fixed period (or periods) and then permanently rotate out of the observation set.

There exist an array of designs combining individuals observed at some time points and individuals observed at all time points of the study set of time points. The simplest such design is a two-phase sample in which the observations at the second of two time points is a subsample of those observed at time one. The book edited by Kasprzyk et al. (1989) contains an excellent discussion of various aspects of panel surveys. Duncan and Kalton (1987) and Schreuder, Gregorie, and Wood (1993) discuss different types of repeated surveys and the objectives of such surveys. The largest fraction of the survey literature on repeated surveys has been devoted to rotating surveys. An early study is that of Jessen (1942), also see Cochran (1942). Patterson (1950) investigated estimation for rotating samples. Patterson's work was followed by a number of authors, including Eckler (1955), Rao and Graham (1964), Gurney and Daly (1965), Raj (1965), Smith (1978), Wolter (1979), Jones (1980), Huang and Ernst (1981), Breau and Ernst (1983), Kumar and Lee (1983), Singh (1996) and Yansaneh and Fuller (1998).

We review generalized least squares estimation for partial overlap surveys. We then compare some designs for studies directed toward longitudinal dynamics. Because such studies are multiple purpose, we cannot

[1] Fuller, Wayne A., Iowa State University, 218 Snedecor Hall, Statistical Laboratory, Ames, Iowa USA 50011-1210.

hope to develop a design that is optimal for all objectives. A part of our investigation will be to identify the trade-offs.

2. SUPPLEMENTED PANEL DESIGNS

To introduce generalized least squares estimation, consider a simple three period survey in which one-fourth of the units are observed on all three periods and each of three sets of one-fourth of the units is observed in exactly one of the three periods. Thus, if the total sample size is n , then 0.5 n of the units are observed at each time point. Let $(Y_1, Y_2, Y_3,)$ denote the value of a characteristic observed at times one, two, and three respectively. Assume that the correlation between observations at time i and time j on the same element is that of a first order autoregressive process with parameter ρ . Assume simple random sampling for the selection of all samples.

Let $(\bar{y}_{11}, \bar{y}_{21}, \bar{y}_{31})'$ denote the estimated mean at time one, two and three, of the sample elements that are observed all three periods. Let $(\bar{y}_{12}, \bar{y}_{23}, \bar{y}_{34})'$ denote the sample means for the three periods for the sample elements that are observed once. We call these six estimators the elementary estimators. See Gurney and Daly (1965). Let $\mu = (\mu_1, \mu_2, \mu_3)'$ denote the population means for the three periods. Then we can write

$$y = X\mu + e \tag{1}$$

where $y' = (\bar{y}_{11}, \bar{y}_{21}, \bar{y}_{31}, \bar{y}_{12}, \bar{y}_{23}, \bar{y}_{34})$, $X' = (I_3, I_3)$, the covariance matrix of e is

$$\Sigma_{ee} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 & 0 & 0 \\ \rho_2 & \rho_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \sigma^2 \tag{2}$$

and σ^2 is the variance of a mean of $n/4$ observations. It follows that the best linear unbiased estimator of μ using this amount of information is

$$\hat{\mu}_g = (X'\Sigma_{ee}^{-1} X)^{-1} X'\Sigma_{ee}^{-1} \bar{y} \tag{3}$$

and

$$V\{\hat{\mu}_g\} = (X'\Sigma_{ee}^{-1} X)^{-1} \tag{4}$$

We compare the covariance matrix (4) with the covariance matrix of a pure panel survey in which the same 0.5 n units are observed on all three periods. The covariance matrix for the pure panel design is

$$V\left\{\hat{\mu}_{\text{panel}}\right\} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} \frac{\sigma^2}{2} .$$

The comparison is for populations in which $\rho_1 = \rho$ and $\rho_2 = \rho^2$. Table 1 contains variances for several functions of means for ρ ranging from 0 to 0.99, relative to the corresponding variance for the pure panel design. If $\rho = 0$, the variance of the best linear unbiased estimator of the three means is the variance of the simple mean at each period. If the total number of elements is n , the variance at each time period is $2n^{-1}\sigma^2$. In the remaining discussion we assume $\sigma^2 = 1$. If observations on the same element are correlated ($\rho \neq 0$), the variance of the estimated means for the supplemented panel design is less than $2n^{-1}$. The limit of the correlation is one. This can occur for characteristics such as the age of an individual at a fixed point. The lower bound for the variance of the supplemented panel design at $\rho = 1$ is n^{-1} because this is the number of different individuals in the study. Correspondingly, the limit of the relative efficiency for period means of the supplemented panel, relative to the pure panel, is 2.0. The variances of the estimated means for the supplemented panel design for the first and last period are the same. The variance of the middle period is somewhat smaller because the middle observation has one period correlation with the first and third observations on the same element.

Table 1. Variances of Functions of the Estimators for Three Period Design With 50% New Observations at Each Time Relative to Variances of Pure Panel Design.

ρ	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_1 - \bar{y}_2$	$\bar{y}_1 - \bar{y}_3$
-0.70	0.838	0.755	0.674	1.253
-0.50	0.929	0.875	0.768	1.143
0	1.000	1.000	1.000	1.000
0.50	0.929	0.875	1.304	1.143
0.70	0.838	0.755	1.488	1.253
0.90	0.660	0.595	1.773	1.681
0.99	0.520	0.510	1.973	1.951

If the correlation is positive, the variance of the difference of two means is smaller for the pure panel survey than for the supplemented panel design. As ρ approaches one, the variance of the difference of two means approaches zero for both designs. Thus, for example, the variance of period-to-period change for the pure panel is only 2% of the variance of the mean if $\rho = 0.99$. The pure panel design has an efficiency approaching twice that of the supplemented panel design as ρ approaches one. This efficiency holds for any multiple period study.

