SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2002

•

VOLUME 28

•

NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2002 • VOLUME 28 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

June 2002

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN $47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 x 2 issues); Other Countries, CDN $20 ($10 x 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 28, Number 1, June 2002

CONTENTS

In This Issue 1

Waksberg Invited Paper Series

WAYNE A. FULLER
Regression Estimation for Survey Samples 5

Special Section "Remembering Leslie Kish"

GRAHAM KALTON
Leslie Kish's Impact on Survey Statistics 25

LESLIE KISH
New Paradigms (Models) for Probability Sampling 31

CHARLES H. ALEXANDER
Still Rolling: Leslie Kish's "Rolling Samples" and The American Community Survey 35

JEAN-MICHEL DURR and JEAN DUMAIS
Redesign of the French Census of Population 43

Regular Papers

IAN CAHILL and EDWARD J. CHEN
Benchmarking Parameter Estimates in Logit Models of Binary Choice and Semiparametric Survival Models 51

STEVEN T. GARREN and TED C. CHANG
Improved Ratio Estimation in Telephone Surveys Adjusting for Noncoverage 63

YVES TILLÉ
Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement 77

JUN SHAO and SHAIL BUTANI
Variance Estimation for the Current Employment Survey 87

MICHAEL P. COHEN
Implementing Rao-Shao Type Variance Estimation with Replicate Weights 97

RICHARD VALLIANT
Variance Estimation for the General Regression Estimator 103

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains the second in an annual invited paper series in honour of Joseph Waksberg. A brief description of the series and a short biography of Joseph Waksberg were given in the June 2001 issue of the journal. The author of the Waksberg Invited Paper for 2002 is Wayne Fuller. I would like to thank the members of the Committee, Graham Kalton (chair), Chris Skinner, David Binder and Paul Biemer, for having chosen such a distinguished statistician, who has made profound contributions to many areas of statistical theory and practice, as the author of the second paper in the Waksberg Invited Paper Series.

In his paper entitled "Regression Estimation for Survey Samples" Wayne Fuller presents a broad overview of historical and recent developments in the use of regression models in surveys for estimation, weight calibration and non-response adjustment. After a brief introduction and historical background, he discusses the use of regression models for estimation in complex surveys from a design based perspective. He follows this with an exploration of the model based perspective. Other topics discussed are the use of regression models for multinomial data, techniques available when auxiliary variables are available for every unit of the population, and regression to account for the effects of non-response in surveys. Finally, consideration of a few practical aspects of applications rounds out this insightful overview of an important area of inference from survey data to which Wayne Fuller himself has made many important contributions.

This issue also contains a special section "Remembering Leslie Kish" which includes four papers, one by Leslie Kish himself containing some of his last thoughts on the topics of combining samples and surveys. Two of the other papers discuss implementations of Leslie Kish's idea of rolling censuses. These two papers were also presented at the Statistics Canada Symposium 2001 in a special session entitled "Remembering Leslie Kish".

The first paper in the special section, by Graham Kalton, presents an inspiring overview of Kish's contributions to many areas of statistics. Many of the problems that Kish worked on are put into historical perspective and their practical importance is emphasized.

The paper by Kish presents ideas that he was still working on at the time of his death in October 2000. I am grateful to Graham Kalton and Jack Gambino for making editorial corrections to the paper, but it is presented largely as it was at the time of Kish's death. In this paper he argues that, just as statistics represented a new paradigm in the scientific method, and survey sampling required a new paradigm in statistics, so rolling samples and multi-population surveys require new paradigms in survey methods. We can only speculate as to what the final paper would have been like had Kish lived.

Alexander describes the American Community Survey, planned to be introduced by the U.S. Census Bureau in coming years as a replacement for the decennial census long form. This is a very large survey based very much on the idea of rolling samples and censuses that Kish introduced more than twenty years ago. This paper discusses the concepts, frame, sampling design, and cumulation of samples and weighting.

The final paper in the special section, by Durr and Dumais, describes the new rolling census being introduced in France to replace their more traditional census. In this rolling census, every small commune will be surveyed once within a five year period; larger communes will be divided into five rotation groups, each rotation group being surveyed in one of the five years. This paper describes objectives, design and estimation procedures for the rolling census.

In their article, Cahill and Chen develop an approach to exploit data from multiple surveys and epochs by benchmarking the parameter estimates of logit models of binary choice and semi-parametric survival models. Estimates obtained from a survey rich in explanatory variables are benchmarked to information from a survey with significant historical depth. Cahill and Chen demonstrate how the method can be applied, using the maternity leave module of the LifePaths dynamic microsimulation project at Statistics Canada.

Garren and Chang consider the problem of the non-telephone population in telephone surveys using random digit dialing. Using Public Use Microdata Samples, the propensity that a household owns a phone is estimated using generalized linear regression and is used during estimation. Asymptotic biases and variances are presented for both the non-poststratified and poststratified estimators incorporating and not incorporating the estimated propensity. These four estimators are further compared through a simulation study.

The article by Tillé develops an estimator that can be used to avoid the problem of empty post-strata that can occurs with the usual post-stratified estimator. The idea involves using a conditionally weighted estimator and conditioning on ranks in the population of an auxiliary variable known for all units of this population. In this way, the sizes of the post-strata are set in the sample and random in the population. The next step is to calculate the mean of the conditionally weighted estimators to obtain greater stability. The estimator obtained is calibrated on distribution, linear and exactly unbiased. A simulation study is used to show that the proposed estimator is more robust than the generalized regression estimator when the relation of the variable of interest and the auxiliary variable is not linear. Lastly, the article proposes an approximate estimator of the variance verified using simulations.

Shao and Butani consider the problem of estimating variances for imputed survey estimators. They show that the resulting variances can be estimated in two parts, the first of which can be estimated using a grouped half-sample method that incorporates adjustments to take imputation into account. As the estimation of the second part may entail many derivations, Shao and Butani propose an adjustment to the grouped half-sample method that leads to approximately unbiased variance estimates.

In his paper Cohen describes a method to implement Rao and Shao's jackknife method of estimating variances to account for imputation using replicate weights. Rao and Shao's method involves calculation, for each jackknife replicate, adjusted values of imputed data points. The method can be used with either mean imputation or hot deck imputation. Cohen's method involves adding extra rows to the replicate weight file. For each imputed value, one extra row is added for each respondent in the same imputation class.

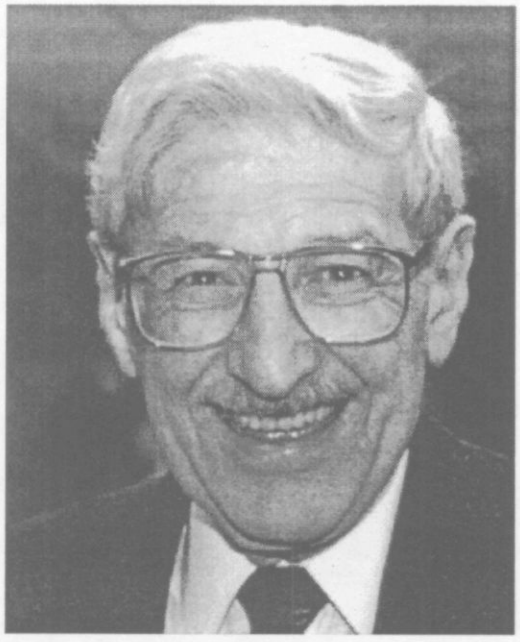
In the last paper of this issue, Valliant studies several variance estimators for the General Regression (GREG) estimator. The interest is in finding variance estimators that, under certain conditions, are approximately unbiased for both the design-variance and the model-variance even if the model that motivates the GREG has an incorrect variance parameter. A key feature of these robust estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. It is shown that the delete-one jackknife implicitly includes the leverage adjustments and is a good choice from either the design-based or model-based perspective. A simulation study shows that these variance estimators have small bias and produce confidence intervals with near-nominal coverage rates.

M.P. Singh

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 3-4
Statistics Canada

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association*. The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*.



JOSEPH WASKBERG

2002 WAKSBERG INVITED PAPER

Author : Wayne A. Fuller

Wayne A. Fuller is Emeritus Distinguished Professor in Statistics and Economics at Iowa State University. He has published approximately 100 articles in more than twenty journals and is author of the texts *Introduction to Statistical Time Series* and *Measurement Error Models*. As a member of the Survey Group at Iowa State University, he had primary responsibility for developing estimation procedures for a large longitudinal national survey called the *U.S. National Resources Inventory*. His research interests in survey sampling include regression estimation, small area estimation, imputation, and multiple phase sampling. He currently chairs the Advisory Committee on Statistical Methods of Statistics Canada.

MEMBERS OF THE WASKBERG PAPER SELECTION COMMITTEE (2002-2003)

David A. Binder (Chair), *Statistics Canada*
J. Michael Brick, *Westat, Inc.*
David R. Bellhouse, *University of Western, Ontario*
Paul Biemer, *Research Triangle Institut, U.S.A.*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)

Past Authors:

Gad Nathan (2001)

Nominations:

Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, D.A. Binder, at Statistics Canada, 3rd, floor R.H. Coats Bldg. Tunneys' Pasture, Ottawa, Ontario, Canada, K1A 0T6, by e-mail binderdav@statcan.ca or by fax (613) 951-5711. Nominations and suggestions for topics must be received by December 6, 2002.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 5-23
Statistics Canada

Regression Estimation for Survey Samples

WAYNE A. FULLER[1]

ABSTRACT

Regression and regression related procedures have become common in survey estimation. We review the basic properties of regression estimators, discuss implementation of regression estimation, and investigate variance estimation for regression estimators. The role of models in constructing regression estimators and the use of regression in nonresponse adjustment are explored.

KEY WORDS: Auxiliary information; Calibration; Least squares; Design consistency; Linear prediction.

1. INTRODUCTION

Design and estimation in survey sampling involve the use of information about the study population to construct efficient procedures. While design and estimation are intimately related, with estimators depending on the design, the two topics are often treated somewhat separately in the survey sampling literature. We follow tradition first studying estimation treating the design as given. The estimation task is to combine the available information about the population, with the sample data to produce good representations of characteristics of interest.

Regression estimation is one of the important procedures that use population information or information from a larger sample, to construct estimators with good efficiency. The information, sometimes called *auxiliary information*, may have been used in the design or may not have been available at the design stage. In surveys of the human population, the information often comes from official sources such as the national census. Similar sources may provide information for other types of surveys. For example, in a survey of land use the total surface area, the area owned by the national government, and the area in permanent water bodies may be available from national data archives.

Three distinct situations can be identified with respect to the nature of the auxiliary information that is available. In the first, the values of the auxiliary vector $\mathbf{x}$ are known for each element in the population at the time of sample selection. In this case the auxiliary variable can be used in designing the sample selection procedure.

In the second situation all values of the vector $\mathbf{x}$ are known, but a particular value cannot be associated with a particular element until the sample is observed. In this case, the auxiliary information cannot be used in design, but a wide range of estimation options are available once the observations are available. For example, the population census may give the age-sex distribution of the population, but a list of individuals and their characteristics is not available to non governmental institutions selecting samples.

In the third situation, only the population mean of $\mathbf{x}$ is known, or known for a large sample. In this case, the auxiliary information cannot be used in design and the estimation options are limited. For example the U.S. Department of Agriculture might release an estimate of the total number of animals of a particular type on farms on a particular date. Our discussion concentrates on this situation.

Two estimation situations can also be identified. In one, a single variable and a parameter, or a very small number of parameters, is under consideration. The analyst is willing to invest a great deal of effort in the analysis, has a well formulated population model, and is prepared to support the estimation procedure on the basis of the reasonableness of the model. In the second situation, a large number of analyses of a large number of variables is anticipated. No single model is judged adequate for all variables. The prototypical example of the second situation is the case in which a data set is prepared by the survey sampler to be analyzed by others. Because the person preparing the data set does not have knowledge of the analysis variables, emphasis is placed on the use of estimators that can be defended with minimal recourse to models.

Regression estimators fall in the class of linear estimators. Linear estimators have a particular advantage in survey sampling because once the weights are calculated they are appropriate for any analysis variable. Several properties of estimators will be examined in our discussion. Given a model, we accept the classical goal of minimizing the mean square error in a class of estimators. That class may be the class of linear estimators that are unbiased under the model, but the class may be further restricted.

Estimators that are scale and location invariant can be used in general settings. Mickey (1959) suggested that the term regression estimator be restricted to linear estimators that are location and scale invariant. While we may not adhere strictly to this definition, we support the distinction

[1] Wayne Fuller, Emeritus Distinguished Professor, Iowa State University, 221 Snedecor Hall, Ames, IA 50011-1210, U.S.A.

between estimators that are location and scale invariant and those that are not. We consider location invariance to be important for sampling designs where the unit of interest for analysis is also the sampling unit. For cluster and two stage designs in which weights are constructed for primary sampling unit totals, location invariance is less important.

Models play an important role in the construction of regression estimators. It is desirable that the estimators retain good properties if the model specification is not exact. Therefore properties conditional on the realized finite population, as well as properties under the model, are important.

Linear estimators that reproduce the known means of the auxiliary variables are said to be calibrated. This is a desirable property in that, for example, the marginals of tables with an auxiliary variable as an analysis variable agree with known totals. If the auxiliary variable is of no analytic interest, then calibration is less important.

2. BACKGROUND

The earliest references to the use of regression in survey sampling include Jessen (1942) and Cochran (1942). Regression in similar contexts would certainly have been used earlier and Cochran (1977, page 189) mentions a regression on leaf area by Watson (1937). It is interesting that Jessen's use of regression was essentially composite estimation where regression was used to improve estimates for two time points given samples at each point with some common elements in the two samples. Cochran (1942) gave the basic theory for regression in survey sampling relying heavily on linear model theory. He showed that the linear model did not need to hold in order for the regression estimator to perform well. He derived an expression for the $O(n^{-1})$ bias and an $O(n^{-2})$ approximation for the variance. He also showed that for the model with regression passing through the origin and error variances proportional to x, the ratio estimator is the generalized least squares estimator.

Regression estimation attracted theoretical interest in the 1950's, often in the form of studies of the bias. See Mickey (1959). Brewer (1963) is an early reference that considers linear estimation using a superpopulation model to determine an optimal procedure. He was concerned with finding the optimal design for the ratio estimator and discussed the possible conflict between an optimal design under the model and a design that is less model dependent. See also Brewer (1979). Royall (1970) argued for the use of models, that the conditional properties that are important are those conditional on the auxiliary information in the sample, and that the design should be chosen to optimize those properties. Royall and his coworkers, *e.g.*, Royall and Cumberland (1981), studied the conditional properties of regression estimators, conditional on the realized sample of auxiliary variables.

A great deal of research was conducted in the 1970's and 1980's on the general nature of the regression estimator in survey samples and on the degree to which the model prediction approach can be reconciled with the design perspective. Fuller (1973, 1975) gave the large sample properties of a vector of regression coefficients computed from a survey sample. Isaki (1970) studied regression estimators and the results were published in expanded versions in Isaki and Fuller (1982) and Fuller and Isaki (1981). It was shown that a regression estimator constructed under a model is design consistent for the population mean if the model contains certain variables. Cassel, Särndal and Wretman (1976) considered both model and design principles in estimator construction and suggested the term "generalized regression estimator" for design consistent estimators of the total of the form

$$\hat{T}_{y,\mathrm{GREG}} = \hat{T}_{y,\mathrm{HT}} + (T_{x,N} - \hat{T}_{x,\mathrm{HT}})\hat{\boldsymbol{\beta}},$$

where $\hat{T}_{y,\mathrm{HT}}$ and $\hat{T}_{x,\mathrm{HT}}$ are the Horvitz-Thompson estimators of the totals of y and x, respectively, $T_{x,N}$ is the know population total of x and $\hat{\boldsymbol{\beta}}$ is an estimated regression coefficient. Särndal (1980), Wright (1983), and Särndal and Wright (1984) discussed classes of regression estimators. The text by Särndal, Swensson and Wretman (1992) contains an extensive discussion of regression estimation and Mukhopadhyay (1993) is a review.

It was the 1970's before the use of regression for general purpose, multiple characteristic, surveys appeared and it was the 1990's before the use of regression weighting could be called widespread. An early use of regression weights was at Doane Agricultural Services Inc., now Doane Marketing Research. During 1971-1972 a readership study of farmers was conducted under the direction of Mr. John Wilkin in which 6,920 farmers responded. Weights for the respondents were constructed using regression procedures, where the controls came from the U.S. Agricultural Census and from Department of Agriculture sources. Doane provided financial support to Iowa State University to develop a regression weight generation program. To guarantee positive weights in the Doane study, observations with small weights were grouped and assigned a common weight. Grouping continued until the common weight was positive. Later computer programs used modifications of the Huang and Fuller (1978) procedure to guarantee positive weights. Doane has used regression weights for their syndicated market research studies since 1972.

Regression estimation was first used at Statistics Canada in 1988 for the Canadian Labour Force Survey. In 1992 regression estimation was used by the 1991 *Canadian Census of Population* to ensure that the weighted sum of variables collected via the long form (a one in five systematic sample of all households in Canada) was equal to known household and population totals as collected in the 1991 Census. See Bankier, Rathwell and Majkowski (1992) and Bankier, Houle and Luc (1997). The regression estimator is also the key component of the Generalized

Estimation System (GES) developed at Statistics Canada and used in numerous business and social surveys since its release in 1992. The methodology is described in Estevao, Hidiroglou and Särndal (1995). See also Hidiroglou, Särndal and Binder (1995). Regression estimation is now used to construct composite estimators for the Canadian Labour Force Survey. See Singh, Kennedy and Wu (2001), Gambino, Kennedy and Singh (2001) and Fuller and Rao (2001).