We have outlined an estimation scheme for a vector of time means for a scalar *y*-characteristic, but there is no conceptual difficulty in extending the procedure to a vector of *y*-characteristics. However it is not operationally feasible to construct separate estimates for every variable in a study. For applications we suggest the procedure of Fuller (1990). The procedure has been implemented for the United States Current Population Survey. See Lent, Miller, and Cantwell (1996). In that procedure a set of variables from each time period is chosen as the control variables. Generalized least squares or a similar procedure is used to estimate the means (or totals) for the vector of control variables. A set of regression weights is then constructed for the panel portion of the sample using the control variables. An alternative method of constructing the weights for rotating samples has been suggested by Singh (1996).

Under certain conditions, the variance estimation procedure suggested by Fuller (1998) can be applied to estimate the variance of estimates. Under less restrictive conditions, the replication scheme in which replicates are drawn from the entire sampling procedure can be used. See Rao and Sitter (1995) and Sitter (1997).

3. THE NATIONAL RESOURCES INVENTORY

The National Resources Inventory is a survey of the nonfederal land area of the United States conducted by the Natural Resources Conservation Service of the United States Department of Agriculture. It is a large survey of about 300,000 primary sampling units. A primary sampling unit is a segment of land. In the Midwest the segment is 160 acres, but the primary sampling units vary across the country. In the west, there are some that are as big as 640 acres and in the east some segments are on the order of 100 acres. Within the primary sampling unit, points are designated for observation. There are either two or three points per primary sampling unit. The basic survey has been conducted in each of the years 1982, 1987, 1992, and processing is currently underway for 1997 data.

In 1995, a sample of 3,000 segments was selected from the 300,000 for a study that was called the Erosion Update Study. This study was a subsample of the large sample, but the primary sampling units were different. In 26 states, counties were sampling units and for 22 of the states, states were primary sampling units. The 1992 basic NRI sample of 300,000 segments and the 1995 subsample of 3,000 segments form a classical two-phase sample. In the first-phase sample of 300,000 sampling units vector X is observed. In the subsample of 3,000 units, an extended vector (X, Y) is observed.

A second special study was conducted in 1996. The original 1995 sample of 3,000 was augmented by another 1,000 segments to obtain a total of 4,000 segments. A third study was conducted in 1997 in which the 1996 sample was augmented by another 2,000 segments. Thus, we can divide the original 300,000 segments into four subgroups: 294,000 are observed only in 1992; 3,000 are observed in 1992, 1995, 1996, and 1997; 1,000 are observed in 1992, 1996, and 1997; and 2,000 are observed in 1992 and 1997. The vector of subgroup means is for a characteristic y for these data is

$$\bar{y}' = \left[\bar{y}_{92,294}, \bar{y}_{92,3}, \bar{y}_{95,3}, \bar{y}_{95,3}, \bar{y}_{97,3}, \bar{y}_{92,1}, \bar{y}_{96,1}, \bar{y}_{97,1}, \bar{y}_{92,2}, \bar{y}_{97,2}\right]$$

where $\bar{y}_{t,j}$ is the mean for the characteristic at time t on a group of $j(000)$ segments, and μ_t is the mean of y at time t.

The covariance structures for several characteristics observed in the National Resources Inventory have been estimated. See Breidt and Fuller (1998). We use those correlations to study the efficiency of generalized least squares. Under the model used to estimate the correlations, the observation on segment i at time t, denoted by y_{it}, is expressed as a sum of μ_t and e_{it}, where μ_t is the mean at time t and e_{it} is the deviation from the mean. The deviation is assumed to satisfy the model

$$\begin{aligned} e_{it} &= \alpha_{it} + \varepsilon_{it}, \\ \alpha_{it} &= \rho\alpha_{i,t-1} + u_{it}, \end{aligned}$$

where α_{it} is a stationary autoregressive process, $|\rho| < 1$ is the autoregressive parameter, ε_{it} is uncorrelated measurement error distributed with mean zero and variance σ_ε^2, denoted by $(0, \sigma_\varepsilon^2)$ and u_{it} is the $(0, \sigma_u^2)$ uncorrelated error of the autoregressive process. Under the model

$$V\{y_t\} = (1-\rho^2)^{-1}\sigma_u^2 + \sigma_\varepsilon^2 .$$

We let

$$\nu = \left[V\{y_t\}\right]^{-1}\sigma_\varepsilon^2 ,$$

where ν is the fraction of total variance that is due to measurement error. Under the model the correlation between observations h periods apart is

$$\mathrm{Corr}\left(y_{it}, y_{i,t+h}\right) = \rho^{|h|}(1-\nu), \qquad h = \pm 1, \pm 2, \ldots$$

The estimated parameters for several characteristics are given in Table 2.

Table 2. Estimated parameters of error model.

Characteristic	Parameter	
	ρ	ν
USLE	0.955	0.229
Cultivated cropland	0.981	0.014
Noncult. cropland	0.955	0.130
Pasture	0.974	0.031
Rangeland	0.992	0.011
Forest	0.981	0.015
Small built-up	0.957	0.029
Large urban	0.977	0.016
Streams	0.996	0.089
Roads	0.998	0.027

The USLE is the tons of soil loss calculated by the Universal Soil Loss Equation. The computations require a number of factors such as the slope of the land and the cropping practices. Hence, the variable has large relative measurement error with nearly 23 % of total variance attributable to measurement error. The remaining variables are the acres in that land use in the segment. Land uses that are relatively permanent, such as rangeland, streams, and roads have large autoregressive coefficients. There is considerable measurement error associated with the determination of the acres in streams.

One simple estimator of the mean for each period is the mean of segments actually observed during that period. Table 3 contains estimated efficiencies of generalized least squares relative to the simple mean. The generalized least squares estimator for 1992 is the simple mean because all other samples are subsamples of the 1992 sample. The estimator for 1997 is a two-phase estimator with the 1992 sample as the first phase sample. The estimates for 1995 and 1996 use information from more than one other sample.

The smallest efficiency gains are for USLE which has the smallest autocorrelations. The year-to-year correlation for USLE is about 0.74. On the other hand, generalized least squares is much superior to the simple mean for rangeland which has a year-to-year correlation of about 0.98.

The estimated efficiencies for changes are given in the last three columns of Table 3. The gains for changes are generally comparable to those for levels.

Table 3. Estimated relative efficiency of generalized least squares to sample means.

	Year			Change		
Characteristic	1995	1996	1997	$\mu_{92} - \mu_{95}$	$\mu_{95} - \mu_{96}$	$\mu_{96} - \mu_{97}$
USLE	3.67	1.90	1.59	3.70	1.67	1.42
Cultivated cropland	14.42	7.13	4.67	16.32	4.56	6.09
Noncult. cropland	8.29	3.91	2.53	8.76	2.97	3.85
Pasture	5.58	2.91	2.34	5.75	2.07	1.91
Rangeland	24.15	12.74	8.60	30.86	7.51	9.88
Forest	14.51	7.19	4.71	16.45	4.58	6.11
Small built-up	7.58	3.67	2.47	7.95	2.70	3.22
Large urban	12.63	6.18	4.05	14.00	4.06	5.36
Streams	9.83	5.38	4.65	10.59	2.96	2.70
Roads	24.34	13.86	11.13	31.21	6.86	7.04

REFERENCES

Breau, P. and Ernst, L. R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

Breidt, F. J. and Fuller, W. A. (1998). Comparison of alternative sample rotation designs for the National Resources Inventory. Unpublished manuscript. Iowa State University, Ames, Iowa.

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

Duncan, G. J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

Eckler, A. R. (1987). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-685.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

Fuller, W.A. (1997). Replication variance estimation for two-phase samples. Unpublished manuscript, Iowa State University, Ames, Iowa.

Gurney, M. and Daly, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.

Huang, L. R. and Ernst, L. R. (1981). Comparison of an alternative estimator to the current composite estimator in the current population survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 303-308.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

Jones, R. G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.

Kasprzyk, D., Duncan, G.J., Kalton, G., and Singh, M.P. (1989). *Panel Surveys*, Wiley, New York.

Kumar, S. and Lee, H. (1983). Evaluation of composite estimation for the Canadian Labor Force Survey. *Survey Methodology*, 9, 1-24.

Lent, J., Miller, S. M., and Cantwell, P. J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 130-139.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *The Journal of the Royal Statistical Society, Series B*, 12, 241-255.

Raj, D. (1965). On sampling over two occasions with probability proportionate to size. *Annals of Mathematical Statistics*, 36, 327-330.

Rao, J. N. K. and Graham, J. E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Schreuder, H.T., Gregorie, T.G., and Wood, G.B. (1993). Sampling Methods for Multi-Resource Forest Inventory. Wiley, New York.

Singh, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Smith, T. M. F. (1978). Principles and problems in the analysis of repeated surveys. Pages 201-216. In N. Krishnan Namboodiri, ed. *Survey Sampling and Measurement*. Academic Press, New York.

Wolter, K. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Yansaneh, I. S. and Fuller, W. A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*. To appear.

ACKNOWLEDGMENTS

This research was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Service and Iowa State University.

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

MULTIVARIATE LOGISTIC REGRESSION FOR DATA FROM COMPLEX SURVEYS

Paul-André Salamin[1]

ABSTRACT

Multivariate logistic regression, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is useful in the analysis of longitudinal data as it allows for dependent repeated observations of a categorical variable and for incomplete response profiles. We show how the method can be extended to deal with data from complex surveys and we illustrate it on data from the Swiss Labour Force Survey. The effect of the sampling weights on the parameter estimates and their standard errors is considered.

KEY WORDS: Multivariate binary data; Multivariate Logistic Regression; Complex Surveys.

1. INTRODUCTION

Some of the surveys conducted by the Swiss Federal Statistical Office (SFSO) are of a longitudinal nature, i.e. surveys where the units of a sample are observed repeatedly through time. In this type of surveys, the observations taken at different times are dependent. Also, some of the response profiles can be incomplete, either by design as in a rotating panel or due to non-response. When the observed response is categorical, we found the multivariate logistic model useful for analysing this type of data (Salamin 1997, 1998).

In Section 2 we indicate how multivariate logistic regression can be extended to deal with data from complex surveys. In Section 3 the method is illustrated on data from the Swiss Labour Force Survey. Finally, in Section 4 we present our conclusions and indicate directions for further work.

2. MULTIVARIATE LOGISTIC REGRESSION

The multivariate logistic model, introduced by Glonek and McCullagh (1995), can handle multivariate responses of either nominal or ordinal type and either discrete or continuous explanatory variables. Here we consider only multivariate binary responses and discrete predictors. The multivariate logistic model is an example of a generalised linear model, see McCullagh and Nelder (1989). Its link function, also called the multivariate logistic transformation, expresses the joint distribution of the response profiles in terms of marginal logits, marginal log odds ratios and contrasts of marginal log odds ratios.

We consider a binary variable observed at d occasions. We denote by y the vector whose components are the profile indicator functions, i.e. for any given response profile i, the corresponding component of the vector y is equal to 1 if the profile i has been observed, and to 0 otherwise. The vector of the probabilities of the response profiles is denoted by π. The multivariate logistic regression models are defined to be those of the form $\eta = X\,\beta$, where X is a matrix of explanatory variables, β is a vector of unknown parameters and η is the image of π under the multivariate logistic transformation. For one observation, the kernel of the log likelihood is given by

[1] Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Schwarztorstrasse 96, CH-3003 Berne, Switzerland

$$l(\beta; y) = y^T \log \pi(\beta)\,,$$

where $\pi(\beta)$ is the inverse of the link function. The score vector and the information matrix are given by

$$s(\beta; y, X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} y$$

and

$$I(\beta; X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} D\pi(\beta)\,,$$

where $D\pi(\beta)$ is the Jacobian matrix of $\pi(\beta)$.