Bethlehem and Keller (1987) report on the use of regression estimation at the Netherlands Central Bureau of Statistics (now Statistics Netherlands) in a program called LIN WEIGHT. Nieuwenbrock, Renssen and Hofman (2000) describe the software package Bascula, that has replaced LIN WEIGHT. Deville, Särndal and Sautory (1993) describe a computer program CALMAR developed at Institut National de la Statistique et des Etudes Economiques (I. N. S. E. E.) that computes weights of the regression type with options for different objective functions. A program developed at Statistics Sweden and called CLAN97 is documented in Anderson and Nordberg (1998). Folsom and Singh (2000) discuss a procedure developed at the Research Triangle Institute.

3. THE CLASSICAL LINEAR MODEL

The classical linear model is the foundation for survey regression estimation, but the survey situation requires certain adaptations. To introduce regression estimation for survey samples, we review the classical linear model. Assume

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,\ i = 1, 2, \cdots, n,$$
$$e_i \sim \mathrm{NI}(0, \sigma_e^2), \tag{3.1}$$

where e_i is independent of the k-dimensional row vectors $\mathbf{x}_j$ for all i and j, and $\boldsymbol{\beta}$ is the unknown parameter column vector. We will also use matrix representations for the sample quantities. Thus, for a sample of n elements,

$$\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_n') \text{ and } \mathbf{y}' = (y_1, y_2, \cdots, y_n).$$

Given a sample of size n and treating the $\mathbf{x}_i$ as fixed, the best (minimum mean squared error) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\sum_{i\in A}\mathbf{x}_i' y_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{3.2}$$

where A is the set of indexes of the sample elements and we assume, as we will throughout, that the matrix to be inverted is nonsingular. If the e_i are not normally distributed, $\hat{\boldsymbol{\beta}}$ is the estimator with smallest variance in the class of linear unbiased estimators. The estimator of a linear combination of the coefficients, say $\theta_\alpha = \sum_{j=1}^k \alpha_j\beta_j$, can be written as

$$\hat{\theta}_\alpha = \sum_{i\in A} w_{\alpha i}\, y_i$$

where the weights, $w_{\alpha i}$, minimize the Lagrangean

$$\sum_{i\in A} w_{\alpha i}^2 + \sum_{j=1}^k \lambda_j\left(\sum_{i\in A} w_{\alpha i}\, x_{ij} - \alpha_j\right)$$

and the λ_j are Lagrange multipliers. The variance of $\hat{\theta}_\alpha$ is

$$V\{\hat{\theta}_\alpha\} = V\left\{\sum_{i\in A} w_{\alpha i}e_i\right\} = \sum_{i\in A} w_{\alpha i}^2\sigma_e^2$$

because the weights are functions of the $\mathbf{x}_i$ and not of y_i.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1} V\left\{\sum_{i\in A}\mathbf{b}_i'\right\}\left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}$$
$$= V\left\{\sum_{i\in A}\mathbf{c}_i\right\} \tag{3.3}$$

where $\mathbf{b}_i' = \mathbf{x}_i' e_i$ and $\mathbf{c}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' e_i$. Because e_i is independent of $\mathbf{x}_j$ for all i and j,

$$V\left\{\sum_{i\in A}\mathbf{b}_i'\right\} = \sum_{i\in A} V\{\mathbf{b}_i'\} = \sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\sigma_e^2$$

and we obtain the familiar expression,

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\sigma_e^2.$$

The usual unbiased estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained by replacing σ_e^2 with the unbiased estimator of σ_e^2 obtained as the mean square of the residuals, $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$. An estimator of the covariance matrix that estimates $V\{\sum_{i\in A}\mathbf{b}_i'\}$ directly is

$$\tilde{V}_b\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\sum_{i\in A}\hat{\mathbf{b}}_i'\hat{\mathbf{b}}_i\left(\sum_{i\in A}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}$$
$$= \sum_{i\in A}\hat{\mathbf{c}}_i'\hat{\mathbf{c}}_i, \tag{3.4}$$

where $\hat{\mathbf{b}}_i' = \mathbf{x}_i'\hat{e}_i$ and $\hat{\mathbf{c}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'\hat{e}_i$. In the same way

$$\hat{V}_b\{\hat{\theta}_\alpha\} = \sum_{i\in A} w_{\alpha i}^2\hat{e}_i^2 \tag{3.5}$$

is a linear combination of the elements of (3.4) and is a consistent estimator of $V\{\hat{\theta}_\alpha\}$. The estimator (3.4) is a consistent estimator of $V\{\hat{\boldsymbol{\beta}}\}$ when the covariance matrix of the e_i is a diagonal matrix with bounded elements. Thus it is a more robust estimator. However, the estimator (3.4) is biased downward because the variance of $\hat{e}_i$ is usually less than the variance of e_i. Two methods are available for reducing the bias. The first is to make a degrees-of-freedom adjustment by multiplying $\tilde{V}_b\{\hat{\boldsymbol{\beta}}\}$ by $(n-k)^{-1}n$, where k is the dimension of $\mathbf{x}_i$. An alternative adjustment is to replace $\hat{e}_i$ with

$$\tilde{e}_i = (1 - \psi_{ii})^{-0.5}\hat{e}_i,$$

where ψ_{ii} is the i-th diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. See Horn, Horn and Duncan (1975), Royall and Cumberland (1978) and Cook and Weisberg (1982, section 2.2).

If we observe the value $\mathbf{x}_i$ for an element, but do not observe y_i, then the best predictor of y_i for that element is $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Likewise, if we know the sum of $\mathbf{x}_i$ for a set of $\mathbf{x}$'s, then the best predictor for the sum of the y_i is the sum of $\mathbf{x}_i\hat{\boldsymbol{\beta}}$. Thus, given a set of N elements that satisfy model (3.1), a set of observations $(y_i, \mathbf{x}_i)$ on a subset denoted by A, and the known values of $\mathbf{x}_i$ for the remaining $N-n$ elements,

$$\hat{Y}_{N-n,\text{reg}} = \sum_{i\in\bar{A}} \hat{y}_i = \sum_{i\in\bar{A}} \mathbf{x}_i\hat{\boldsymbol{\beta}},$$

where $\bar{A}$ is the set of elements for which y is not observed, is the best predictor of the sum of the unobserved y's. See Goldberger (1962), Brewer (1963), Royall (1970), Harville (1976) and Graybill (1976, section 12.2). Hence

$$\hat{T}_{y,\text{reg}} = \sum_{i\in A} y_i + \hat{Y}_{N-n,\text{reg}} \tag{3.6}$$

is the best predictor for the total of N observations.

If the first element in the x-vector is always one, we can partition the x-vector as $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$ and write the regression estimator of the mean as

$$\bar{y}_{\text{reg}} = N^{-1}\hat{T}_{y,\text{reg}} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} = \bar{y}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})\hat{\boldsymbol{\beta}}_1, \tag{3.7}$$

where $\hat{\boldsymbol{\beta}}$ of (3.2) is partitioned as $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1')'$ and $(\bar{y}_n, \bar{\mathbf{x}}_n)$ is the vector of simple sample means. We call $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ the regression estimator of the mean.

Given the model (3.1), the expected value of the mean of y for the finite population of N elements generated by the model is $\bar{\mathbf{x}}_N\boldsymbol{\beta}$ and $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ is an unbiased estimator of the finite population mean. This, we believe, is the point at which regression estimation for the finite population mean under more complex designs begins.

4. DESIGN BASED ESTIMATION

The development of this section treats the finite population as a sample realization from an infinite population. The use of such models has a long history in survey sampling. Some references through 1970 are Cochran (1939, 1942, 1946), Deming and Stephan (1941), Madow and Madow (1944), Yates (1949), Godambe (1955), Hájek (1959), Rao, Hartley, and Cochran (1962), Konijn (1962), Brewer (1963), Godambe and Joshi (1965), Hanurav (1966), Ericson (1969), Isaki (1970), and Royall (1970).

To discuss the large sample properties of regression estimators we consider sequences of finite populations and associated probability samples. The set of indices of the elements in the Nth finite population is $U_N = \{1, \cdots, N\}$, where $N = 1, 2, \cdots$. Associated with the ith element of the Nth population is a row vector of characteristics $\mathbf{z}_{iN} = (y_{iN}, \mathbf{x}_{iN})$. Let

$$\mathbf{F}_N = \left[(y_{1N}, \mathbf{x}_{1N}), (y_{2N}, \mathbf{x}_{2N}), \cdots, (y_{NN}, \mathbf{x}_{NN})\right]$$

be the set of vectors for the N-th finite population. The subscript N on the vectors will often be omitted. The finite population mean is

$$\bar{\mathbf{z}}_N = (\bar{y}_N, \bar{\mathbf{x}}_N) = N^{-1}\sum_{i=1}^{N}(y_i, \mathbf{x}_i). \tag{4.1}$$

We denote the set of indices appearing in the sample selected from the Nth finite population by A_N.

When the finite population is a sample from an infinite superpopulation, the probability properties of a sample are determined by the properties of the superpopulation and the properties of the probability mechanism used to select the sample. One can consider the unconditional properties, the properties conditional on the particular finite population, or the properties conditional on some part of the realized sample.

Properties conditional on the finite population depend primarily on the survey design and are often called design properties. Thus an estimator $\hat{\theta}$ is said to be design consistent for the finite population parameter θ_N if, for all $\varepsilon > 0$,

$$\lim_{N,n\to\infty} \text{prob}\left\{|\hat{\theta} - \theta_N| > \varepsilon \mid \mathbf{F}_N\right\} = 0,$$

where the notation means that we condition on the realized finite population $\mathbf{F}_N$ and, hence, the probability is with respect to the design.

Assume the finite population is generated as independent selections from a superpopulation for which $E\{\mathbf{z}_i'\mathbf{z}_i\}$ is positive definite, where $\mathbf{z}_i = (y_i, \mathbf{x}_i)$. We define a superpopulation vector of least squares regression coefficients by

$$\boldsymbol{\beta} = \left[E\{\mathbf{x}_i'\mathbf{x}_i\}\right]^{-1}E\{\mathbf{x}_i'y_i\}. \tag{4.2}$$

Given a sample of n observations on $\mathbf{z}_i$ we define the $n\times(k+1)$ matrix $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$ of observations, where the ith row of $\mathbf{Z}$ is $(y_i, \mathbf{x}_i)$. If we assume the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{4.3}$$

$$E\{\mathbf{u}, \mathbf{u}\mathbf{u}'\} = (\mathbf{0}, \boldsymbol{\Phi}),$$

the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}. \tag{4.4}$$

The model (4.3) serves as motivation for estimators of the form (4.4) but we shall consider estimators where $\boldsymbol{\Phi}$ is a general symmetric positive definite weight matrix, not necessarily the covariance matrix of the errors.

We give the large sample properties of the vector of estimated regression coefficients (4.4) following Fuller (1975). See also Hidiroglou (1974), Scott and Wu (1981), and Robinson and Särndal (1983).

Assume the superpopulation has eighth moments and that the sample design is such that the error in the Horvitz-Thompson estimator of the mean is $O_p(n^{-1/2})$, where the Horvitz-Thompson estimator of the mean is

$$\bar{\mathbf{z}}_{HT} = (\bar{y}_{HT}, \bar{\mathbf{x}}_{HT}) = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i \qquad (4.5)$$

and π_i is the selection probability for element i. Then the error in the vector of regression coefficients is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1} \bar{\mathbf{b}}_{HT}' + O_p(n^{-1}), \qquad (4.6)$$

where

$$\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1} \mathbf{Q}_{xyN}, \qquad (4.7)$$

$$(\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}) = E\left\{(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) \mid \mathbf{F}_N\right\}, \qquad (4.8)$$

$$(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) = n^{-1} (\mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{X}, \mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{y}),$$

$$\bar{\mathbf{b}}_{HT} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i, \qquad (4.9)$$

$\mathbf{b}_i' = n^{-1} N \pi_i \zeta_i e_i$, $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_N$, and ζ_i is column i of $\mathbf{X}' \boldsymbol{\Phi}^{-1}$. By (4.9) the error in the estimator of $\boldsymbol{\beta}_N$ is approximately the error in a Horvitz-Thompson estimator of the mean. In result (4.6), the $\boldsymbol{\beta}_N$ is defined as a function of the expected values of the sample quantities $(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy})$. Thus $\boldsymbol{\beta}_N$ is not necessarily the ordinary least squares finite population regression coefficient. The vector $\mathbf{b}_i$ of (4.9) is the generalization of the vector $\mathbf{b}_i$ of (3.3). If the limiting distribution of the properly standardized Horvitz-Thompson estimator is normal, and if there is a design consistent estimator of the variance of the Horvitz-Thompson estimator, then it is possible to construct tests and confidence intervals for the coefficients. Assume the design is such that

$$\mathbf{V}_{\bar{z}\bar{z}}^{-1/2} (\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N) \mid \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \qquad (4.10)$$

as $N, n \to \infty$, where $\mathbf{V}_{\bar{z}\bar{z}}$ is the covariance matrix of $\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N$. If $\mathbf{V}_{\bar{z}\bar{z}}$ is $O(n^{-1})$ and the estimator $\hat{\mathbf{V}}_{\bar{z}\bar{z}}$ is consistent for $\mathbf{V}_{\bar{z}\bar{z}}$, then

$$\left[\hat{V}\{\hat{\boldsymbol{\beta}}\}\right]^{-1/2} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) \mid \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \qquad (4.11)$$

where

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{V}}_{\bar{b}\bar{b}} \hat{\mathbf{Q}}_{xx}^{-1}, = \hat{V}\{\bar{\mathbf{c}}_{HT}'\}, \qquad (4.12)$$

$\hat{\mathbf{V}}_{\bar{b}} = \hat{V}\{\bar{\mathbf{b}}_{HT}'\}$ is the estimated design variance of $\bar{\mathbf{b}}_{HT}$ calculated with $\hat{\mathbf{b}}_i' = n^{-1} N \pi_i \zeta_i' \hat{e}_i$ $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, and $\hat{V}\{\bar{\mathbf{c}}_{HT}'\}$ is the estimated design variance of $\bar{\mathbf{c}}_{HT}'$ calculated with $\hat{\mathbf{c}}_i' = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{b}}_i'$. The limiting properties hold for stratified samples and for stratified two stage samples under mild restrictions on the sequence of populations.

By analogy to (3.7), a regression estimator of the finite population mean is obtained by evaluating the estimated regression function at the population mean of $\mathbf{x}$ to obtain

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}, \qquad (4.13)$$

where $\hat{\boldsymbol{\beta}}$ is of the form (4.4) with a general $\boldsymbol{\Phi}$ matrix. The estimator can be written as $\mathbf{w}'\mathbf{y}$, where the vector of weights can be constructed by minimizing the Lagrangean

$$\mathbf{w}' \boldsymbol{\Phi} \mathbf{w} + (\mathbf{w}' \mathbf{X} - \bar{\mathbf{x}}_N) \boldsymbol{\lambda}$$

and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers.

If there is a column vectors $\boldsymbol{\gamma}$ such that

$$\mathbf{X}\boldsymbol{\gamma} = \boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J} \qquad (4.14)$$

for all possible samples, where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \cdots, \pi_n)$ and $\mathbf{J}$ is an n-dimensional column vector of ones, then the regression estimator $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ of (4.13) with $\hat{\boldsymbol{\beta}}$ defined in (4.4) is a design consistent estimator of $\bar{y}_N$. It follows from (4.11) that

$$\left[\bar{\mathbf{x}}_N \hat{V}\{\hat{\boldsymbol{\beta}}\} \bar{\mathbf{x}}_N'\right]^{-1/2} \left(\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} - \bar{y}_N\right) \xrightarrow{L} N(0, 1). \qquad (4.15)$$

The requirement of (4.14) that $\boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J}$ be in the column space of $\mathbf{X}$ is crucial for design consistency. Simple ways to satisfy this requirement are to let one column of $\mathbf{X}$ be the column of ones and to use a multiple of $\mathbf{D}_\pi$ as $\boldsymbol{\Phi}$, or to let one column of $\mathbf{X}$ be the elements π_i^{-1} and set $\boldsymbol{\Phi} = \mathbf{I}$, or to let one column of $\mathbf{X}$ be the elements π_i and set $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$. If $\mathbf{X}$ is composed of the single column vector with elements π_i and if $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$, then the estimator (4.13) reduces to the Horvitz-Thompson estimator of (4.5) for fixed size designs. If $\mathbf{X} = \mathbf{J}$ and $\boldsymbol{\Phi} = \mathbf{D}_\pi$, the estimator (4.13) reduces to the ratio estimator,

$$\bar{y}_\pi = \left(\sum_{i \in A} \pi_i^{-1}\right)^{-1} \sum_{i \in A} \pi_i^{-1} y_i, \qquad (4.16)$$

which is location and scale invariant.

To see the nature of the estimator when (4.14) is satisfied, let, with no loss of generality, $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}_1)$, where $\mathbf{x}_0 = \boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J}$ and $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i})$. Then

$$\bar{y}_{reg} = \bar{x}_{0,N} \bar{x}_{0,\pi}^{-1} \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{x}_{0,N} \bar{x}_{0,\pi}^{-1} \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_1, \qquad (4.17)$$

where

$$\hat{\boldsymbol{\beta}}_1 = \left[(\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})' \boldsymbol{\Phi}^{-1} (\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})\right]^{-1} \times (\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})' \boldsymbol{\Phi}^{-1} \mathbf{y},$$

$\hat{\boldsymbol{\mu}}_{x1} = \bar{x}_{0,\pi}^{-1} \bar{\mathbf{x}}_{1,\pi}$, and $(\bar{y}_\pi, \bar{\mathbf{x}}_\pi)$ is defined in (4.16). The ratios, such as $\bar{x}_{0,\pi}^{-1} \bar{y}_\pi$, can also be written as ratios of Horvitz-Thompson estimators. If $\mathbf{J}$ is in the column space of $\mathbf{X}$, estimator (4.17) is location invariant. If $\boldsymbol{\Phi} = \mathbf{D}_\pi$, then $\bar{x}_{0,\pi}^{-1} \bar{x}_{0,N} = 1$, and

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_1, \qquad (4.18)$$

where

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_{i\in A} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})' \pi_i^{-1} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})\right]^{-1} \times \sum_{i\in A} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})' \pi_i^{-1} (y_i - \bar{y}_\pi). \quad (4.19)$$

Also, when $\boldsymbol{\Phi} = \mathbf{D}_\pi$, the $\boldsymbol{\beta}_N$ of (4.7) is the population regression coefficient

$$\boldsymbol{\beta}_N = \left[\sum_{i\in U} \mathbf{x}_i' \mathbf{x}_i\right]^{-1} \sum_{i\in U} \mathbf{x}_i' y_i. \quad (4.20)$$

Because the regression estimator of the mean is a linear combination of regression coefficients, it is a regression coefficient for a linear combination of the original x-variables. To see this, let $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i}) = (1, \mathbf{x}_{1,i})$, and define a new vector with one in the first position and a second vector with population mean equal to zero obtained by subtracting the original population mean $\bar{\mathbf{x}}_{1,N}$ from the original $\mathbf{x}_{1,i}$ vector. Let $\mathbf{q}_i = (1, \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})$ be the transformed vector. Then the transformed regression model is

$$y_i = \mathbf{q}_i \boldsymbol{\gamma} + e_i, \quad (4.21)$$

where the finite population coefficient vector is

$$\boldsymbol{\gamma}_N = (\bar{y}_N, \boldsymbol{\beta}_{1,N}')' = \left(\sum_{i\in U} \mathbf{q}_i' \mathbf{q}_i\right)^{-1} \sum_{i\in U} \mathbf{q}_i' y_i. \quad (4.22)$$

The expression for the regression estimator of the mean becomes

$$\bar{y}_{\text{reg}} = \bar{\mathbf{q}}_N \hat{\boldsymbol{\gamma}} = \hat{\gamma}_0, \quad (4.23)$$

where $\hat{\boldsymbol{\gamma}}$ is obtained from (4.4) with $\mathbf{q}_i$ replacing $\mathbf{x}_i$. Because the estimator is a linear estimator of the form $\mathbf{w}'\mathbf{y}$, we can write

$$\bar{y}_{\text{reg}} = \sum_{i\in A} w_i y_i = \sum_{i\in A} \pi_i^{-1} g_i y_i, \quad (4.24)$$

where $w_i = \pi_i^{-1} g_i$. Furthermore, the estimated variance from (4.12) is

$$\hat{V}\{\bar{y}_{\text{reg}}\} = \hat{V}\{\hat{\gamma}_0\} = \hat{V}\left\{\sum_{i\in A} \pi_i^{-1} (g_i \hat{e}_i)\right\}, \quad (4.25)$$

where it is understood that the estimated design variance of (4.25) is computed for the variable $g_i \hat{e}_i$, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}}$ is defined in (4.4) The variance estimator (4.25) is a direct generalization of expression (3.5). By transforming the variables so that the population mean of the auxiliary vector is zero, the first element of the regression vector is the regression estimator of the mean and the first element of (4.12) is an estimator of the variance of the regression estimator that contains a component due to estimating $\boldsymbol{\beta}$. This was pointed out in Hidiroglou, Fuller, and Hickman (1978). Also, see Särndal (1982). Särndal, Swensson and Wretman (1989) suggested the g-factor terminology for the calculation of the estimated variance of a regression estimated total.

From (4.17), we can write

$$\bar{y}_{\text{reg}} = \bar{x}_{0,N} \bar{x}_{0,\pi}^{-1} \left[\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1N} - (\bar{y}_N - \bar{\mathbf{x}}_{1,N} \boldsymbol{\beta}_{1N})\right] + O_p(n^{-1}), = \bar{e}_\pi + O_p(n^{-1}),$$

where $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}$. Hence, the variance of the regression estimator can be estimated with

$$\hat{V}\{\bar{e}_\pi\} = \hat{V}\left\{\left(\sum_{i\in A} \pi_i^{-1}\right)^{-1} \sum_{i\in A} \pi_i^{-1} \hat{e}_i\right\}, \quad (4.26)$$

where $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$. Because (4.25) is as easy to compute as (4.26), and is applicable when $\bar{\mathbf{x}}_{1,\pi} - \bar{\mathbf{x}}_{1,N}$ is not $O_p(n^{-1/2})$, the estimator (4.25) is recommended.

The variance of the regression estimator can also be computed using the jackknife or other replication methods, and the use of replication methods is becoming more common. See Frankel (1971), Kish and Frankel (1974), Woodruff and Causey (1976), Royall and Cumberland (1978), and Duchesne (2000). Yung and Rao (1996) showed that (4.25) is identical to a jackknife linearization estimator for stratified multistage designs.

The approach to regression estimation associated with (4.18) and (4.19) falls completely within a design formulation. No models of the population, beyond the existence of moments, are used, through one might argue that one would only consider regression when one feels there is some linear correlation between $\mathbf{x}_{1,i}$ and y_i.

The estimator (4.19) is a very natural estimator because the estimated regression coefficient is a design consistent estimator of the population regression coefficient. It is mildly annoying that (4.18) does not always yield the smallest large sample design variance for the estimated mean. Treating $\hat{\boldsymbol{\beta}}_1$ of (4.18) as a fixed vector, the value that minimizes the variance of the linear combination of means is

$$\boldsymbol{\beta}_{1,\text{dopt}} = \left[V\{\bar{\mathbf{x}}_{1,\pi} \mid \mathbf{F}_N\}\right]^{-1} C\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi \mid \mathbf{F}_N\}. \quad (4.27)$$

See Cochran (1977, page 201), Fuller and Isaki (1981), Montanari (1987, 1999) and Rao (1994). If there is a design consistent estimator of the variance of $\bar{\mathbf{x}}_{1,\pi}$, then the $\boldsymbol{\beta}_{1,d}$ that minimizes the estimated variance

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1,d}\}, \quad (4.28)$$

denoted by $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$, is a consistent estimator of $\boldsymbol{\beta}_{1,\text{dopt}}$. It follows that the estimator

$$\bar{y}_{d,\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_{1,\text{dopt}} \quad (4.29)$$

has the minimum limit variance for design consistent estimators of the form $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \boldsymbol{\beta}_{1,d}$. Also

$$\left[\hat{V}\{\bar{e}_\pi\}\right]^{-1/2} (\bar{y}_{d,\text{reg}} - \bar{y}_N) \xrightarrow{L} N(0, 1), \quad (4.30)$$

where $\hat{V}\{\bar{e}_\pi\}$ is the estimator of (4.26) constructed with $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \mathbf{x}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$.

In a large sample sense, (4.29) answers the question of how to construct a regression estimator with optimum design properties. In practice a number of questions remain. The estimator is obtained under the assumption of a large sample and a vector $\mathbf{x}$ of fixed dimension. In practice there may be a number of potential auxiliary variables and if a large number are included in the regression, terms excluded in the large sample approximation become important. This is particularly true for cluster samples where the number of primary sampling units in the sample is small. In such cases, the number of degrees-of-freedom in $\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\}$ is small and the inverse can be unstable. These issues are discussed further in section 9.

The estimator $\boldsymbol{\beta}_{1,\text{dopt}}$ of (4.29) is linear in y for most designs. See Rao (1994). For example, for a stratified design with simple random sampling within strata,

$$\hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\} = \sum_{h=1}^{H} K_h \sum_{j=1}^{n_h} (\mathbf{x}_{1,hj} - \bar{\mathbf{x}}_{1,h})'(y_{hj} - \bar{y}_h), \quad (4.31)$$

where

$$K_h = W_h^2(1 - f_h)(n_h - 1)^{-1} n_h^{-1} = N^{-2}\pi_h^{-2}(1 - f_h)(n_h - 1)^{-1} n_h,$$

$N^{-1}N_h = W_h$, N_h is the size of stratum h, $f_h = \pi_h = N_h^{-1} n_h$, and n_h is the sample size in stratum h. It follows that the weights associated with estimator (4.29) are

$$w_{hi} = N^{-1}\pi_h^{-1} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \times \left[\sum_{t=1}^{H} K_t \sum_{j=1}^{n_h} (\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t})'(\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t})\right]^{-1} \times K_h(\mathbf{x}_{1,hi} - \bar{\mathbf{x}}_{1,h})'. \quad (4.32)$$

See also Särndal (1996). The weights of (4.32) can be constructed by minimizing $\sum_{hi\in A} w_{hi}^2 K_h^{-1}$ subject to the constraints

$$\sum_{i\in A_h} w_{hi} = N^{-1}N_h, \; h = 1, 2, \cdots, H,$$

and

$$\sum_{hi\in A} w_{hi}\mathbf{x}_{1,hi} = \bar{\mathbf{x}}_{1,N},$$

where A_h is the set of sample elements in stratum h.

The estimator of (4.19) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$ is a function of Horvitz-Thompson estimators of population moments. The estimator (4.17) with $\boldsymbol{\Phi}^{-1} = \text{diag}\{K_t\}$, the diagonal matrix with K_t on the diagonal for elements in stratum t, and dummy variables for stratum effects, gives the estimator of the mean in the class

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_1$$

with the smallest estimated design variance. If the true slopes in the strata are the same and if the selection probabilities are proportional to the square roots of the within-stratum variances, then the use of $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$ gives a smaller small sample MSE than the use of $\boldsymbol{\Phi}^{-1} = \text{diag}\{K_t\}$ because the sum of $w_{hi}^2\sigma_h^2$ is smaller. Fuller and Isaki (1981) noted that the design-optimum estimator is often well approximated by the estimator constructed with $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$.

We have introduced regression estimation for the mean, but it is often the totals that are estimated and totals that are used as controls. Consider the regression estimator of the total of y defined by

$$\hat{T}_{y,\text{reg}} = \hat{T}_{y,\pi} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\hat{\boldsymbol{\beta}}_{y\cdot x}, \quad (4.33)$$

where $\mathbf{T}_{x,N}$ is the known total of $\mathbf{x}$ and $(\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi})$ is a vector of design consistent estimators of $(T_{y,N}, \mathbf{T}_{x,N})$. By analogy to (4.28), the estimator of the optimum $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{y\cdot x} = \left[\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}\right]^{-1}\hat{C}\{\hat{\mathbf{T}}'_{x,\pi}, \hat{T}_{y,\pi}\}, \quad (4.34)$$

where $\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}$ is a design consistent estimator of the variance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{C}\{\hat{T}'_{x,\pi}, \hat{T}_{y,\pi}\}$ is a design consistent estimator of the covariance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{T}_{y,\pi}$.

The estimator of the total is $N\bar{y}_{\text{reg}}$ for simple random sampling, but the exact equivalence may not hold in more complicated samples, because in such situations the estimated mean may be a ratio estimator. However, if the regression estimator of the two totals is constructed using (4.34), the ratio of the two estimated totals has large sample variance equal to that of the regression estimator of the mean. To see this write the error in the regression estimated totals of y and u as

$$\hat{T}_{y,\text{reg}} - T_{y,N} = \hat{T}_{y,\pi} - T_{y,N} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\boldsymbol{\beta}_{y\cdot x,N} + O_p(Nn^{-1})$$

and

$$\hat{T}_{u,\text{reg}} - T_{u,N} = \hat{T}_{u,\pi} - T_{u,N} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\boldsymbol{\beta}_{u\cdot x,N} + O_p(Nn^{-1}), \quad (4.35)$$

where we are assuming $\hat{T}_{y,\pi} - T_{y,N}$, $\hat{\boldsymbol{\beta}}_{y\cdot x} - \boldsymbol{\beta}_{y\cdot x,N}$ and the corresponding quantities for u, to be $O_p(Nn^{-1/2})$ and $O_p(n^{-1/2})$, respectively. Then the error in $\hat{T}_{u,\text{reg}}^{-1}\hat{T}_{y,\text{reg}}$ is

$$\hat{T}_{u,\text{reg}}^{-1}\hat{T}_{y,\text{reg}} - T_{u,N}^{-1}T_{y,N} = T_{u,N}^{-1}\Big[\left(\hat{T}_{y,\pi} - T_{y,N}\right) - R_N\left(\hat{T}_{u,\pi} - T_{u,N}\right) + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\left(\boldsymbol{\beta}_{y\cdot x,N} - R_N\boldsymbol{\beta}_{z\cdot x,N}\right)\Big] + O_p(Nn^{-1}), \quad (4.36)$$

where $R_N = T_{u,N}^{-1} T_{y,N}$. If we construct the regression estimator for R_N starting with $\hat{R} = \hat{T}_{u,\pi}^{-1} \hat{T}_{y,\pi}$, we have

$$\hat{R}_{\text{reg}} = \hat{R} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\hat{\boldsymbol{\beta}}_{R\cdot x}, \tag{4.37}$$

where

$$\hat{\boldsymbol{\beta}}_{R\cdot x} = \left[\hat{V}\left\{\hat{\mathbf{T}}_{x,\pi}\right\}\right]^{-1} \hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}', \hat{R}\right\}$$

and

$$\hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}, \hat{R}\right\} = \hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}, T_{u,N}^{-1}\left(\hat{T}_{y,\pi} - R_N \hat{T}_{u,\pi}\right)\right\}.$$

It follows that the large-sample-design-optimum coefficient for the ratio is $T_{u,N}^{-1}(\boldsymbol{\beta}_{y\cdot x,N} - R_N \boldsymbol{\beta}_{u\cdot x,N})$ and the ratio of design-optimum regression estimators is the large sample design-optimum regression estimator of the ratio.

5. MODELS AND REGRESSION ESTIMATION

In this section we assume that the analyst postulates a detailed superpopulation model. Assume also that the sample is an unequal probability sample or (and) the specified error covariance structure is not a multiple of the identity matrix. Then, only in special cases will the design optimal estimator of (4.29) agree with the best estimator constructed under the model, conditioning on the sample x-values. To investigate this possible conflict, write the model for the population in matrix notation as

$$\begin{aligned} \mathbf{y}_U &= \mathbf{X}_U \boldsymbol{\beta} + \mathbf{e}_U \\ \mathbf{e}_U &\sim \left(\mathbf{0}, \boldsymbol{\Sigma}_{eeUU}\right), \end{aligned} \tag{5.1}$$

where $\mathbf{y}_U = (y_1, y_2, \cdots, y_N)'$, $\mathbf{e}_U = (e_1, e_2, \cdots, e_N)'$ and $\mathbf{X}_U = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_N')'$. It is assumed that $\boldsymbol{\Sigma}_{eeUU}$ is known or known up to a multiple. The model for a sample of n observations is

$$\begin{aligned} \mathbf{y}_A &= \mathbf{X}_A \boldsymbol{\beta} + \mathbf{e}_A, \\ \mathbf{e}_A &\sim \left(\mathbf{0}, \boldsymbol{\Sigma}_{eeAA}\right), \end{aligned}$$

where $\mathbf{y}_A = (y_1, y_2, \cdots, y_n)'$, $\mathbf{e}_A = (e_1, e_2, \cdots, e_n)'$, $\mathbf{X}_A = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_n')'$, and we index the sample elements by 1, 2, ..., n, for convenience. We have used the subscript U to identify population quantities, and the subscript A to identify sample quantities, but we will often omit the subscript A to simplify the notation. For example, we may sometimes write the $n \times n$ covariance matrix as $\boldsymbol{\Sigma}_{ee}$. The unknown finite population mean is

$$\bar{y}_N = \bar{\mathbf{x}}_N \boldsymbol{\beta} + \bar{e}_N. \tag{5.2}$$

Under model (5.1), the best linear, conditionally unbiased predictor of $\theta_N = \bar{y}_N$, conditional on $\mathbf{X}$ is

$$\begin{aligned} \hat{\theta} = N^{-1}\Bigg[&\sum_{i \in A} y_i + (N-n)\bar{\mathbf{x}}_{N-n}\hat{\boldsymbol{\beta}} \\ &+ \mathbf{J}_{N-n}' \boldsymbol{\Gamma}_{\bar{A}A}\left(\mathbf{y}_A - \mathbf{X}_A \hat{\boldsymbol{\beta}}\right)\Bigg], \end{aligned} \tag{5.3}$$

where $\boldsymbol{\Gamma}_{\bar{A}A} = \boldsymbol{\Sigma}_{ee\bar{A}A}\boldsymbol{\Sigma}_{eeAA}^{-1}$, $\bar{\mathbf{x}}_{N-n} = (N-n)^{-1}(N\bar{\mathbf{x}}_N - n\bar{\mathbf{x}}_n)$, $\boldsymbol{\Sigma}_{ee\bar{A}A} = E\{\mathbf{e}_{\bar{A}}\mathbf{e}_A'\}$,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Sigma}_{eeAA}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{eeAA}^{-1}\mathbf{y},$$

$\mathbf{e}_{\bar{A}} = (e_{n+1}, e_{n+2}, \cdots, e_N)$, $\mathbf{J}_{N-n}$ is an $N-n$ dimensional column vector of ones, $\bar{\mathbf{x}}_n$ is the simple sample mean, and $\bar{A}$ is the set of elements in U that are not in A. See Royall (1976). Under the model,

$$\hat{\theta} - \bar{y}_N = \mathbf{C}_{x\bar{A}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + N^{-1}\mathbf{J}_{N-n}'(\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{e}_A - \mathbf{e}_{\bar{A}})$$

and

$$\begin{aligned} V\left\{\hat{\theta} - \bar{y}_N | \mathbf{X}_A\right\} = &\mathbf{C}_{x\bar{A}} V\{\hat{\boldsymbol{\beta}}\}\mathbf{C}_{x\bar{A}}' \\ &+ N^{-2}\mathbf{J}_{N-n}'\left(\boldsymbol{\Sigma}_{ee\bar{A}\bar{A}} - \boldsymbol{\Gamma}_{\bar{A}A}\boldsymbol{\Sigma}_{eeA\bar{A}}\right)\mathbf{J}_{N-n}, \end{aligned} \tag{5.4}$$

where

$$\mathbf{C}_{x\bar{A}} = N^{-1}\left[(N-n)\bar{\mathbf{x}}_{N-n} - \mathbf{J}_{N-n}'\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{X}_A\right].$$

Design consistency of estimator (5.3) and the situations in which the model estimator reduces to the Horvitz-Thompson estimator have been considered by, among others, Isaki (1970), Royall (1970, 1976), Scott and Smith (1974), Cassel, Särndal, and Wretman (1976, 1979, 1983), Zyskind (1976), Tallis (1978), Isaki and Fuller (1982), Wright (1983), Pfefferman (1984), Tam (1986), Brewer, Hanif and Tam (1988), Montanari (1999), and Gerow and McCulloch (2000).