The extension of multivariate logistic regression to the case of a sample drawn from a finite population is a straightforward application of the results of Binder (1983). We consider a finite population U of size N and we assume that a vector of profile indicator variables and a design matrix are associated to each element of U. Let S be a probability sample of size n drawn from the population U. We denote by w the weights associated to S. We define the population score vector as

$$s_U(\beta) = \sum\nolimits_{k\in U} s(\beta; y_k, X_k)\,.$$

For any value of β, the population score vector is a vector of totals. Using the sampling weights, we can then estimate this population score vector from the sample. Following Binder (1983), we assume that we can compute a consistent estimator of the covariance matrix, under the sampling design, of this estimated population score vector. The estimated population score vector can then be used to define an estimator of β as the solution of the equation

$$s_w(\hat{\beta}) = \sum\nolimits_{k\in S} w_k s(\beta; y_k, X_k) = 0\,.$$

An estimator of the covariance matrix, under the sampling design, of the estimator of β is given by

$$\operatorname{cov}_w(\hat{\beta}) = I_w(\hat{\beta})^{-1} \hat{\operatorname{cov}}_w(s_w(\hat{\beta})) I_w(\hat{\beta})^{-1}\,, \tag{1}$$

where

$$I_w(\beta) = -\sum\nolimits_{k\in S} w_k I(\beta; X_k) \tag{2}$$

is an estimator of the population information matrix.

3. ILLUSTRATION

3.1 Description of the data

A detailed description of the sampling design and weighting procedure for the Swiss Labour Force Survey (SLFS) can be found in Hulliger *et al.* (1997). Here we just recall some of the relevant aspects of this survey. The SLFS collects information on employment of resident persons of age 15 or more in Switzerland. Starting in the second quarter of 1991, a sample of about 16000 persons are interviewed each year. The survey is designed as a rotating panel with a time-in-sample of 5 years. During the start-up phase, i.e. from 1992 to 1996, approximately one fifth of the original sample was rotated out each year and replaced by a renewal sample. The units in the renewal samples then stay in the panel for a full period of 5 years. The sample was drawn according to a two-stage stratified design, with households as primary units and persons as secondary units. The 42 strata in the first stage are defined in term of regions and type of communes. The sample is allocated proportionally to the size of the strata, with the exception of some regions that were over-sampled. In the second stage, a target person is selected with equal probabilities among the eligible persons of the household. The final weights are obtained from the inverse of the inclusion probabilities after adjustment for non-response and calibration on some known population characteristics. The "parallel group" jackknife is used for variance estimation. With this type of jackknife, the strata are first divided into G groups of equal size and the jackknife replicates are obtained by deleting in turn one group of observations from all the strata. The variable of interest is the employment status. This is a nominal variable with three categories defined as "employed", "unemployed" and "out of the labour force".

3.2 Example

In this example we use the observations of the employment status, for the years 1992 to 1995, obtained from the individuals in the sample of the canton of Vaud. Here, we define the employment status as a binary variable taking the value 1 if an individual is employed and 2 if an individual is unemployed or out of the labour force. For simplicity, we have also eliminated the individuals that dropped out of the panel. An attractive feature of multivariate logistic regression, however, is the possibility of incorporating into a single model the different types of response profiles arising from the rotating panel. Thus the individuals which dropped out of the panel could also have been incorporated into the analysis.

In what follows, the emphasis is on the impact of the sampling weights. In particular, it is of interest to check if the weights, which were developed primarily for cross-sectional estimation, change the estimates of the parameters of multivariate logistic models, and to quantify the increase of the variability of the parameter estimates due to the sampling weights. In order to facilitate the comparison with estimates based on the unweighted data, the sampling weights are normalised to sum to the sample size.

We consider a model in which we have one parameter for each of the marginal logits and for each of the log odds ratios. The parameters of order 3 and 4 are set to 0. The parameter estimates and their standard errors are given in Table 1. For the weighted data, the estimate of the standard error can be computed either ignoring the sampling weights, in which case it is denoted by *SD*, or taking the sampling weights into account, in which case it is denoted by *SD.w*. Note that *SD* is computed as the square root of the diagonal of the inverse of the estimated population information matrix (2), and that *SD.w* is based on (1).

Table 1: Parameter estimates and standard errors

Parameter	Unweighted data		Weighted data		
	β	*SD*	β	*SD*	*SD.w*
logit 92	0.6362	0.0352	0.6223	0.0356	0.0407
logit 93	0.5573	0.0338	0.5720	0.0342	0.0389
logit 94	0.5411	0.0325	0.5574	0.0329	0.0356
logit 95	0.4716	0.0320	0.4984	0.0323	0.0365
log OR 92 93	4.2579	0.1465	4.0113	0.1394	0.1647
log OR 93 94	4.1111	0.1310	3.9258	0.1269	0.1484
log OR 94 95	4.5561	0.1389	4.3527	0.1334	0.1549
log OR 92 94	3.8372	0.1442	3.5633	0.1369	0.1654
log OR 93 95	3.7913	0.1334	3.5891	0.1290	0.1527
log OR 92 95	3.5774	0.1530	3.2882	0.1447	0.1643

Overall, we notice that the parameter estimates are not strongly affected by the weights: the differences between the estimates based on the unweighted or weighted data are within twice their standard errors. This is due to the circumstance that, at least for the example considered here, the SLFS weights do not distort the profile counts by much. Comparing *SD* and *SD.w* for the weighted data, we can assess the effect of the weights on the variability of the parameter estimates. We observe a fairly consistent increase of about 15%. As in the SLFS only one person per household is selected, we are not expecting a large cluster effect and the moderate increase in the variability of the parameter estimates is plausible. Also, it is consistent with values obtained in similar surveys at the SFSO.