The estimator (5.3) reduces to $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ if there is an $\boldsymbol{\eta}$ such that

$$\mathbf{X}_A\boldsymbol{\eta} = \boldsymbol{\Sigma}_{eeAA}\mathbf{J}_n + \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n}, \tag{5.5}$$

for all samples with positive probability. If there is also $\boldsymbol{\gamma}$ such that

$$\mathbf{X}_A\boldsymbol{\gamma} = \boldsymbol{\Sigma}_{eeAA}\mathbf{D}_\pi^{-1}\mathbf{J}_n \tag{5.6}$$

for all samples with positive probability, then $\hat{\theta}$ of (5.3) is design consistent, where $\mathbf{D}_\pi$ was defined for (4.14). Given a $\mathbf{k}$ such that

$$\mathbf{X}_A\mathbf{k} = \boldsymbol{\Sigma}_{eeAA}\left(\mathbf{D}_\pi^{-1}\mathbf{J}_n - \mathbf{J}_n\right) - \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n}, \tag{5.7}$$

then $\hat{\theta}$ of (5.3) is expressible as

$$\hat{\theta} = \bar{y}_\pi + \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi\right)\hat{\boldsymbol{\beta}} \tag{5.8}$$

and if the design is such that $\bar{y}_\pi$ is design consistent for $\bar{y}_N$, $\hat{\theta}$ of (5.8) is design consistent for $\bar{y}_N$.

We call a regression model of the form (5.1) for which (5.5) and (5.6), or (5.7), holds a full model. If (5.6) or (5.7) does not hold, we call the model a reduced model or a restricted model. We cannot expect the conditions for a full model to hold for every analysis variable in a general purpose survey because $\boldsymbol{\Sigma}_{ee}$ will be different for different

y's. Therefore, given a reduced model, one might search for a good model estimator in the class of design consistent estimators.

To construct a design consistent estimator of the form $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ when model (5.1) is a reduced model, we can add a vector satisfying (5.7) to the $\mathbf{X}$-matrix to create a full model. There are two possible situations associated with this approach. In the first, the population mean (or total) of the added variable is known. With known mean, one can construct the usual regression estimator and the usual design variance estimation formulas are appropriate.

To describe an estimation procedure for the situation in which the population mean of the added variable is not known, let $\mathbf{q} = (q_1, q_2, \cdots, q_n)'$ denote the added vector, where $\mathbf{q}$ is the vector on the right side of the equality in (5.7). Let $\mathbf{H} = (\mathbf{X}, \mathbf{q})$, where $\mathbf{X}$ is the matrix of auxiliary variables with known population mean vector, $\bar{\mathbf{x}}_N$. We write the full model for the sample as

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta}_{y\cdot h} + \mathbf{e}, \tag{5.9}$$

where $\mathbf{e} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{ee})$. The best linear conditionally unbiased estimator of $\boldsymbol{\beta}_{y\cdot h}$ is

$$\hat{\boldsymbol{\beta}}_{y\cdot h} = \left(\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y}. \tag{5.10}$$

If the coefficient for $\mathbf{q}$ in (5.9) is not zero, it is not possible to construct a conditionally unbiased estimator of $\bar{\mathbf{h}}_N\boldsymbol{\beta}_{y\cdot h}$ because the $\bar{q}_N$ component of $\bar{\mathbf{h}}_N$ is unknown. However, because $\hat{\boldsymbol{\beta}}_{y\cdot h}$ is unbiased for $\boldsymbol{\beta}_{y\cdot h}$, it is possible to construct a conditionally unbiased estimator of any linear function of $\boldsymbol{\beta}_{y\cdot h}$. Thus, it is natural to replace the unknown $\bar{q}_N$ with the "best available" estimator of $\bar{q}_N$, and a reasonable choice is the regression estimator,

$$\bar{q}_{\text{reg}} = \bar{q}_\pi + \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi\right)\hat{\boldsymbol{\beta}}_{q\cdot x}, \tag{5.11}$$

where $\hat{\boldsymbol{\beta}}_{q\cdot x} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{q}$. Then the estimator (5.3) becomes

$$\hat{\theta} = \bar{y}_\pi + \left[\left(\bar{\mathbf{x}}_N, \bar{q}_{\text{reg}}\right) - \left(\bar{\mathbf{x}}_\pi, \bar{q}_\pi\right)\right]\hat{\boldsymbol{\beta}}_{y\cdot h} \tag{5.12}$$

The estimator (5.12) can be expressed in the familiar regression estimator form,

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{y\cdot x}. \tag{5.13}$$

That is, the regression estimator of the finite population mean of y based on the full model, but with the mean of q_i unknown and estimated with the regression estimator, is the regression estimator with $\boldsymbol{\beta}_{y\cdot x}$ estimated by the generalized least squares regression of y on $\mathbf{x}$ using the covariance matrix $\boldsymbol{\Sigma}_{ee}$. See Park (2002). The estimator is conditionally model unbiased under the reduced model containing only $\mathbf{x}$ if the reduced model is true. If the population coefficient for q_i is not zero, the reduced model is not true. Then the estimator is conditionally model biased, but the estimator is unbiased for the finite population mean under the full model and an unbiased design, because

$$E\left\{\bar{y}_{\text{reg}} - \bar{y}_N\right\} = E\left\{E\left[\bar{y}_{\text{reg}} - \bar{y}_N \mid \mathbf{H}\right]\right\}$$

$$= E\left\{(0, \bar{q}_{\text{reg}} - \bar{q}_N)\boldsymbol{\beta}_{y\cdot h}\right\} \doteq 0, \tag{5.14}$$

where $\bar{y}_{\text{reg}}$ is defined in (5.12) and the approximation is due to the approximate design expectation of the regression estimator $\bar{q}_{\text{reg}}$.

The estimator (5.13) is a linear estimator, where the vector of weights, $\mathbf{w}$, minimizes the Lagrangean

$$\mathbf{w}'\boldsymbol{\Sigma}_{ee}\,\mathbf{w} + \left[\mathbf{w}'\,\mathbf{H} - \left(\bar{\mathbf{x}}_N, \bar{q}_{\text{reg}}\right)\right]\boldsymbol{\lambda}. \tag{5.15}$$

The estimator is location invariant if the column of ones is in the column space of $\mathbf{X}$.

Because the variable q is the variable whose omission from the full model can produce a bias, it seems prudent to test the coefficient of q before using the reduced model to construct an estimator for the mean of y. This can be done using a model estimator of the variance,

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}_{y\cdot h} \mid \mathbf{H}\right\} = \left(\mathbf{H}'\,\hat{\boldsymbol{\Sigma}}_{ee}^{-1}\,\mathbf{H}\right)^{-1}$$

or using the design estimator of variance of (4.12). See Du Mouchel and Duncan (1983) and Fuller (1984).

A working specification for $\boldsymbol{\Sigma}_{ee}$ may be particularly appropriate for two-stage samples, see Royall (1976, 1986) and Montanari (1987). A reasonable model is that in which there is common correlation among items in the same primary sampling unit and zero correlation between units in different primary sampling units. Because the associated $\boldsymbol{\Sigma}_{ee}$ is block diagonal of a particular form, it is relatively easy to invert and hence the estimator based on such a working $\boldsymbol{\Phi}$ is relatively easy to construct. The regression estimator using a $\boldsymbol{\Phi}$ with a non zero correlation for units in the same primary sampling unit is a combination of the estimator based on primary sampling unit totals and that based on elements. See Fuller and Battese (1973). Thus, the use of such a $\boldsymbol{\Phi}$ can avoid variance problems associated with the use of primary sampling unit totals.

6. MAXIMUM LIKELIHOOD AND RAKING RATIO

The theoretical foundation for the regression estimators discussed in section 3 and section 4 is maximum likelihood estimation for the linear model with normal errors. We now consider the likelihood for multinomial variables. Given a simple random sample from a multinomial defined by the entries in a two way table, the logarithm of the likelihood, except for a constant, is

$$\sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij}\log p_{ij}, \tag{6.1}$$

where a_{ij} is the estimated fraction in cell ij, p_{ij} is the population fraction in cell ij, r is the number of rows, and c is the number of columns. If (6.1) is maximized subject to the restriction $\sum\sum p_{ij} = 1$, one obtains the maximum

likelihood estimators $\hat{p}_{ij} = a_{ij}$. If the marginal row fractions $p_{i\cdot,N}$ and the marginal column fractions $p_{\cdot j,N}$ are known, it is natural to maximize the likelihood subject to these constraints by using the Lagrangean

$$\sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij} \log p_{ij} + \sum_{i=1}^{r} \lambda_i \left(\sum_{j=1}^{c} p_{ij} - p_{i\cdot,N} \right) + \sum_{j=r+1}^{r+c} \lambda_j \left(\sum_{i=1}^{r} p_{ij} - p_{\cdot j,N} \right), \quad (6.2)$$

where λ_i, $i = 1, 2, ..., r$, are for the row restrictions and λ_j, $j = 1, 2, \cdots, c$, are for the column restrictions. There is no explicit expression for the solution to (6.2) and there may be no solution if there are too many empty cells. A procedure that produces estimates close to the maximum likelihood solution is that called *raking ratio* or *iterative proportional fitting*. The procedure iterates, first making ratio adjustments for the row restrictions, then making ratio adjustments for the column restrictions, then making a ratio adjustments for the row restrictions, *etc.* The method is generally credited to Deming and Stephan (1940). See, for example, Bishop, Fienberg and Holland (1975, Chapter 3).

Deville and Särndal (1992) considered a class of objective functions of the form $\sum_{i\in A} G(w_i, \alpha_i)$, where $G(w, \alpha)$ is a measure of distance between an initial weight α_i and a final weight w_i. The objective function is minimized subject to the constraints

$$\sum_{i\in A} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N. \quad (6.3)$$

Deville and Särndal (1992) used the term *calibrated* to describe weights satisfying (6.3). If the initial weight is $\alpha_i = (\sum \pi_j^{-1})^{-1} \pi_i^{-1}$ and if one is the first element of $\mathbf{x}_i$, the solution to the minimization problem is approximated by a regression estimator of the mean of the form

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}, \quad (6.4)$$

where

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i\in A} \mathbf{x}_i' \varphi_{ii}^{-1} \mathbf{x}_i\right]^{-1} \sum_{i\in A} \mathbf{x}_i' \varphi_{ii}^{-1} y_i,$$

and φ_{ii} is the second derivative of $G(w, \alpha)$ with respect to w evaluated at $(w, \alpha) = (\alpha_i, \alpha_i)$. Using this approach, Deville and Särndal (1992) showed that the maximum likelihood and raking ratio estimators have the same limiting distribution as the regression estimator (4.18) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$. To obtain the raking ratio weights they used the objective function

$$\sum_{i\in A} \left[w_i \log \alpha_i^{-1} w_i + \alpha_i - w_i \right], \quad (6.5)$$

and to obtain the maximum likelihood weights they used the objective function

$$\sum_{i\in A} \left[w_i - \alpha_i - \alpha_i \log \alpha_i^{-1} w_i \right]. \quad (6.6)$$

Deville, Särndal and Sautory (1993) investigated four estimators in the class. Although weights constructed using different functions could differ considerably, the authors concluded that estimates were quite similar, a result consistent with the theory. Singh and Mohl (1996) and Théberge (1999, 2000) discuss estimators with the calibration property.

7. POPULATION OF AUXILIARY VECTORS KNOWN AT ESTIMATION STEP

If the **x**-vector is known for all of the population elements, the number of possible regression-type estimators is greatly expanded. Most procedures involve the fitting of an approximating function for the relationship between y and the auxiliary variables. The most used procedure is to assign the population elements to categories on the basis of the auxiliary data and to use these categories as post strata. This procedure is equivalent to approximating the expected value of y given **x** by a step function. The estimator is formally equivalent to the regression estimator (4.19) where the **x**-vector is a vector of indicator variables for post-stratum membership.

The application of the procedure often requires the development of criteria to use in forming the post strata. Typically the post strata are formed so that each post stratum contains a minimum number of sample elements and so that the weights for any post stratum are not overly large. Estimation with post strata and the formation of post strata have been studied by Fuller (1966), Holt and Smith (1979), Tremblay (1986) Kalton and Maligalig (1991), Little (1993), Eltinge and Yansaneh (1997), and Lazzeroni and Little (1998), among others. Holt and Smith (1979) argued for the use of a conditional variance estimator for post stratification.

Given the population of **x**-vectors, one can use the sample to estimate a functional relationship between y and **x** and then predict the unobserved y. If the procedure is to be design consistent, then a condition similar to (4.14) must hold. One way to ensure design consistency is to require the fitted model to satisfy

$$\sum_{i\in A} \pi_i^{-1} \left[y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \right] = 0, \quad (7.1)$$

where $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ is the model estimated value for the i-th observation.

Firth and Bennett (1998) pointed out that some nonlinear models satisfy (7.1). If the initial model does not satisfy (7.1), an estimated intercept term can be added to create an expanded full model,

$$\tilde{f}_F(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) = f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) + \left(\sum_{i\in A} \pi_i^{-1} \right)^{-1} \sum_{i\in A} \pi_i^{-1} \left[y_i - f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \right].$$

This is a direct extension of the ideas of difference estimation to the nonlinear case. See Isaki (1970), Cassel, Särndal and Wretman (1976) and Wright (1983). A closely related approach was suggested by Wu and Sitter (2001) in which the fitted function $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ is used as the auxiliary variable in a linear regression estimator.

A number of "local" procedures, other than step functions, can be used to approximate the functional relationship between $\mathbf{x}$ and y. Spline functions and polynomials are linear models that fall within the class of section 4. Estimators that use some kind of local smoothing to estimate population quantities have been considered for finite populations from a model viewpoint by Kuo (1988), Dorfman (1993), Dorfman and Hall (1993), Chambers (1996), and Chambers, Dorfman and Wehrly (1993). Breidt and Opsomer (2000) showed that estimators based on local polynomial regression are design consistent. Firth and Bennett (1998) also considered local fit models.

8. REGRESSION ESTIMATION AND NONRESPONSE

Regression estimation is frequently a part of procedures used to adjust data for unit nonresponse. Regression can be justified on the basis of a model such as (3.1) or on the basis that regression can adjust for unequal response probabilities. See Cassel, Särndal and Wretman (1979, 1983), Little (1982, 1986), Bethlehem (1988), Kott (1994), Fuller, Loughin and Baker (1994) and Fuller and An (1998).

Consider an estimator of the population regression vector of the form (4.4) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$ constructed with the responding units. Denote the estimator by $\tilde{\boldsymbol{\beta}}$ and let p_i be the conditional probability of observing unit i given that the unit is selected for the sample. Then under regularity conditions, the estimator $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of

$$\boldsymbol{\gamma}_N = \left(\sum_{i \in U} \mathbf{x}_i' p_i \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' p_i y_i. \tag{8.1}$$

The population mean of y can be expressed as

$$\bar{y}_N = \bar{\mathbf{x}}_N \boldsymbol{\gamma}_N + \bar{a}_N \tag{8.2}$$

where $a_i = y_i - \mathbf{x}_i \boldsymbol{\gamma}_N$ and $\bar{a}_N$ is the population mean of the a_i. The regression estimator $\bar{y}_{reg} = \bar{\mathbf{x}}_N \tilde{\boldsymbol{\beta}}$ will be consistent for $\bar{y}_N$ if the probability limit of $\bar{a}_N$ is zero. The probability limit of $\bar{a}_N$ will be zero if the sequence of finite populations is a sequence of random samples from an infinite population in which

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + e_i, \tag{8.3}$$

and the e_i of the sample are independent of $\mathbf{x}_i$ with $E\{e_i | \mathbf{x}_i\} = 0$.

Alternatively, a sufficient condition for $\bar{a}_N$ to be zero is the existence of a column vector $\boldsymbol{\xi}$ such that

$$\mathbf{x}_i \boldsymbol{\xi} = p_i^{-1} \tag{8.4}$$

for $i = 1, 2, \cdots, N$. Thus, if the reciprocal of the response probability is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y. One way in which (8.4) can be satisfied is for the elements of $\mathbf{x}_i$ to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup.

If (8.4) holds and if the probability of responding is independent from unit to unit, then the estimated variance based on (4.12) is an appropriate estimator for the variance of the regression estimator of the mean. It is particularly important that a variance estimator of the form (4.12) or (4.25), and not of the form (4.26) be used, because $\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi$ is, in general, not $O_p(n^{-1/2})$ in the presence of nonresponse. Singh and Folsom (2000) make a similar argument for the variance estimator (4.25) when using regression to adjust for coverage error.

Often a preliminary adjustment to the selection probabilities is made for nonresponse and this is followed by regression estimation. The most frequently used response adjustment is to form adjustment cells (post strata) and to ratio adjust the weights of respondents in the cell so that the sum of the weights is equal to the estimated (or known) total for the cell. See, for example, Little and Rubin (1987, page 250). Procedures using an estimated response probability function are discussed by Cassel, Särndal and Wretman (1983), Rosenbaum and Rubin (1983), Folsom and Witt (1994). Fuller and An (1998), and Folsom and Singh (2000). Brick, Waksberg and Keeter (1996) use an estimated contact probability to adjust for frame coverage.

To consider procedures based on estimated response probabilities, assume that the inverse of the response probability for individual i is given by

$$p_i^{-1} = g(\mathbf{z}_i; \boldsymbol{\theta}^0), \tag{8.5}$$

where $\mathbf{z}_i$ is a vector of variables that can be observed for both respondents and nonrespondents, $\boldsymbol{\theta}^0$ is the true value of $\boldsymbol{\theta}$, and $g(\mathbf{z}_i; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ with continuous first and second derivatives in an open set containing $\boldsymbol{\theta}^0$ for all $\mathbf{z}_i$. The vector $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ is observed, and we assume that p_i is bounded below by a positive number.

Let δ_i be the indicator variable with $\delta_i = 1$ if a response is obtained and $\delta_i = 0$ if a response is not obtained. Using the vector $(\delta_i, \mathbf{z}_i)$, the parameter $\boldsymbol{\theta}^0$ of the response probability function is estimated. Assume that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O_p(n^{-1/2})$, where $\hat{\boldsymbol{\theta}}$ is the estimator of $\boldsymbol{\theta}$. Let $\boldsymbol{\beta}_N$ denote the finite population regression vector for the regression of y on $\mathbf{x}$. Let

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' y_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i, \tag{8.6}$$

where π_i are the selection probabilities and $\hat{p}_i^{-1} = g(\mathbf{z}_i; \hat{\boldsymbol{\theta}})$. Under conditions of the type used in section 4,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = \mathbf{M}_{xx}^{-1} \sum_{i \in A} \delta_i \pi_i^{-1} p_i^{-1} \mathbf{x}_i' a_i \left[1 + p_i \mathbf{g}_{1,i} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \right) \right] + O_p(n^{-1}),$$

where $\mathbf{g}_{1,i}$ is the row vector of first derivatives of $g(\mathbf{z}_i;\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ and $\mathbf{M}_{xx} = \sum_{i\in A} \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} p_i^{-1} \delta_i$. If $\mathbf{g}_{1,i}$ is uncorrelated with a_i, then the term involving $\mathbf{g}_{1,i} a_i$ is $O_p(n^{-1})$ and the variance estimator constructed as if $g(\mathbf{z};\boldsymbol{\theta}^0)$ is known is appropriate. The conditions are satisfied if $\mathbf{z}_i$ is a subvector of $\mathbf{x}_i$ and $\mathbf{z}_i$ defines imputation cells (adjustment cells) with equal response rates within a cell.

9. PRACTICAL CONSIDERATIONS

If the regression weights are to be used in a general purpose survey, no individual weight used in estimating a total should be less than one. Also, it seems reasonable, on robustness grounds, to avoid very large weights. We discuss some procedures that have been developed to accomplish these objectives.

A number of algorithms produce positive weights with a high probability. Raking ratio procedures produces positive weights for most data configurations. Deville, Särndal and Sautory (1993) discuss the extension of raking ratio to general x-variables and extensions to include bounds on the weights.

Tillé (1998) suggested the use of approximate conditional probabilities, conditional on $\bar{\mathbf{x}}_\pi$, to compute an estimator. His approximation can be extended to produce regression weights that are positive with high probability. Let $\bar{\mathbf{x}}_\pi^{(i)}$ be an estimator obtained by deleting element i, or primary sampling unit i, and modifying the remaining weights so that $\bar{\mathbf{x}}_\pi^{(i)}$ is unbiased, or consistent to the same order as $\bar{\mathbf{x}}_\pi$, for the population mean of all elements excluding i. The estimator $\bar{\mathbf{x}}_\pi^{(i)}$ can be the estimator used to construct jackknife deviates. Let $\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}$ be an estimator of the covariance matrix of $\bar{\mathbf{x}}_\pi$ and let $\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}(i)}$ be an estimator of the conditional covariance matrix of $\bar{\mathbf{x}}_\pi$ conditional on $i \in A$. Then, in large samples $\bar{\mathbf{x}}_\pi$ and $\bar{\mathbf{x}}_\pi^{(i)}$ are approximately normally distributed and an estimator of the probability that i is in the sample given the estimated mean $\bar{\mathbf{x}}_\pi$, is

$$\begin{aligned}\hat{\pi}_{i|A} &= \hat{P}\left\{i \in A \mid \mathbf{F}_N, \bar{\mathbf{x}}_\pi\right\} \\ &= \pi_i \left|\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}\right|^{1/2} \left|\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}(i)}\right|^{-1/2} \exp\left\{0.5\left(\mathbf{G}_{\bar{x}\bar{x}} - \mathbf{G}_{\bar{x}\bar{x}(i)}\right)\right\} \quad (9.1)\end{aligned}$$

where

$$\mathbf{G}_{\bar{x}\bar{x}} = \left(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N\right) \hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}^{-1} \left(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N\right)',$$

$$\mathbf{G}_{\bar{x}\bar{x}(i)} = \left(\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N^{(i)}\right) \hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}(i)}^{-1} \left(\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N^{(i)}\right)',$$

and $\bar{\mathbf{x}}_N^{(i)} = (N-1)^{-1}(N\bar{\mathbf{x}}_N - \mathbf{x}_i)$. For simple random sampling, Tillé (1998) showed that the estimator

$$\bar{y}_{p\pi} = N^{-1} \sum_{i\in A} \pi_{i|A}^{-1} y_i, \quad (9.2)$$

where $\pi_{i|A}$ is the conditional probability calculated under the normality assumptions, is approximately equal to the regression estimator. Because the estimator is not calibrated, we suggest a calibrated version obtained by computing the regression estimator with $\hat{\pi}_{i|A}$ as initial weights. The difference between (9.2) and the regression estimator constructed with initial weights $\hat{\pi}_{i|A}$ is $O_p(n^{-1})$. Hence, there is a good chance that the regression weights so constructed will be positive. The variance estimator $\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}(i)}$ is relatively simple to compute for stratified samples but may require considerable computation for other cases. Thus one may choose to approximate $\boldsymbol{\Sigma}_{\bar{x}\bar{x}(i)}$.

Given that the regression weights are being constructed by minimizing an objective function, one can add restrictions to the problem to place bounds on the weights. Huang and Fuller (1978) gave an interative procedure equivalent to constructing a $\boldsymbol{\Phi}$ matrix at each step that reduces the weight on observations whose current weight deviates from the average by a large absolute amount.

To discuss additional procedures associated with quadratic objective functions, assume we have a working covariance matrix, denoted by $\boldsymbol{\Phi}_{ee}$, for the model (5.1) that is to be used to construct a regression estimator. Let $\boldsymbol{\alpha}$ be the column vector of initial weights and assume $\boldsymbol{\Phi}_{ee}\boldsymbol{\alpha}$ is in the column space of $\mathbf{X}$. Then the weights that minimize the conditional model variance are the weights that minimize $\mathbf{w}'\boldsymbol{\Phi}_{ee}\mathbf{w}$ or, equivalently, that minimize

$$(\mathbf{w} - \boldsymbol{\alpha})' \boldsymbol{\Phi}_{ee} (\mathbf{w} - \boldsymbol{\alpha}) \quad (9.3)$$

subject to the constraint

$$\mathbf{w}'\mathbf{X} = \bar{\mathbf{x}}_N. \quad (9.4)$$

Given an objective function, we can add restrictions on the w_i such as

$$L_1 \le w_i \le L_2, \quad i \in A, \quad (9.5)$$

where L_1 and L_2 are nonnegative constants. Minimizing (9.3), subject to the constraints (9.4) and (9.5) is a quadratic programming problem. The use of quadratic programming was suggested by Husain (1969) and was used by Isaki, Tsay and Fuller (2000).

If a large number of control variables are used, it may not be possible to construct weights satisfying the calibration constraints and also falling within reasonable bounds. The practitioner is faced with making compromises. The most common practice is to drop variables from the model. See Bankier, Rathwell and Maijkowski (1992) and Silva and Skinner (1997). To discuss an alternative procedure, consider the situation in which some of the constraints are required but others can be relaxed. Let the matrix of observations on the auxiliary variables be partitioned as $(\mathbf{X}_0, \mathbf{X}_2)$, where $\mathbf{X}_0$ is the set of variables for which exact constraints are required and $\mathbf{X}_2$ is the set for which the constraints can be relaxed. Assume $\boldsymbol{\Phi}_{ee}\boldsymbol{\alpha}$ is in the column space of $\mathbf{X}_0$. Then a generalization of (9.3) and (9.4) is the function

$$(\mathbf{w} - \boldsymbol{\alpha})' \boldsymbol{\Phi}_{ee} (\mathbf{w} - \boldsymbol{\alpha}) + (\mathbf{w}'\mathbf{X}_2 - \bar{\mathbf{x}}_{2,N}) \boldsymbol{\Psi} (\mathbf{w}'\mathbf{X}_2 - \bar{\mathbf{x}}_{2,N})' \quad (9.6)$$

and the constraint

$$\mathbf{w}'\mathbf{X}_0 - \bar{\mathbf{x}}_{0,N} = \mathbf{0}, \tag{9.7}$$

where $\boldsymbol{\Phi}_{ee}$ and $\boldsymbol{\Psi}$ are positive definite symmetric matrices and $\bar{\mathbf{x}}_N = (\bar{\mathbf{x}}_{0,N}, \bar{\mathbf{x}}_{2,N})$. The $\mathbf{w}$ that minimizes (9.6) subject to (9.7) minimizes the mean squared error of the unbiased linear predictor of $\bar{\mathbf{x}}_N\boldsymbol{\beta}$ under the mixed model

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e},$$

where $\boldsymbol{\beta}_2 \sim (\mathbf{0}, \boldsymbol{\Psi})$, $e \sim (\mathbf{0}, \boldsymbol{\Phi}_{ee})$, the random vector $\boldsymbol{\beta}_2$ is independent of $\mathbf{e}$, and $\boldsymbol{\beta}_0$ is a fixed vector. See Lazzeroni and Little (1998) for the use of random models for post stratification.

The vector $\mathbf{w}'$ that minimizes (9.6) subject to restriction (9.7) is

$$\mathbf{w}' = \boldsymbol{\alpha}' + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\,\mathbf{H}_{x\psi x}^{-1}\mathbf{X}'\boldsymbol{\Phi}_{ee}^{-1}, \tag{9.8}$$

where

$$\mathbf{H}_{x\psi x} = \begin{pmatrix} \mathbf{X}_0'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_0 & \mathbf{X}_0'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_0 & \boldsymbol{\Psi}^{-1} + \mathbf{X}_2'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_2 \end{pmatrix}. \tag{9.9}$$

The estimator can be written

$$\bar{y}_{rreg} = \mathbf{w}'\mathbf{y} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\,\hat{\boldsymbol{\theta}}, \tag{9.10}$$

where $\hat{\boldsymbol{\theta}} = \mathbf{H}_{x\psi y}^{-1}\mathbf{X}'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{y}$. See Henderson (1963), Robinson (1991), and Rao (2002, Chapter 6).

Husain (1969) considered (9.6) for a simple random sample from a normal distribution with $\mathbf{X}_0 = \mathbf{J}$, $\boldsymbol{\Phi}_{ee} = \mathbf{I}$, and $\boldsymbol{\Psi}^{-1} = \gamma^{-1}\hat{\boldsymbol{\Sigma}}_{x,22}$, where $\hat{\boldsymbol{\Sigma}}_{x,22}$ is the estimated covariance matrix of $\bar{\mathbf{x}}_{2,\pi}$, and γ is a constant to be determined. For this case, Husain showed that the optimal γ is

$$\gamma_{opt} = \left[k_2(1 - R^2)\right]^{-1}(n - k_2 - 2)\,R^2, \tag{9.11}$$

where k_2 is the dimension of $\mathbf{x}_2$ and R^2 is the squared multiple correlation coefficient. Bardsley and Chambers (1984) considered the function (9.6), the division of $\mathbf{x}_i$ into two components, and studied the behavior of the estimator from a model perspective. The procedure associated with (9.5), (9.6) and (9.7) was used by Isaki, Tsay and Fuller (2000). In that application, the vector $\bar{\mathbf{x}}_{0,N}$ contained marginal totals of a multiway table and $\bar{\mathbf{x}}_{2,N}$ contained totals for interior cells. Rao and Singh (1997) studied a closely related estimator in which tolerances are given for the difference between the final estimates for elements of $\bar{\mathbf{x}}_{2,N}$ and the corresponding elements of $\bar{\mathbf{x}}_{2,N}$.

Park (2002) extended Husain's optimality results to a more general $\boldsymbol{\Psi}$. The $\mathbf{x}_2$ vector can be transformed so that $\hat{V}\{\bar{\mathbf{x}}_{2,\pi}\}$ for the transformed vector is a diagonal matrix and so that $\tilde{\mathbf{X}}_2'\boldsymbol{\Phi}_{ee}^{-1}\tilde{\mathbf{X}}_2$ is a diagonal matrix, where $\tilde{\mathbf{X}}_2$ is the part of $\mathbf{X}_2$ that is orthogonal to $\mathbf{X}_0$ in the metric $\boldsymbol{\Phi}_{ee}$. That is,

$$\tilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_0\left(\mathbf{X}_0'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\boldsymbol{\Phi}_{ee}^{-1}\mathbf{X}_2.$$

Then the diagonal $\boldsymbol{\Psi}$ that minimizes the approximate variance has elements

$$\psi_{ii} = (m_{ii}\,V_{\beta\beta ii})^{-1}\,\beta_i^2, \tag{9.12}$$

where m_{ii} is the ith element of the diagonal matrix $\tilde{\mathbf{X}}_2'\boldsymbol{\Phi}_{ee}^{-1}\tilde{\mathbf{X}}_2$ and $V_{\beta\beta ii}$ is the variance of $\hat{\beta}_i$ in the transformed scale. To implement the procedure one must estimate the population parameters or choose realistic values for a general purpose $\boldsymbol{\Psi}$. If one postulates a super-population random model for $\boldsymbol{\beta}$, then the β_i^2 of (9.12) is replaced with $E\{\beta_i^2\}$, where the expectation is the model expectation.

10. COMMENTS

Regression estimation is a flexible and powerful tool for the incorporation of auxiliary information into the estimation process. Closely related procedures, such as raking ratio, have large sample properties equivalent to those of regression estimators. The linearity of such estimators is of paramount importance because it permits the construction of a general purpose data set that provides very good estimators for a wide range of parameters.

Given a concentrated interest in a single y-variable, efficiency gains may be possible by postulating a particular set of auxiliary variables and a particular error covariance matrix. Because of the simple nature of the design consistency requirement, it is easy to test such models for design consistency.

ACKNOWLEDGEMENTS

This research was partially supported by Cooperative Agreement 43-3AEU-0-80064 between Iowa State University, the U. S. National Agricultural Statistics Service and the U. S. Bureau of the Census. I am deeply indebted to Mingue Park for assistance in literature review, for comments on and repair of theoretical results, and for use of material from his thesis. I thank Michael Hidiroglou, J.N.K. Rao, Harold Mantel, and Jean Opsomer for useful comments on drafts of the manuscript. I thank the Associate Editor for numerous comments that improved the presentation.

APPENDIX

This appendix contains theorems supporting the limiting properties of the regression estimators discussed in section 4.