4. CONCLUSION

We have proposed an extension of multivariate logistic regression to data from complex surveys and demonstrated its use on data from the SLFS. As far as the computation of point estimates is concerned, the standard method needs only minor adjustments. The estimation of the variance of the point estimates requires the estimation of the variance of a total under the sampling design. This can be complicated. For example, in the SLFS this is done by the parallel group jackknife, a fairly computer intensive procedure. However, for the data that we examined, we found a fairly consistent increase of the variability due to the sampling weights. This can be used to correct results from the standard analysis, which are easier to obtain.

Although the results we obtained seem reasonable, the validity of the proposed method rests on asymptotic arguments and it would be desirable to investigate the properties of the method more thoroughly.

REFERENCES

Binder, D.A. (1983). On the Variance of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-292.

Glonek, G.F.V., and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, 57, 533-546.

Hulliger, B., Ries, A., Comment, T., and Bender, A. (1997). Weighting the Swiss Labour Force Survey. In: *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switzerland,* C. Malaguerra, S. Morgenthaler and E. Ronchetti, Eds. Basel: Birkhäuser Verlag.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, (2nd ed.), London: Chapman and Hall.

Salamin, P.-A. (1997). Longitudinal Analysis of Swiss Labour Force Data. In: *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switzerland,* C. Malaguerra, S. Morgenthaler and E. Ronchetti, Eds. Basel: Birkhäuser Verlag.

Salamin, P.-A. (1998). Longitudinal Analysis of Swiss Labour Force Data By Multivariate Logistic Regression. *Survey Methodology* (submitted for publication).

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

PARAMETER ESTIMATION FOR A FINITE MIXTURE OF DISTRIBUTIONS FOR DICHOTOMOUS LONGITUDINAL DATA: COMPARING ALGORITHMS

J.-F. Beaumont and A. Demnati[1]

ABSTRACT

For longitudinal data, mixed models are often used, since they allow analysts to take account of the correlation between different observations from the same individual. The finite mixture model may be considered as a special case of a mixed model. In this document, attention will be given to the maximum likelihood method. The maximization of the likelihood function for a finite mixture of distributions is generally more difficult than in the usual case of a single distribution and can require considerable time. The objective of this project was therefore primarily to identify the one or more algorithms that best meet the criteria of run time and of efficiency in finding the solution. To achieve this objective, a simulation study was carried out. Only the situation in which the dependent variable is dichotomous was considered. This situation is very useful in practice, since among other things it can be used to model discrete durations, such as the length of time in "low income" status.

KEY WORDS: Finite mixture of distributions; Mixed model; EM algorithm; Newton-Raphson; Logistic model.

1. INTRODUCTION

For longitudinal data, mixed models (Zeger, Liang and Albert, 1988) are often used, since they allow analysts to take account of the correlation between different observations from the same individual. These models are said to be mixed because they include both fixed and random parameters (or effects). The finite mixture model may be considered as a particular case of a mixed model. The only difference lies in the way of specifying the distribution of the random parameters (Beaumont and Demnati, 1998). For a mixture, the random parameters are discrete and do not come from a known parametric distribution such as a normal distribution.

Models in which it is assumed that individuals come from a finite mixture of distributions are becoming increasingly common, and they provide a useful and attractive tool for analysts. For longitudinal data, these models allow for taking account of the correlation of observations over time. There are other factors that may encourage the use of such models. Among other things, they can make it possible to take account of the effect of a missing variable, since categories of that variable that could not be observed may be associated with the distributions that make up the mixture. These models may also be used for the purpose of clustering (McLachlan and Basford, 1988).

This paper will focus solely on the maximum likelihood method. Furthermore, only the situation in which the dependent variable is dichotomous will be considered. This situation is very practical to consider, since among other things, it can be used to model discrete durations, such as the length of time in "low income" status.

In Section 2, an example of an application is briefly presented. Section 3 serves primarily to show that the likelihood function for a finite mixture of distributions is flatter than in the usual case of a single distribution.

[1] J.-F. Beaumont and A. Demnati, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

In Section 4, different algorithms for maximizing this likelihood function (or its natural logarithm) are examined. Section 5 describes and presents the findings of a simulation study for comparing algorithms in terms of total run time (related to the number of iterations to converge) and efficiency in finding the solution. The last section offers a brief conclusion.

2. EXAMPLE

In this section, an example of an application is briefly presented using the longitudinal administrative databank (LAD) of the Small Area and Administrative Data Division of Statistics Canada. That databank is obtained from Revenue Canada and is designed on the basis of individuals' income tax returns. In fact, LAD is a Bernoulli sample in which each return has one chance in ten of being selected. It currently covers the years 1982 to 1995.

In this example, we are interested in modelling the length of time in "low income" status. With LAD, the data are available only on an annual discrete basis. The mixed logistic model (Zeger, Liang and Albert, 1988) therefore seems appropriate. This model allows us to take account of the correlation between episodes in "low income" status for a same individual as well as the correlation between the observations of a single episode. Furthermore, the model allows for the use of independent variables that vary over time. For example, duration is incorporated into the model by means of independent dummy variables.

In this application, we assumed a model in which individuals come from a mixture of two distributions and in which only the intercept is random. The estimates were obtained by the maximum likelihood method. It was found that the probability of moving out of "low income" status depends on a number of factors, one of which is obviously length of time. In addition, there is strong evidence against the hypothesis that individuals come from a single distribution, meaning that there are likely at least two groups of individuals in the population. These groups may correspond to categories of a variable that we were unable to observe. With LAD, for example, we have no information concerning individuals' education level. If it is believed that this variable affects the probability of moving out of "low income" status, the unobserved variable could be interpreted as being education level. The unobserved variable could also be interpreted as being each individual's motivation to move out of "low income" status. The interpretation of this variable is therefore difficult and should not be taken lightly.

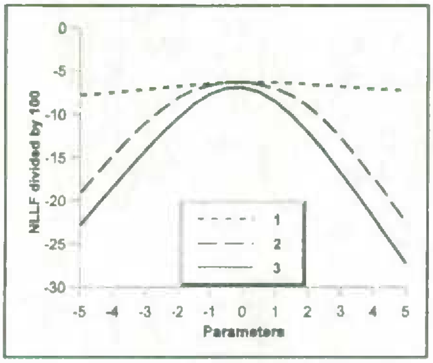
3. LIKELIHOOD FUNCTION

In what follows, it will be assumed that the individuals come from a superpopulation which is a finite mixture of distributions and that the sampling design is ignorable. The likelihood function is given in Beaumont and Demnati (1998), among other sources. There are a number of factors that make this function more difficult to maximize than the one for the usual case of a single distribution. One such factor is that this function has frequently more than one maximum, which in some cases makes it ineffective to use the commonly used algorithms, such as Newton-Raphson. Furthermore, the likelihood function for a mixture is generally flatter than in the case of a single distribution. To show this phenomenon, we simulated a data set in which the dependent variable is dichotomous and in which there is no independent variable except for the intercept. The data set contains 100 individuals with 10 observations per individual, and the data come from a mixture of two distributions. We then compared the natural logarithm of the likelihood function (NLLF) for the following three cases: (1) We choose the mixed logistic model and assume a mixture of two distributions. We then estimate the parameter associated with the distribution of the random effect and consider known the other parameters. (2) We choose the mixed logistic model and assume a mixture of two distributions. We then estimate the intercept (fixed parameter) and consider known the other parameters. (3) We choose the usual logistic model (a single distribution) and estimate the intercept.

From examining Figure 1, it may easily be seen that the likelihood functionis flatter if we assume a mixture of two distributions (cases 1 and 2) than if we assume a single distribution (case 3). This phenomenon is even more pronounced when we attempt to estimate the parameter associated with the distribution of the random effect (case 1). Thus, when we assume a finite mixture of distributions, the convergence criterion must be sufficiently strict, since it is generally based on the relative change in the NLLF between two successive iterations. If this criterion is not adequate, the convergence might appear to be achieved, whereas the solution would still be relatively far off.

It should lastly be noted that in practice, all the algorithms that will be presented in the next section are generally good for finding the global maximum in such simple cases (a single parameter to be estimated and the two distributions relatively well-separated).

Figure 1: Comparison of NLLF divided by 100 for cases 1, 2 and 3



4. ALGORITHMS

The algorithms that will be considered for maximizing the likelihood function may be grouped into two categories: algorithms of the Newton-Raphson type, described in Section 4.1, and algorithms of the EM (expectation-maximization) type, described in Section 4.2. In Section 4.3, we will discuss another way to approach the maximization problem, which consists in combining two algorithms. For a more detailed description of these algorithms, the reader is referred to Beaumont and Demnati (1998).

Each of the algorithms was programmed using the NLIN procedure of SAS (SAS Institute Inc., 1990), which is designed to minimize a sum of squares. Thus they will be easier to compare in terms of run time. Furthermore, to ensure that the NLLF increases with each iteration, this procedure performs sub-iterations. A sub-iteration consists in reducing by half the change in a given iteration. These sub-iterations are carried out so long as the NLLF does not increase and the maximum number of sub-iterations has not been reached.

4.1 Algorithms of the Newton-Raphson type

With this type of algorithms, the maximization of the NLLF is obtained by solving the system of non-linear equations resulting from the first derivatives of the NLLF with respect to each of the parameters to be estimated. The Newton-Raphson (NR) algorithm is well-known and is often used to solve such a system of equations. It calls for calculating the first and second derivatives of the NLLF. This algorithm converges quickly if the initial values are not too far from the solution. However, it very often becomes incapable of finding the global maximum of the likelihood function when the initial values are not sufficiently close to the solution. In addition, calculation of the second derivatives can take considerable computer time, especially when the model considered contains many parameters or when there are a large number of observations.

If providing the second derivatives is impractical or if calculating them involves a great deal of run time, they can be approximated using the first derivatives (Beaumont and Demnati, 1998). The algorithm for this is called NRH1.

4.2 EM-type algorithms

The EM algorithm, developed by Dempster, Laird and Rubin (1977), serves to maximize a likelihood function where the data are incomplete. An iteration of this algorithm consists in (i) calculating the expectation of the NLLF for the complete data, conditional on the values observed and the current values of the parameters, and (ii) maximizing this expectation. For a finite mixture of distributions, the complete data are constituted by the observed values and the unobserved variable. An important feature of the EM algorithm is that the NLLF does not decrease over the course of the iterations.

By means of a simulation study, Rai and Matthews (1993) show that the run time of the EM algorithm could be greatly shortened by carrying out a single Newton-Raphson cycle at the maximization step when this maximization requires an iterative process. They name this algorithm EM1. Another way to reduce the run time is to avoid calculating second derivatives in the maximization. This reduction will be especially great if there are a large number of observations and parameters to be estimated. Once again, the second derivatives may be approximated by means of the first derivatives (Beaumont and Demnati, 1998), and the algorithm EM1H1 is obtained.

4.3 Combination of algorithms

The EM algorithm is known for requiring a great number of iterations before converging. However, it has the property of having the likelihood function that does not decrease over the course of the iterations, which makes the algorithm useful, especially if the iterative process is started with poor initial values. The EM1 and EM1H1 algorithms do not have this property, but in practice, they will seldom produce a decrease in the likelihood function. By contrast, algorithms of the Newton-Raphson type take fewer iterations to converge but are much more sensitive to the choice of the initial values.

The approach here therefore consists in starting the iterative process with an EM-type algorithm and finishing with an algorithm of the Newton-Raphson type. Consequently, two convergence criteria are used: a less strict one determining when to stop the EM-type algorithm, and a stricter one for determining when to stop the iterative process. Therefore, the stricter the first criterion, the closer one gets to the EM-type algorithm, and the less strict it is, the closer one gets to the algorithm of the Newton-Raphson type. In what follows, the focus is solely on the following combination: EM1H1NRH1 (EM1H1 in combination with NRH1).