Theorem A.1. Let $\{U_N, \mathrm{F}_N, A_N, n_N: N = k + 3, k + 4, \ldots\}$ be a sequence of finite populations and samples, where $\mathbf{F}_N$ is a sample from an infinite population with eighth moments, A_N is the sample of size n_N selected from the Nth population. Let $\hat{\boldsymbol{\beta}}$ be defined by (4.4) of the text, and let

$$\hat{\mathbf{Q}}_{zz} = n^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z},$$

where $\boldsymbol{\Phi}$ is a positive definite symmetric $n \times n$ matrix that may be a function of $\mathbf{X}$ but not of y, $\mathbf{Z}$ is defined following (4.2), and we omit the subscript N on sample quantities. Assume $\hat{\mathbf{Q}}_{zz}$ is positive definite with probability one. If $\boldsymbol{\Phi}$ is random, assume the rows of $\boldsymbol{\Phi}^{-1}\mathbf{Z}$ have bounded fourth moments. Assume the selection probabilities satisfy

$$0 < K_1 < Nn^{-1}\pi_i < K_2,$$

where π_i are the selection probabilities. Assume the sample design is such that for any $\mathbf{z}$ with bounded fourth moments

$$\left[(\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N)', (\hat{\mathbf{Q}}_{zz} - \mathbf{Q}_{zzN})\right] \mid \mathbf{F}_N = O_p(n^{-1/2}), \quad \text{(A.1)}$$

where

$$\bar{\mathbf{z}}_{HT} = (\bar{y}_{HT}, \bar{\mathbf{x}}_{HT}) = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{z}_i, \quad \text{(A.2)}$$

$\mathbf{Q}_{zzN} = E\{\hat{\mathbf{Q}}_{zz} \mid \mathbf{F}_N\}$, $\bar{\mathbf{z}}_N$ is the finite population mean of $\mathbf{z}$, $\mathbf{Q}_{zzN}$ is a positive definite matrix for the Nth population, and the limit of $\mathbf{Q}_{zzN}$ is positive definite. Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1}\bar{\mathbf{b}}_{HT}' + O_p(n^{-1}), \quad \text{(A.3)}$$

where $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1}\mathbf{Q}_{xyN}$, $\bar{\mathbf{b}}_{HT} = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i$, $\mathbf{b}_i' = n^{-1}N\pi_i\zeta_i' e_i$,

$$\mathbf{Q}_{zzN} = \begin{pmatrix} Q_{yyN} & \mathbf{Q}_{yxN} \\ \mathbf{Q}_{xyN} & \mathbf{Q}_{xxN} \end{pmatrix}, \quad \text{(A.4)}$$

$e_i = y_i - \mathbf{x}_i\boldsymbol{\beta}_N$, and ζ_i' is column i of $\mathbf{X}'\boldsymbol{\Phi}^{-1}$. Assume the design is such that

$$\mathbf{V}_{\bar{z}\bar{z}}^{-1/2}\left\{\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N \mid \mathbf{F}_N\right\} \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad \text{(A.5)}$$

as $n_N \to \infty$ for any $\mathbf{z}$ with finite fourth moments, where $\mathbf{V}_{\bar{z}\bar{z}}$ is the covariance matrix of $\bar{\mathbf{z}}_{HT} - \bar{\mathbf{z}}_N$. Assume that $\mathbf{V}_{\bar{z}\bar{z}}$ is $O(n^{-1})$ and that the design admits an estimator $\hat{\mathbf{V}}_{\bar{z}\bar{z}}$ such that

$$n\left(\hat{\mathbf{V}}_{\bar{z}\bar{z}} - \mathbf{V}_{\bar{z}\bar{z}}\right) \mid \mathbf{F}_N = o_p(1) \quad \text{(A.6)}$$

for any $\mathbf{z}$ with bounded fourth moments. Then

$$[\hat{V}\{\hat{\boldsymbol{\beta}}\}]^{-1/2}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N] \mid \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad \text{(A.7)}$$

where

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1}\hat{\mathbf{V}}_{\bar{b}\bar{b}}\hat{\mathbf{Q}}_{xx}^{-1}, \quad \text{(A.8)}$$

$\hat{\mathbf{V}}_{\bar{b}\bar{b}} = \hat{V}\{\bar{\mathbf{b}}_{HT}'\}$ is the estimated design variance of $\bar{\mathbf{b}}_{HT}'$ calculated with $\hat{\mathbf{b}}_i' = n^{-1}N\pi_i\zeta_i'\hat{e}_i$ and $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$.

Proof. The error in $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = \left(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y} - \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\boldsymbol{\beta}_N\right]$$
$$= \hat{\mathbf{Q}}_{xx}^{-1}(n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e}).$$

Now $\hat{\boldsymbol{\beta}}$ is a generalized least squares estimator. Therefore

$$\hat{\mathbf{e}}'\boldsymbol{\Phi}^{-1}\mathbf{X} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Phi}^{-1}\mathbf{X} = \mathbf{0}$$

and $\mathbf{Q}_{xyN} - \boldsymbol{\beta}_N'\mathbf{Q}_{xxN} = \mathbf{Q}_{exN} = \mathbf{0}$. By assumption (A.1)

$$\hat{\mathbf{Q}}_{ex}' = n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e} = O_p(n^{-1/2}).$$

Thus

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1}\left(n^{-1}\sum_{i\in A}\zeta_i' e_i\right) + O_p(n^{-1})$$
$$= \mathbf{Q}_{xxN}^{-1}\left(N^{-1}\sum_{i\in A}\pi_i'\mathbf{b}_i'\right) + O_p(n^{-1}).$$

The $\mathbf{b}_i$ have bounded fourth moments by the assumptions. Thus, by assumption (A.5)

$$\mathbf{V}_{\beta\beta}^{-1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}),$$

where

$$\mathbf{V}_{\beta\beta} = \mathbf{Q}_{xxN}^{-1}\mathbf{V}_{\bar{b}\bar{b}}\mathbf{Q}_{xxN}^{-1}$$

and $\mathbf{V}_{\bar{b}\bar{b}} = V\{\bar{\mathbf{b}}_{HT}\}$. Now

$$n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\hat{\mathbf{e}} = n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e} + n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}\right)$$
$$=: N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i' + N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i',$$

where

$$\mathbf{h}_i' = n^{-1}N\pi_i\zeta_i'\mathbf{x}_i\boldsymbol{\delta}_\beta$$

and $\boldsymbol{\delta}_\beta = \boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}$. For any fixed $\boldsymbol{\delta}$, by (A.6), the estimated variance of $N^{-1}\sum_{i\in A}\pi_i^{-1}(\mathbf{b}_i' + \mathbf{h}_i')$ is consistent for the variance of the estimator of the mean of $\mathbf{b}+\mathbf{h}$. By assumption, the elements of $\zeta_i'\mathbf{x}_i$ have fourth moments. For a fixed $\boldsymbol{\delta}$ the variance of $\bar{\mathbf{h}}_{HT}$ is $O(n^{-1})$. For $\boldsymbol{\delta} = \boldsymbol{\delta}_\beta$,

$$\hat{V}\{\bar{\mathbf{h}}_{HT}'\} = o_p(n^{-1}),$$

and

$$\hat{V}\{\bar{\mathbf{b}}_{HT}'\} = V\{\bar{\mathbf{b}}_{HT}'\} + o_p(n^{-1})$$

because $\boldsymbol{\delta}_\beta = O_p(n^{-1/2})$. Result (A.7) then follows from the asymptotic normality of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N$.

Theorem A.2. Let $\mathbf{y}' = (y_1, y_2, \cdots, y_n)$ and $\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_1', \cdots, \mathbf{x}_n')$. Let $\boldsymbol{\Phi}$ be a nonsingular symmetric $n \times n$ matrix and let $\boldsymbol{\Phi}_N$ be a nonsingular symmetric $N \times N$ matrix. Let

$$\bar{y}_\pi, \bar{\mathbf{x}}_\pi, n^{-1}(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}) \text{ and } n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}$$

be design consistent estimators for finite population characteristics $\bar{y}_N, \bar{\mathbf{x}}_N, \mathbf{Q}_{xxN}$ and $\mathbf{Q}_{xyN}$, respectively, where

$$\left[\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}\right] = \left[N^{-1}\mathbf{X}_N'\boldsymbol{\Phi}_N^{-1}\mathbf{X}_N, N^{-1}\mathbf{X}_N\boldsymbol{\Phi}_N^{-1}\mathbf{y}_N\right]. \quad (A.9)$$

Let $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1}\mathbf{Q}_{xyN}$. Let there be a sequence of column vectors $\{\boldsymbol{\gamma}_N\}$ such that

$$\mathbf{X}\boldsymbol{\gamma}_N = \boldsymbol{\Phi}\mathbf{D}_\pi^{-1}\mathbf{J}_n \quad (A.10)$$

for all possible samples, where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \cdots, \pi_n)$ and $\mathbf{J}_n$ is an n-dimensional column vector of ones. Then, the regression estimator $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ with

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}, \quad (A.11)$$

is a design consistent estimator of $\bar{y}_N$.

Proof. If $\hat{\boldsymbol{\beta}}$ is defined by (A.11), then by the properties of generalized least squares estimators,

$$\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\Phi}^{-1}\mathbf{X} = \mathbf{0}.$$

If (A.10) holds, then

$$\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\mathbf{D}_\pi^{-1}\mathbf{J} = \left(\sum_{i\in A}\pi_i^{-1}\right)\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\hat{\boldsymbol{\beta}}\right) = 0.$$

It follows that $\bar{y}_{\text{reg}}$ is design consistent because

$$\begin{aligned} 0 &= p\lim_{N\to\infty}\left\{\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\hat{\boldsymbol{\beta}}_n\right)\middle|\,\mathbf{F}_N\right\} \\ &= p\lim_{N\to\infty}\left\{\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\boldsymbol{\beta}_N\right)\middle|\,\mathbf{F}_N\right\} \\ &= p\lim_{N\to\infty}\left\{\left(\bar{y}_N - \bar{\mathbf{x}}_N\boldsymbol{\beta}_N\right)\middle|\,\mathbf{F}_N\right\}. \end{aligned}$$

Theorem A.3. Let a sequence of populations and samples be as defined in Theorem A.1. Let $\mathbf{z}_i$ be a vector of the form $\mathbf{z}_i = (y_i, 1, \mathbf{x}_{1,i})$ and let $\mathbf{z}_{1,i} = (y_i, \mathbf{x}_{1,i})$. Assume $\bar{\mathbf{z}}_{1,\pi}$ is a design consistent estimator of the population mean $\bar{\mathbf{z}}_{1,N}$ with nonsingular covariance matrix

$$V\{\bar{\mathbf{z}}_{1,\pi}\,|\,\mathbf{F}_N\} = O(n^{-1}) \quad (A.12)$$

and

$$n^{1/2}(\bar{\mathbf{z}}_{1,\pi} - \bar{\mathbf{z}}_{1,N})\,|\,\mathbf{F}_N \xrightarrow{L} N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\bar{z}\bar{z}}\right), \quad (A.13)$$

where $\boldsymbol{\Sigma}_{\bar{z}\bar{z}}$ is the limit of $n\,V\{\bar{\mathbf{z}}_{1,\pi}\,|\,\mathbf{F}_N\}$. Assume there is an estimator of the variance of $\bar{\mathbf{z}}_{1,\pi}$, denoted by $\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\}$, such that

$$p\lim_{N\to\infty} n^{1+\delta}\left(\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\} - V\{\bar{\mathbf{z}}_{1,\pi}\,|\,\mathbf{F}_N\}\right) = \mathbf{0} \quad (A.14)$$

for some $\delta > 0$. Let $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$ be the vector that minimizes

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}_{1,d}\} \quad (A.15)$$

and let $\boldsymbol{\beta}_{1,\text{dopt}}$ be the vector that minimizes $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}_{1,d}\}$. Let $\bar{y}_{d,\text{reg}}$ be defined by (4.29). Then $\bar{y}_{d,\text{reg}}$ has the minimum limit variance for design consistent estimators of the form $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\boldsymbol{\beta}_{1,d}$. Also

$$\left[\hat{V}\{\bar{e}_\pi\}\right]^{-1/2}(\bar{y}_{d,\text{reg}} - \bar{y}_N) \xrightarrow{L} N(0,1), \quad (A.16)$$

where $\hat{V}\{\bar{e}_\pi\}$ is the estimator of (A.14) constructed with $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$.

Proof. The estimator

$$\hat{\boldsymbol{\beta}}_{1,\text{dopt}} = \left[\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\}\right]^{-1}\hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\}$$

minimizes the estimated variance of (A.15), and, by assumption (A.14), the estimated variance is consistent for the true variance. Hence, $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$ is design consistent for $\boldsymbol{\beta}_{1,\text{dopt}}$ and $\boldsymbol{\beta}_{1,\text{dopt}}$ minimizes $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}\}$. Therefore, no estimator of the form (4.29) has a limit distribution with smaller variance.

Now

$$\begin{aligned} \bar{y}_{d,\text{reg}} - \bar{y}_N &= \bar{y}_\pi - \bar{y}_N - (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\text{dopt}} \\ &= \bar{e}_\pi + o_p(n^{-1/2}), \end{aligned}$$

where $e_i = y_i - \bar{y}_N - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\boldsymbol{\beta}_{1,\text{dopt}}$. Therefore the variance of the limiting distribution of $n^{1/2}(\bar{y}_{d,\text{reg}} - \bar{y}_N)$ is the variance of $n^{1/2}(\bar{e}_\pi - \bar{e}_N)$. By assumption (A.14), the estimator $\hat{V}\{\bar{\mathbf{z}}_\pi\boldsymbol{\gamma}\}$ is a consistent variance estimator of $V\{\bar{\mathbf{z}}_\pi\boldsymbol{\gamma}\}$ for any fixed $\boldsymbol{\gamma}$. Because $\hat{\boldsymbol{\beta}}_{1,\text{dopt}} - \boldsymbol{\beta}_{1,\text{dopt}} = o_p(1)$, the estimated variance based on $\hat{e}_i$ converges to the estimated variance based on e_i and (A.16) holds.

REFERENCES

ANDERSON, C., and NORDBERG, L. (1998). A user's guide to CLAN97. Statistics Sweden, Orebro, Sweden.

BANKIER, M.D., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working Paper-Methodology Branch, Census Operations Section, Social Survey Methods Division. Statistics Canada, Ottawa.

BANKIER, M.D., HOULE, A.M. and LUC, M. (1997). Calibration estimation in the 1991 and 1996 Canadian census. Statistics Canada (draft), 8 pages.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

BREWER, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

BREWER, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W., HANIF, M. and TAM, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.

BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.

CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1979). Prediction theory for finite populations when model-based and design-based principles are combined. *Scandinavian Journal of Statistics*, 6, 97-106.

CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, (Eds. W.G. Madow, I. Olkin, and D. Rubin). New York: Academic Press, 3, 143-160.

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, Inc.

COOK, R.D., and WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMING, W.E., and STEPHAN, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 45-49.

DEVILLE, J., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J., SÄRNDAL, C.-E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DORFMAN, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.

DORFMAN, A.H., and HALL, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1475.

DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.

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

ELTINGE, J.L., and YANSANEH, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U. S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.

ESTEVAO, V., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

FOLSOM, R.E., and WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 428-433.

FOLSOM, R.E., and SINGH, A.C. (2000). The generalized exponential model for a unified approach to sampling weight calibration for outlier weight treatment, nonresponse adjustment and post-stratification. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 598-603.

FRANKEL, M.R. (1971). Inference from survey samples: An empirical investigation. Institute for Social Research, University of Michigan, Ann Arbor.

FULLER, W.A. (1966). Estimation employing post strata. *Journal of the America Statistical Association*, 61, 1172-1183.

FULLER, W.A. (1973). Regression for sample surveys. Paper presented at meeting of International Statistical Institute. August, 1973, Vienna, Austria.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā Series C*, 37, 117-132.

FULLER, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.

FULLER, W.A., and AN, A.B. (1998). Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.

FULLER, W.A., and BATTESE, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

FULLER, W.A., and ISAKI, C.T. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling*, (D. Krewski, J. N. K. Rao and R. Platek, Eds.), New York: Academic Press, 199-226.

FULLER, W.A., LOUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey, *Survey Methodology*, 20 75-85.

FULLER, W.A., and RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology*, 27, 45-52.

GAMBINO, J., KENNEDY, B. and SINGH, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.

GEROW, K., and MCCULLOCH, C.E. (2000). Simultaneously model unbiased, design-unbiased estimation. *Biometrics* 56, 873-878.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369-375.

GRAYBILL, F.A. (1976). *Theory and application of the linear model*. Wadsworth, Belmont, CA.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pestovani Matematiky*, 84, 387-423.

HANURAV, T.V. (1966). Some aspects of unified sampling theory. *Sankhyā, Series A*, 28, 175-204.

HENDERSON, C.R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, 141-163. National Academy Sciences, National Research Council Publication 982, Washington, DC.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theroem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Ph.D. thesis, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1978). *Super Carp*, (sixth edition, 1980) Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., SÄRNDAL, C.-E. and BINDER, D. A. (1995). Weighting and estimation in business surveys. *Business Survey Methods*, (Eds. Cox, Binder, Chinnappa, Colledge and Kott) New York: John Wiley & Sons, Inc., 477-502.

HOLT, D., and SMITH, T.M. F. (1979). Post Stratification. *Journal of the Royal Statistical Society, Serie. A*, 142, 33-46.

HORN, S.D., HORN, R.A. and DUNCAN, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the social statistics section*, American Statistical Association, 300-305.

HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

ISAKI, C.T. (1970). Survey designs utilizing prior information. Unpublished Ph.D. thesis. Iowa State University.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

ISAKI, C.T., TSAY, J.H. and FULLER, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*, 26, 31-42.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agriculture Experiment Station Research Bulletin*. 304

KALTON, G., and MALIGALIG, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference*, U. S. Bureau of the Census, 409-428.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

KONIJN, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association*, 57, 590-606.

KOTT, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

KUO, L. (1988). Classical and prediction approaches to estimating distribution function from survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 280-285.

LAZZERONI, L.C., and LITTLE, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

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

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

MONTANARI, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

MONTANARI, G.E. (1999). A study on the conditional properties of finite population mean estimators. *Metron*, 57, 21-35.

MUKHOPADHYAY, P. (1993). Estimation of a finite population total under regression models: A review. *Sankhyā*, 55, 141-155.

NIEUWENBROEK, N., RENSSEN, R. and HOFMAN, L. (2000). Towards a generalized weighting system. In *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia.

PARK, M. (2002). Regression estimation of the mean in Survey Sampling. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

PFEFFERMANN, D. (1984). Note on large sample properties of balanced samples. *Journal of the Royal Statistical Society, Series B*, 46, 38-41.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

RAO, J.N.K. (2002). *Small Area Estimation Theory and Methods*, New York: John Wiley & Sons, Inc.

RAO, J.N.K, HARTLEY, H.O. and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.

RAO, J.N.K., and SINGH, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.

ROBINSON, G.K. (1991). The BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-32.

ROBINSON, P.M., and SÄRNDAL, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā, Series B*, 45, 240-248.

ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika*, 70, 41-55.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Association*, 71, 657-664.

ROYALL, R.M. (1986). The prediction approach to robust variance estimation in two stage cluster sampling. *Journal of the American Statistical Association*, 81, 119-123.

ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). The finite population linear regression estimator and estimators of its variance, an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

SÄRNDAL, C.-E. (1980). On π-weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistics Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÄRNDAL, C.-E., and WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*. 11, 146-156.