5. SIMULATION STUDY

In order to compare the different algorithms described in Section 4, we conducted a simulation study. The main objective of the study was to evaluate the algorithms in terms of the number of iterations required to achieve convergence, total run time including data reading time, and efficiency in finding the solution.

To conduct this study, we simulated a number of data sets in which the dependent variable was dichotomous and in which only the intercept was random. Thus we used the mixed logistic model. The data sets contain 1,000 individuals, with 12 observations per individual (12 time periods) , and the data are from a mixture of two distributions. We then tested the behaviour of the different algorithms for the following four cases: (A) well-separated distributions and good initial values, (B) well-separated distributions and poor initial values, (C) distributions close together and good initial values, and (D) distributions close together and poor initial values.
The convergence criterion is based on the relative change in the NLLF between two successive iterations.

Convergence is considered to be achieved when the change is less than $\epsilon = 1\times10^{-10}$. For EM1H1NRH1, this criterion is applied for a first time to end EM1H1 with $\epsilon = 1\times10^{-3}$ and for a second time to end the iterative process with $\epsilon = 1\times10^{-10}$. The maximum number of iterations is set at 151.

In Section 5.1, the mixed logistic model with 25 independent variables is considered. In Section 5.2, the algorithms NRH1 and EM1H1NRH1 are compared, varying the number of independent variables from 0 to 60. All the calculations were made using a portable Pentium 90 computer. The reader is referred to Beaumont and Demnati (1998) for further details concerning the simulation study.

5.1 Model with 25 independent variables

We first considered the mixed logistic model with 25 dichotomous independent variables. The values of these variables may vary over time and are obtained by means of a Bernoulli distribution with probability of success equal to 0.5. We simulated one data set in which the distributions are well-separated and another in which the distributions are close together. Table 1 shows the results for five algorithms studied. The run time is measured in seconds.

Table 1: Comparison of algorithms for a model with 25 independent variables

	CASE A		CASE B		CASE C		CASE D	
METHOD	ITER	TIME	ITER	TIME	ITER	TIME	ITER	TIME
NR	3	647.9	X	X	4	678.88	X	X
NRH1	8	126.82	12	171.86	9	139.24	15	191.2
EM1	19	1,027.6	35	1,675.51	151	5,483.77	151	5,602.67
EM1H1	14	112.76	18	148.19	151	976.3	151	1154.7
EM1H1NRH1	8	143.75	12	185.81	9	186.87	18	223.1

Note: X indicates that the global maximum of the likelihood function was not attained. It is assumed that this maximum is the one that corresponds to the largest value observed from the NLLF among all the algorithms considered and the two types of initial values.

The main finding that emerges from Table 1 is: the algorithms that have to calculate second derivatives are much slower. For example, the NR algorithm is the one that takes the fewest iterations to converge when it begins the iterative process with good initial values, whereas it is far from being the fastest. It should be noted that this algorithm did not converge toward the global maximum when the initial values were poor (cases B and D).

When the distributions are well-separated, the EM1H1 algorithm is faster in terms of run time. On the other hand, it reaches the maximum number of iterations (151) when the distributions are close together. However, it has the following advantages: not much time is required for each iteration, and most of the change in the likelihood function takes place in the first few iterations. For case D, for example, the NLLF equals -7,901.02 at the beginning, -5,331.37 after the fifth iteration, -5,326.14 after the twentieth iteration and -5,320.27 when the global maximum is reached. The EM1H1 algorithm is therefore quite efficient when combined with an algorithm that requires fewer iterations to converge, such as NRH1.

Only two algorithms converged for all four cases: NRH1 and EM1H1NRH1. The algorithm EM1H1NRH1 is somewhat slower than NRH1 but reads the data twice (being a combination of two procedures), which makes it appear less attractive.

5.2 Comparison of NRH1 and EM1H1NRH1 for case D

In order to determine which algorithm to choose between NRH1 and EM1H1NRH1, we simulated one data set containing only the intercept, another with five independent variables that can vary over time, another with 10 independent variables that can vary over time, and so forth, up to 60 independent variables. We thus obtained 13 data sets for which the values of the independent variables are obtained by means of a Bernoulli distribution with probability of success equal to 0.5. We settled for examining case D only, which is a realistic scenario in practice. Figure 2 shows the number of iterations required to converge according to the number of independent variables for the two algorithms considered. Figure 3 shows the same chart, except that the variable on the vertical axis is the run time rather than the number of iterations.

An analysis of figures 2 and 3 shows that the number of independent variables has a greater impact on run time than the number of iterations required to converge. The algorithm NRH1 is faster, especially when there are a large number of independent variables. It should again be noted that the run time includes twice the data reading time for EM1H1NRH1. On the other hand, the NRH1 algorithm is less efficient for finding the global maximum, since it did not achieve that maximum for 4 of the 13 simulated samples, whereas EM1H1NRH1 achieved the global maximum in all cases. This is why the curves for the NRH1 algorithm are not continuous.

Figure 2: Number of iterations according to number of independent variables

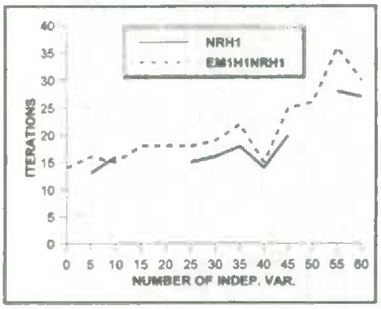
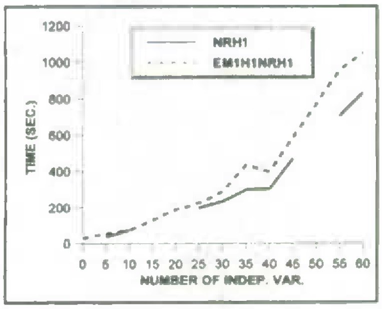


Figure 3: Run time according to number of independent variables



In Table 1 in Section 5.1, we saw that the EM1H1 algorithm was the fastest when distributions were well-separated and the NRH1 algorithm was faster when the distributions were closer together. Thus, EM1H1NRH1, which is a combination of these two algorithms, is an attractive compromise with respect to run time. It is also the most efficient algorithm for finding the global maximum.