SCOTT, A., and SMITH, T.M.F. (1974). Linear superpopulation models in survey and sampling. *Sankhyā, C*, 36, 143-146.

SCOTT, A., and WU, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.

SILVA, P.L.D.N., and SKINNER, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.

SINGH, A.C., and FOLSOM, R.E. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 610-615.

SINGH, A.C., KENNEDY, B. and WU, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating design. *Survey Methodology*, 27, 33-44.

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

TALLIS, G.M. (1978). Note on robust estimation infinite populations. *Sankhyā C*, 40, 136-138.

TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.

THÉBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

THÉBERGE, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26, 99-107.

TILLE, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.

TREMBLAY, V. (1986). Practical Criteria for Definition of Weighting Classes. *Survey Methodology*, 12, 85-97.

WATSON, D. J. (1937). The estimation of leaf area in field crops. *Journal of Agricultural Science*, 27, 474-483.

WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

WU, C., and SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

YATES, F. (1949). *Sampling Methods for Census and Surveys*. London: Griffin.

YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

ZYSKIND, G. (1976). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38, 1092-1109.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 25-29
Statistics Canada

Leslie Kish's Impact on Survey Statistics

GRAHAM KALTON[1]

ABSTRACT

Leslie Kish, one of the pioneers of survey sampling, died on October 7, 2000, at the age of 90. This paper reviews his impact on survey statistics, mainly in terms of his research but also in terms of his promotion of sound probability sampling methods around the world. Kish's research was broad-ranging, covering sampling methods, variance estimation and design effects, nonsampling errors, small area estimation, survey designs across time and space, and observational studies. He promoted probability sampling designs through consultancies in many countries, his writings, and in particular through the highly effective intensive summer Sampling Program for Foreign Statisticians that he established at the Survey Research Center of the University of Michigan.

KEY WORDS: Sample design; Variance estimation; Nonsampling errors; Rolling samples.

1. INTRODUCTION

Leslie Kish, one of the pioneers of survey sampling, died on October 7, 2000, at the age of 90. During his long and productive career, he had a major impact on the field, achieved both through his impressive research contributions and through his extremely successful promotion of the use of scientific probability sampling methods throughout the world, and especially in developing countries. His wide-ranging research always focused on issues of practical importance, and his innovations facilitated the use of effective probability sampling in diverse areas. He promoted the practice of probability sampling through his expository writings (particularly for sociologists and demographers), through his numerous consultancies and advisory services, and through his training of survey statisticians, particularly those from developing countries.

This paper reviews Kish's impact on survey statistics, primarily with respect to his contributions to the advancement of survey sampling and survey research more generally. It is useful to start with a brief account of his career in order to place these contributions in a temporal context. The interview of Kish in 1994 by Frankel and King (1996) is recommended for those interested in more details of Kish's fascinating life. Some of the material in this paper is drawn from that interview.

Kish was born in 1910 in Poprad, which was then part of the Austro-Hungarian Empire and is now in Slovakia. In 1926, he emigrated to the United States with his family. When his father died the following year, he became a laboratory assistant at the Rockefeller Institute for Medical Research, while attending Bay Ridge Evening High School. He graduated from high school in 1930 and enrolled in the College of the City of New York night school, while continuing to work for 54 hours a week at the Rockefeller Institute. His interest in statistics arose out of his work at the Institute, and he studied on his own books by Fisher, Yule, Wallace and Snedecor, Tippett, Pearl, and others. In 1937, he interrupted his education to join the International Brigade to fight for the Loyalist cause in the Spanish Civil War. He returned to the United States in 1939 and earned a B.S. in Mathematics, cum laude, in that year. He was then hired by the U.S. Census Bureau as a Section Head, and subsequently moved to be a Statistician at the United States Department of Agriculture (USDA) Division of Program Surveys. In 1942, he left the Division of Program Surveys for war service, returning there in 1945 after the war. In 1947, he moved with a group of USDA colleagues headed by Rensis Likert to set up the Survey Research Center at the University of Michigan. He remained at the Survey Research Center until his retirement in 1981, when he became a Professor Emeritus. He remained fully active professionally until his death in 2000.

2. RESEARCH

At the start of Kish's career, survey sampling was in its infancy. Much survey research was based on nonprobability samples. Methods for probability sampling were under development and many problems remained to be resolved. While at the USDA, Kish identified three important problems that he pursued at the Survey Research Center (SRC) in developing sampling methods there.

One of these problems was how to have an interviewer randomly select an individual within a sampled household. At the time, probability sampling methods for sampling households had been developed and were being applied in the Current Population Survey, but the CPS collected data on all members of sampled households, so that no selection of persons within households was needed. Kish invented a method for objective respondent selection and wrote it up in a memorandum. He was urged by his colleague Angus Campbell to submit the work for publication, and it resulted

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

in the famous paper that was his first published research (Kish 1949). The widely used method is now known as the Kish selection table.

The second problem that Kish identified was counting nonresponse. He had to argue for counting and reporting nonresponse with probability samples against the concerns of colleagues who felt that to do so would put the SRC at a competitive disadvantage, particularly with organizations using nonprobability methods. He won his case and SRC adopted his approach, which is now fully accepted as standard good practice.

The third problem was that of deep stratification. Standard stratification assumes independence of selections between strata, with the maximum number of strata possible being the number of selections. Particularly when the number of selections is small, as is often the case with the primary sampling units (PSUs) in a multistage design, it can be desirable to obtain greater balance in the sample than standard stratification permits. With Roe Goodman, Kish developed the technique of controlled selection that provides that greater balance by dropping the requirement of independence of selections between strata, while still retaining probability sampling (Goodman and Kish 1950). Kish, who was always concerned to coin good names, preferred to call the technique 'multiple stratification', and he uses that term in his sampling text (Kish 1965a).

Kish's subsequent research in survey statistics was wide-ranging, covering many aspects of survey sampling, nonsampling errors, small area estimation, survey designs across time and space, and observational studies. His many contributions have had a major impact on the development of the practice of survey sampling and of survey research more generally. The following paragraphs outline some of his contributions organized by topic.

Variance estimation. Before the 1970s, the analysis of survey data was severely limited by the analytic tools available, then mostly punch card equipment, such as counter-sorters and tabulators, and hand calculators. Thus, for example, weights – and particularly non-integer weights – were difficult to handle. For this reason Kish examined the use of uniform weights with the Kish selection table, even though unbiased estimation calls for weights proportional to the number of eligible household members.

As a result of the computational difficulties, prior to the 1970s sampling errors were rarely computed in a manner that reflected the complex sample designs typically employed in survey research. A widespread practice was to compute variances as if a simple random sample (SRS) had been drawn. Kish sought to promote the use of appropriate variance estimation methods by social researchers, which he did by illustrating the sizable underestimation that often arises when SRS formulas are applied to clustered samples (Kish 1957). Initially he developed and applied simple computational procedures, emphasizing the simplicity that can be obtained with a paired selection design in which two PSUs are sampled in each stratum (Kish and Hess 1959a; Kish 1968). He coined the term "design effect" for the ratio of the variance of a survey estimate for a given design to the variance of the same estimate obtained from a simple random sample of the same size. He made much use of this concept in his famous *Survey Sampling* book (Kish 1965a), which provides an encyclopedic treatment of practical survey sampling and is still widely read as a Wiley classic. He retained his interest in design effects throughout his career as an important tool in the design and analysis of survey samples (see, for example, Kish 1982, 1995a; Kish, Frankel, Verma and Kaciroti 1995; Kish, Groves and Krotki 1976). An important term in the design effect for a clustered sample is the intra-class correlation, which is featured in Kish's Ph.D. dissertation (Kish 1952) and in a number of his other papers (*e.g.*, Kish 1954, 1961a).

With the development of computers, Kish was quick to see their importance for variance estimation, and with SRC colleagues he developed an early *Sampling Error Program Package* (Kish, Frankel and Van Eck 1972). With his doctoral student Martin Frankel, he also extended the range of statistics for which sampling errors from complex sample designs could be computed (Kish and Frankel 1970, 1974). This highly influential research developed, applied, and evaluated balanced repeated replication (BRR) and jack-knife repeated replication (JRR) methods of variance estimation. It also provided a definition of the population parameters estimated by analytical survey statistics in the finite population context.

Multipurpose surveys. The survey sampling literature deals mostly with an efficient sample design for estimating a single population parameter. Kish recognized the limitation of this approach since virtually all surveys are multipurpose in nature. He wrote several important papers dealing with multipurpose surveys, producing effective compromise designs that provide estimates not only for the population as a whole but also for various domains (Kish 1961b, 1969, 1976; Anderson, Kish and Cornell 1976; Kish and Anderson 1978; Kish 1980; Kish 1988). In recent years, he extended his interests to multipopulation surveys (*e.g.*, Kish 1999, 2002).

Small area estimation. In considering the production of estimates for domains, Kish (1980, 1987a, 1987b) classified domains into major, minor, and mini domains and rare items. Estimates for major domains can be produced from a survey using standard sample-based estimators, particularly if the sample is designed to give sufficient domain sample sizes for this purpose. The sample sizes of most surveys preclude the production of estimates of adequate precision for minor or mini domains that comprise less than, say, one-tenth of the population. Yet, as Kish recognized early on, the demand for up-to-date estimates for small domains, particularly small geographical areas, would expand. This recognition led to his research in two related areas.

When a survey's sample size is too small to produce small area sample-based estimates of adequate precision,

reliance may be placed on statistical models to produce indirect estimates. Much research on small area estimation techniques using this model-dependent approach has been conducted in recent years. In the 1970's, Kish contributed to the development of the field through his direction of three doctoral dissertations at the University of Michigan (Ericksen 1973; Kalsbeek 1973; Purcell and Kish 1979, 1980).

Direct, or sample-based, estimates for small domains are sometimes possible. One obvious source of estimates for domains of any size is a population census, and indeed censuses are a major source of small domain estimates. However, data from a decennial census become out-of-date as the decade progresses. To address this problem, Kish proposed replacing the census by a rotating or rolling sample so that, by spreading the data collection over time, more up-to-date estimates can be produced. He first proposed such a procedure in 1979 (Kish 1979a,b), and wrote many papers on this topic after that (Kish 1981, 1983, 1986, 1990, 1997, 1998, 2002; Kish and Verma 1986), including the issue of how to cumulate sample data over time (Kish 1999). In another paper in this volume, Charles Alexander (2002) provides a detailed review of Kish's work on this topic and its influence on the large-scale continuous survey, the American Community Survey, that the U.S. Census Bureau plans to introduce to replace the long form in the 2010 Census.

Special sample design problems. During the course of his work, Kish encountered a number of specialized sampling problems that often occur and he offered some efficient solutions. The areas to which he contributed include the following:

- *Sampling rare and elusive populations.* One of the most challenging design tasks faced by sampling statisticians is constructing an efficient sample design for a rare or elusive population (such as persons with a rare illness or the homeless). Kish (1965b, 1991) provides insightful reviews of methods for tackling this type of problem.
- *Maximizing overlap.* When a population is sampled repeatedly over time, the issue arises of how to control the sample overlap between one round and the next. A particular example occurs when a master sample of PSUs is used and needs to be updated when new census data become available. Frequently it is desirable to maximize the overlap in the sample of PSUs, while updating measures of size and changing the stratification to reflect current data. Kish and Scott (1971) provide a relatively simple and effective method of satisfying these requirements.
- *Sampling organizations of unequal size.* Some surveys are designed to produce estimates for units at different levels, for instance, for hospitals and for patients. When hospitals vary considerably in their numbers of patients, a design conflict arises between the production of efficient hospital- and patient-level estimates. Kish (1965c) examines this problem and clarifies the issues involved.

Nonsampling errors. Kish clearly recognized the harmful effects that nonsampling errors can have on the quality of survey estimates. Early in his career he collaborated with Jack Lansing to investigate the response errors in respondents' reports of the values of their homes by comparing these reports with estimates made by professional appraisers (Kish and Lansing 1954). In his studies of interviewer variance, he took advantage of the theory on cluster sampling, measuring interviewer variance with the intra-class correlation coefficient, and determining the optimum number of interviews per interviewer based on a simple cluster sample cost model (Kish 1962). With Irene Hess, he conducted a study of noncoverage in area samples of dwelling units. The study was stimulated by a 10 percent noncoverage rate in SRC surveys at that time, and led to improvements that reduced this rate to about 3 percent (Kish and Hess 1958). Also with Irene Hess, he introduced an imaginative replacement procedure for noncontacts in one survey by substituting noncontacts from a previous, similar, survey (Kish and Hess 1959b). For stochastic imputation schemes, Kish was an early proponent of replicating the imputations to reduce imputation variance, in what he termed a repeated replication imputation procedure (RRIP) and what is now known as fractional imputation (Kalton and Kish 1984).

Observational studies. Early in his career, Kish (1959) wrote a widely cited paper on the design of studies to investigate causal relationships, particularly nonrandomized studies. In his writing about this topic he made use of his survey sampling expertise as, for instance, in the relationship between stratification and matching (Anderson, Kish and Cornell 1980). His work developed into his book *Statistical Design for Research* (Kish 1987a) in which he compared surveys, experiments, and observational studies for investigating causal effects in terms of the three R's: realism, randomization and representativeness (see also Kish 1975). He also made clear the importance of assessing both bias and variance in assessing the ability of different study designs to measure causal effects, rather than concentrating on bias as had been common in the literature on this topic.

3. OTHER CONTRIBUTIONS

Kish's seminal and wide-ranging contributions to the methodology of survey statistics are of great importance. Yet of possibly even greater importance are his contributions to the promotion of the use of sound probability sampling methods around the world.

Kish's writings, of course, contributed to the current widespread use of probability sampling methods by emphasizing good practical methods. His three books *Survey Sampling* (Kish 1965a), *Statistical Design for Research* (Kish 1987a), and *Sampling Methods for Agricultural Surveys* (Kish 1989) are all extremely valuable in this respect, as are his expository writings for social scientists.

Kish had a long-standing dedication to assisting developing and transition countries, and that can be seen in many of his activities. He was a sampling consultant to the World Fertility Survey from 1973 to 1983 and he consulted in many countries, he ran a training program for foreign statisticians, and he wrote specifically for statisticians in developing countries. *Sampling Methods for Agricultural Surveys* was, for instance, written for the FAO, particularly for use in developing countries. He contributed a *Questions/Answers* column for the *Survey Statistician*, the newsletter of the International Association of Survey Statisticians, from 1978 to 1994. In that column he provided sound advice on many practical sampling problems that frequently arise but that are not well addressed in the literature. The column was considered so useful that the IASS published the full set of questions and answers in a special volume (Kish 1995b).

Kish was rightly particularly proud of the intensive two-month summer Sampling Program for Foreign Statisticians that he established at the Survey Research Center in 1961. The SPFS has now trained more than 500 survey statisticians from 105 countries. It is significant that Kish chose "Developing samplers for developing countries" as the topic for his 1994 Morris Hansen Memorial Lecture (Kish 1996). To help maintain this important program, the Leslie Kish International Fellows Fund was established at the University of Michigan at a celebration of Kish's 90th birthday. Of all his accomplishments, the SPFS was the one that gave him greatest pleasure.

4. CONCLUDING REMARKS

Leslie Kish is a giant in the field of survey sampling. His contributions were enormous and recognized by many honors. These honors included, among others, President of the International Association of Survey Statisticians in 1983-85, President of the American Statistical Association in 1978 (see Kish 1978, for his Presidential address on "Chance, Statistics and Statisticians"), Honorary Fellow of the International Statistical Institute, Honorary Fellow of the Royal Statistical Society, Honorary Member of the Hungarian Academy of Sciences, Fellow of the American Association for the Advancement of Science, Fellow of the American Academy of Arts and Sciences, recipient of the American Statistical Association's Samuel L. Wilks Award in 1997, recipient of the Mindel Shep Award from the Population Association of America in 1998, recipient of the Methodology Award from the American Sociological Association in 1989, and honorary degrees from the University of Bologna, the Athens University of Economics and Business, and the Eotvos Lorand University in Budapest.

Yet Kish remained down-to-earth, approachable by all. He had a great enthusiasm for many subjects including sport, art, literature, politics, philosophy, and science. He was always concerned with improving the conditions of the world's population. He was particularly interested in young people and one of his favorite sayings was "Keep young by being curious, and have young friends". Undoubtedly his endearing personality played an important part in his great success in promoting sound sampling methods around the world. Ivan Fellegi's excellent obituary in *Survey Methodology* was aptly titled "Leslie Kish – A Life of Giving" (Fellegi 2000). Kish gave so much personally to so many people and so much professionally to the development of survey statistics.

REFERENCES

ALEXANDER, C. H. (2002). Still rolling: Leslie Kish's "rolling samples" and the American Community Survey. *Survey Methodology*, 28, 35-41.

ANDERSON, D.W., KISH, L. and CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

ANDERSON, D.W., KISH, L. and CORNELL, R.G. (1980). On stratification, grouping, and matching. *Scandinavian Journal of Statistics*, 7, 61-66.

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

FELLEGI, I.P. (2000). Leslie Kish – A life of giving. *Survey Methodology*, 26, 119-120.

FRANKEL, M., and KING, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 65-87.

GOODMAN, R., and KISH, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

KALSBEEK, W.D. (1973). A Method for Obtaining Local Postcensal Estimates for Several Types of Variables. Ph. D. Thesis, University of Michigan.

KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, 13(16), 1919-1939.

KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

KISH, L. (1952). On the Differentiation of Ecological Units. Ph.D. Thesis, University of Michigan.

KISH, L. (1954). Differentiation in metropolitan areas. *American Sociological Review*, 19, 388-398.

KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 1954-1965.

KISH, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24, 328-338.