6. CONCLUSION

In the preceding section, we saw that avoiding the calculation of second derivatives could greatly reduce the run time. Hence the EM1H1NRH1 algorithm, which is a compromise between EM1H1 (EM algorithm with one iteration at the maximization step and approximation of second derivatives) and NRH1 (Newton-Raphson algorithm and approximation of second derivatives) is attractive from this standpoint. Furthermore, this algorithm is highly efficient for finding the global maximum of the likelihood function, even when the distributions are close together.

Other algorithms that do not require calculating second derivatives could be combined with EM1H1. Among the potential candidates are the quasi-Newton algorithms. They were not, however, considered in this study since it does not appear possible to use them with the NLIN procedure of SAS. Therefore, the results would not have been comparable with the algorithms studied.

In this document, we focused on the problem of estimating parameters for dichotomous data. However, it is realistic to believe that the conclusions of this study also apply to other types of distributions. Among other things, avoiding the calculation of second derivatives should reduce the run time regardless of the distribution of the data.

ACKNOWLEDGMENTS

The authors first wish to thank the Small Area and Administrative Data Division of Statistics Canada, which made this study possible. They are also grateful to Yanick Beaucage and Hew Gough for their comments, which served to improve the quality of this article.

REFERENCES

Beaumont, J.F., and Demnati, A. (1998). Estimation des paramètres d'un modèle pour un mélange fini de distributions et pour des données longitudinales dichotomiques; une comparaison d'algorithmes. *Working document (upcoming publication)*, Statistics Canada, Ottawa.

Dempster, A.P., Laird, N.M., and Rubin, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society* B, 39, 1-38.

McLachlan, G.J., and Basford, K.E. (1988). *Mixture models: inference and application to clustering*, New York: Marcel Dekker.

Rai, S.N., and Matthews, D.E. (1993). Improving the EM algorithm. *Biometrics*, 49, 587-591.

SAS Institute Inc. (1990). *SAS/STAT User's Guide*, Volume 2, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.

Zeger, S.L., Liang, K., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

CLOSING REMARKS

Proceedings of Statistics Canada Symposium 98
Longitudinal Analysis for Complex Surveys
May 1998

CLOSING REMARKS

Maryanne Webber, Statistics Canada

Thanks, Sylvie. Good Day, Friends.

I promise not to keep you very long, as this has been a very full two days and I am much too humble to think that I could add substantially to the heady stuff that has been offered.

However, I do appreciate the opportunity to say how grateful I am that this conference actually happened, as a step along the long road to mining our longitudinal data sources to their fullest.

I would like to begin by thanking the organizing committee of the 15th symposium, for a truly excellent event.

Mike Hidiroglou, Michel Latouche, Tony Labillois, Sylvie Michaud, Georgia Roberts & Denis Lemire, through hard work and careful planning, have provided us with a great opportunity to exchange ideas on a topic that is of critical importance to many statistical agencies. These people did not give up their day jobs either to plan the conference-I think it was truly a labour of love.

I would next like to thank all the volunteers who worked on this event. You can see on the screen the long list of volunteers who have helped out. Needless to say, it takes an enormous amount of work to successfully plan a conference of this scope, and the volunteers deserve a big, heartfelt thank you.

Next, thanks to our international visitors, who took the time to come and share their experiences with us. It is always a pleasure to hear from colleagues working on panel surveys in other countries who, despite institutional and other differences, share a common understanding of the challenges that such surveys pose.

Finally, I would like to thank our presenters for very stimulating talks on topics that are complex and often difficult to encapsulate in a twenty minute presentation. Judging from the questions and comments of the audience, their topics were well chosen and piqued considerable interest. For all those who had a hand in making this conference a success, I suggest that we offer a round of applause.

Let me take a moment to tell you what I have gained from this symposium. My experience lies more in the area of survey management and defining survey content. Consequently, I found a number of the presentations to be—how shall I say it—a challenge.

I am, however, very much aware that the development of longitudinal data sources requires teamwork by methodologists, subject matter experts, computer analysts and users. Only such an approach will bring success. The program included a fairly wide choice of subjects to underscore the network of disciplines that must help each other in order to achieve the goal.

So...analysts, methodologists, systems analysts all have an important role in ensuring that we really do draw the benefits from these surveys. The agenda for this Symposium underscored this fact by covering not just analytical techniques and data quality issues but also processing concerns, data access, confidentiality protection. These are all major concerns for all of us. Those of us who have lived though the early growing pains of a new longitudinal survey are not likely to ever forget trying to wrestle the time dimension into a data structure, as we learn to "think longitudinally".

In his opening address, Michael Wolfson referred to the new longitudinal data sources in Statistics Canada, on health, children, labour and income dynamics. Others on the drawing board or being implemented covering such topics as school to work transitions and immigration. It seems longitudinal surveys have come into their own, and not just in Canada.

With the European Panel at much the same stage of development as the recent Canadian surveys, and the SIPP redesign in progress, it is important to continue the dialogue, share our experiences in exploiting these new and powerful data sets. So, once again, thanks to all who contributed to this conference. I wish our visitors a safe trip home. Thank you, and I look forward to seeing you again.

AUTHOR LIST
IN ALPHABETIC ORDER

KEY WORD LIST

This list is based on the *key words* provided by the authors in their abstracts. Note that some papers are not referred to in this list because the authors did not provide any key word.