KISH, L. (1961a). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*, 4, 201-209.

KISH, L. (1961b). Efficient allocation of a multi-purpose sample. *Econometrica*, 29, 363-385.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

KISH, L. (1965a). *Survey Sampling*. New York: John Wiley & Sons, Inc.

KISH, L. (1965b). Selection techniques for rare traits. *Genetics, Epidemiology, and Chronic Diseases,* Public Health Service Publication, No. 1173.

KISH, L. (1965c). Sampling organizations and groups of unequal sizes. *American Sociological Review*, 20, 564-572.

KISH, L. (1968). Standard errors for indexes from complex samples. *Journal of the American Statistical Association*, 63, 512-529.

KISH, L. (1969). Design and estimation for subclasses, comparisons, and analytical statistics. *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc.

KISH, L. (1975). Representation, randomization and control. *Quantitative Sociology*, (Eds. H.M.Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon and V. Capecchi). New York: Academic Press.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, A*, 139, 80-95.

KISH, L. (1978). Chance, statistics, and statisticians. *Journal of the American Statistical Association*, 73, 1-6.

KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.

KISH, L. (1979b). Rotating samples instead of censuses. *Asian and Pacific Census Forum* (East-West Center, Honolulu), 6, 1-13.

KISH, L. (1980). Design and estimation for domains. *The Statistician*, 29, 209-222.

KISH, L. (1981). Using Cumulated Rolling Samples.Washington: Library of Congress.

KISH, L. (1982). Design effects. *Encyclopedia of Statistics*, New York: John Wiley & Sons, Inc.

KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright). New York: Academic Press.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.

KISH, L. (1987a). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.

KISH, L. (1987b). Discussion. *Small Area Statistics*, (Ed. R. Platek) New York: John Wiley & Sons, Inc.

KISH, L. (1988). Multipurpose sample design. *Survey Methodology*, 14, 19-32.

KISH, L. (1989). *Sampling Methods for Agricultural Surveys*. Rome: FAO.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-71 and 93-94.

KISH, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339-347.

KISH, L. (1995a). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L. (1995b). *Questions/Answers from the Survey Statistician 1978-1994*. Libourne: International Association of Survey Statisticians.

KISH, L. (1996). Developing samplers for developing countries. *International Statistical Review*, 64, 143-162.

KISH, L. (1997). Periodic and rolling samples and censuses. *Statistics and Public Policy*, (Ed. B.D. Spencer). New York: Oxford University Press.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KISH, L. (1999). Combining/cumulating population surveys. *Survey Methodology*, 25, 129-138.

KISH, L. (2002). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, 102, 109-118.

KISH, L., and ANDERSON, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 24-34.

KISH, L., and FRANKEL, M. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, B*, 36, 1-37.

KISH, L., FRANKEL, M.R. and VAN ECK, M. (1972). *SEPP: Sampling Error Programs Package*. Ann Arbor: Institute for Social Research.

KISH, L., FRANKEL, M.R., VERMA, V. and KACIROTI, N. (1995). Design effects for correlated $(P_i - P_j)$. *Survey Methodology*, 21, 117-124.

KISH, L., GROVES, R.M. and KROTKI, K. (1976). Sampling Errors for Fertility Surveys. Occasional Paper No. 17, World Fertility Survey.

KISH, L., and HESS, I. (1958). On noncoverage of sample dwellings. *Journal of the American Statistical Association*, 53, 509-524.

KISH, L., and HESS, I. (1959a). On variances of ratios and their differences in multi-stage samples. *Journal of the American Statistical Association*, 54, 416-446.

KISH, L., and HESS, I. (1959b). A replacement procedure for reducing the bias of nonresponse. *The American Statistician*, 13, 17-19.

KISH, L., and LANSING, J.B. (1954). Response error in estimating the value of homes. *Journal of the American Statistical Association*, 49, 520-538.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

KISH, L., and VERMA, V. (1986). Complete censuses and samples. *Journal of Official Statistics*, 2, 381-96.

PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (small domains). *International Statistical Review*, 48, 3-18.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 31-34
Statistics Canada

New Paradigms (Models) for Probability Sampling

LESLIE KISH[1]

1. STATISTICS AS A NEW PARADIGM

In several sections I discuss new concepts in diverse aspects of sampling, but I feel uncertain whether to call them new paradigms or new models or just new methods. Because of my uncertainty and lack of self-confidence, I ask the readers to choose that term with which they are most comfortable. I prefer to remove the choice of that term from becoming an obstacle to our mutual understanding.

Sampling is a branch of and a tool for statistics, and the field of statistics was founded as a new paradigm in 1810 by Quetelet (Porter 1987; Stigler 1986). This was later than the arrival of some sciences: of astronomy, of chemistry, of physics. "At the end of the seventeenth century the philosophical studies of cause and chance...began to move close together... During the eighteenth and nineteenth centuries the realization grew continually stronger that aggregates of events may obey laws even when individuals do not." (Kendall 1968). The predictable, meaningful, and useful regularities in the behavior of population aggregates of unpredictable individuals were named "statistics" and were a great discovery.

Thus Quetelet and others computed national (and other) birth rates, death rates, suicide rates, homicide rates, insurance rates, *etc.* from individual events that are unpredictable. These statistics are basic to fields like demography and sociology. Furthermore, the ideas of statistics were taken later during the nineteenth century also into biology by Frances Galton and Karl Pearson, and into physics by Maxwell, and were developed greatly both in theory and applications.

Statistics and statisticians deal with the effects of chance events on empirical data. The mathematics of chance had been developed centuries earlier for gambling games and for errors of observation in astronomy. Also data have been compiled for commerce, banking, and government. But combining chance with real data needs a new theoretical view, a new paradigm. Thus statistical science and its various branches arrived late in history and in academia, and they are products of the maturity of human development (Kish 1985).

The populations of random individuals comprise the most basic concept of statistics. It provides the foundation for distribution theories, inferences, sampling theory, experimental design, *etc.* And the statistics paradigm differs fundamentally from the deterministic outlook of cause and effect, and of precise relations in the other sciences and mathematics.

2. THE PARADIGM OF SAMPLING

The Representative Method is the title of an important monograph, almost a century after the birth of statistics and over a century ago now, which is generally accepted as the birth of modern sampling (Kiaer 1895). That term has been used in several landmark papers since then (Jensen 1926; Neyman 1934; Kruskal and Mosteller 1979a, 1979b, 1979c, 1980). The last authors agree that the term "representative" has been used for so many specific methods and with so many meanings that it does not denote any single method. However, as Kiaer used it, and as it is still used generally, it refers to the aims of selecting a sample to represent a population specified in space, in time, and by other definitions, in order to make statistical inferences from the sample to that specified population. Thus a national representative sample demands careful operations for selecting the sample from all elements of the national population, not only from some arbitrary domain such as a "typical" city or province, or from some subset, either defined or undefined.

The scientifically accepted method for survey sampling is probability sampling, which assures known positive probabilities of selection for every element in the frame population. The frame provides the equivalent of listings of sampling units for each stage of selection. The sampling frame for the entire population is needed for mechanical operations of random selection. This is the basis for statistical inferences from the sample statistics to the corresponding population statistics (parameters) (Hansen, Hurwitz and Madow 1953a, 1953b). This insistence on inferences based on selections from frame populations is a different paradigm from the unspecified or model based approaches of most statistical analyses.

It took a half century from Kiaer's paper to the wide acceptance of survey sampling. In addition to neglect and passive resistance, there was a great deal of active opposition by national statistical offices which distrusted sampling methods to replace the complete counts of censuses. Some even preferred the "monograph method," which offered complete counts of a "typical" or "representative" province or district instead of randomly selected national sample (O'Muircheartaigh and Wong 1981). In addition to political opposition, there were also many opponents among academic disciplines, and among academic statisticians. The tide in favor of probability sampling turned with the report of the UN Statistical Commission led by Mahanalobis and Yates (United Nations

[1] Printing of this paper has been kindly authorized by Rhea Kish, 1050 Wall St. #9A, Ann Arbor, MI 48105, e-mail: rheakk@umich.edu.

Statistical Office 1950). Five influential textbooks between 1949 and 1954 started a flood of articles with both theory and wide applications.

The strength, the breadth, and the duration of resistance to the concepts and use of probability sampling of frame populations implies that this was a new paradigm that needed a new outlook both by the public and the professionals.

3. COMPLEX POPULATIONS

The need for strict probability selection from a population frame for inferences from the sample to a *finite* population is but one distinction of survey sampling. But even more important and difficult problems are caused by the complex distributions of the elements in all the populations. These complexities present a great contrast with the simple model of independence that is assumed, explicitly or implicitly, by almost all statistical theory, all mathematical statistics.

The assumption of independent or uncorrelated observations of variables or elements underlies mathematical statistics and distribution theory. We need not distinguish here between independently and identically distributed (IID) random variables and "exchangeability," and "superpopulations." The simplicity underlying each of those models is necessary for the complexities of the mathematical developments.

Simple models are needed and used for early stages and introductions in all the sciences: for example, perfect circular paths for the planets or $d = gt^2/2$ for freely dropping objects in frictionless situations. But those models fail to meet the complexities of the actual physical world. Similarly, independence of elements does not exist in any *population* whether human, animal, plant, physical, chemical, biological. The simple independent models may serve well enough for small *samples*; and the Poisson distribution of deaths by horsekicks in the Prussian Army in 43 years has often served as an example (precious because rare) (Fisher 1926).

There have also been attempts to construct theoretical populations of IID elements; perhaps the most famous was the classic "collective" of Von Mises (1931); but they do not correspond to actual populations. However, with great effort tables of random numbers have been constructed that have passed all tests. These have been widely used in modern designs of experiments and sample surveys. *Replication* and *randomization* are two of the most basic concepts of modern statistics following the concept of populations.

The simple concept of a population of independent elements does not describe adequately the complex distributions (in space, in time, in classes) of elements. Clustering and stratification are common names for ubiquitous complexities. Furthermore, it appears impossible to form models that would better describe actual populations. The distributions are much too complex and they are also different for every survey variable. These complexities and differences have been investigated and presented now in thousands of computations of "design effects."

Survey sampling needed a new paradigm to deal with the complexities of all kinds of populations for many survey variables and a growing list of survey statistics. This took the form of robust designs of selections and variance formulas that could use a multitude of sample designs, and gave rise to the new discipline of survey sampling. The computation of "design effects" demonstrated the existence, the magnitude, and the variability of effects due to the complexities of distributions not only for means but also for multivariate relations, such as regression coefficients. The long period of disagreements between survey samplers and econometricians testifies to the need for a new paradigm.

4. COMBINING POPULATION SAMPLES

Samples of national populations always represent subpopulations (domains) which differ in their survey characteristics; sometimes they differ slightly, but at other times greatly. These subclasses can be distinguished in the sample with more or less effort. First, samples of provinces are easily separated when their selections are made separately. Second, subclasses by age, sex, occupation, and education can also be distinguished, and sometimes used for poststratified estimates. Third, however, are those subclasses by social, psychological, and attitudinal characteristics, which may be difficult to distinguish; yet they may be most related to the survey variables. Thus, we recognize that national samples are not simple aggregations of individuals from an IID population, but combinations of subclasses from subpopulations with diverse characteristics. The composition of national populations from diverse domains deserves attention, and it also serves as an example for the two types of combinations that follow. Furthermore, these remarks are pertinent to combinations not only of national samples but also of cities, institutions, establishments, *etc.*

In recent years two kinds of sample designs have emerged that demand efforts beyond those of simple national samples: a) periodic samples and b) multipopulation designs. Each of these has emerged only recently, because they had to await the emergence of three kinds of resources: 1. effective demand supported by financial and political resources; 2. adequate institutional technical resources in national statistical offices; 3. new methods. In both types of designs we should distinguish the needs of the *survey methods* (definitions, variables, measurements), which must be harmonized, standardized, from *sample designs*, which can be designed freely to fit national (even provincial) situations, provided they are probability designs

(Kish 1994). Both types have been designed first and chiefly for comparisons: periodic comparisons and multinational comparisons, respectively. But new uses have also emerged: "rolling samples" and multinational cumulations, respectively. Each type of cumulation has encountered considerable opposition, and needs a new outlook, a new paradigm.

"Rolling samples" have been used a few times for local situations (Mooney 1956; Kish, Lovejoy and Rackow 1961). Then they have been proposed several times for national annual samples and as a possible replacement for decennial censuses (Kish 1981, 1990). They are now being introduced for national sample censuses first and foremost by the US Census Bureau (Alexander 1999; Kish 1990). Recommending this new method, I have usually experienced opposition to the concept of averaging periodic samples: "How can you average samples when these vary between periods?" In my contrary view, the greater the variability the less you should rely on a single period, whether the variation is monotonic, or cyclical, or haphazard. Hence I note two contrasting outlooks, or paradigms. Quite often, the opposition disappears after two days of discussion and cogitation.

> "For example, annual income is a readily accepted aggregation, and not only for steady incomes but also for occupations with high variations (seasonal or irregular). Averaging weekly samples for annual statistics will prove more easily acceptable than decennial averaging. Nevertheless, many investors in mutual stock funds prefer to rely more on their ten-year or five-year average earnings (despite their obsolescence) than on their up-to-date prior year's earnings (with their risky "random" variations). Most people planning a picnic would also prefer a 50 year average "normal" temperature to last year's exact temperature. There are many similar examples of sophisticated averaging over long periods by the "naïve" public. That public, and policy makers, would also learn fast about rolling samples, given a chance."
>
> (Kish 1998)

Like rolling samples, combining multipopulation samples also encountered opposition: national boundaries denote different historical stages of development, different laws, languages, cultures, customs, religions, behaviors. How then can you combine them? However, we often find uses and meanings for continental averages; such as European birth and death rates, or South American, or sub-Saharan, or West African rates. Sometimes even world birth, death, and growth rates. Because they have not been discussed, they all usually combined very poorly. But with more adequate theory, they can be combined better (Kish 1999). But first the need must be recognized with a new paradigm for multinational combinations, followed by developing new and more appropriate methods.

5. EXPECTATION SAMPLING

Probability sampling assures for each element in the population ($i = 1, 2, ..., N$) a known positive probability ($P_i > 0$) of selection. The assurance requires some mechanical procedure of chance selection, rather than only assumptions, beliefs, or models about probability distributions. The randomizing procedure requires a practical physical operation that is closely (or exactly) congruent with the probability model (Kish 1965). Something like this statement appears in most textbooks on survey sampling, and I still believe it all. However, there are two questionable and bothersome objections to this definition and its requirements.

The more important of the two objections concerns the frequent practical situations when we face a choice between probability sampling and expectation sampling. These occur often when the easy, practical selection rate for listing units of $1/F$ yields not only the unique probability $1/F$ for elements, but also some with variable k_i/F for the ith element ($i = 1, 2, ..., N$) and with $k_i > 0$. Examples of $k_i > 1$, usually a small integer, occur with duplicate or replicate lists, dual or multiple frames of selection, second homes for households, mobile populations and nomads, farm operators with multiple lots. Examples of $k_i < 1$ are selecting a single adult from households, selecting single dwellings from buildings. In these examples often the k_i can be easily ascertained, and it is cheaper, more convenient and economical to use weighting than attempting to obtain $1/F$ for all the elements. These problems are described in books and articles.

In most cases, we find it more convenient and less expensive to accept the variable probabilities and to counter them with weighting the expected values $1/k_i$ or k_i than to operate another stage of selection. Thus, to paraphrase probability sampling: *expectation sampling* assures for each element in the population ($i = 1, 2, ..., N$) a known positive expected number of selections ($k_i/F > 0$). These procedures are used in practice for descriptive (first order) statistics where the k_i or $1/k_i$ are neither large nor frequent. The treatments for inferential – second order or higher – statistics are more difficult and diverse, and are treated separately in the literature. Note that probability sampling is the special (and often desired) situation when all k_i are 1.

The other objection to the term probability sampling is more theoretical and philosophical and concerns the word "known" in its definition. That word seems to imply belief. Authors from classics like John Venn and M.G. Kendall to modern Bayesians like Dennis Lindley – and beyond at both ends – have clearly assigned "probability" to states of belief and "chance" to frequencies generated by objective phenomena and mechanical operations. Thus, our insistence

on operations, like random number generators, should imply the term "chance sampling." However, I have not observed its use and it also could lead to a philosophical problem: the proper use of good tables of random numbers implies beliefs in their "known" probabilities. I have spent only a modest amount of time on these problems and agreeable discussions with only a few colleagues, who did agree. I would be grateful for further discussions, suggestions and corrections.

6. SOME RELATED TOPICS

We called for recognition of new paradigms in six aspects of survey sampling, beginning with statistics itself. Finally, we note here the contrast of sampling to other related methods. Survey methods include the choice and definition of variables, methods of measurements or observations, control of quality (Kish 1994; Groves 1989).

Survey sampling has been viewed as a method that competes with censuses (annual or decennial), hence also with registers (Kish 1990). In some other context, survey sampling competes with or supplements experiments and controlled observations, and clinical trials. These contrasts also need broader comprehensive views (Kish 1987, section A.1). However, those discussions would take us well beyond our present limits.

REFERENCES

ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Bulletin of the International Statistical Institute*, Helsinki, 52nd session.

FISHER, R.A. (1926). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953a). *Sample Survey Methods and Theory, I – Methods and Applications*, New York: John Wiley & Sons, Inc.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953b). *Methods and Applications, II – Theory*. New York: John Wiley & Sons, Inc.

JENSEN, A. (1926). The representative method in practice, *Bulletin of the International Statistical Institute*, 22, pt. 1, 359-439.

KENDALL, M.G. (1968). Chance. *Dictionary of the History of Ideas*, (Ed. P.P.Wiener), New York: Chas Scribners.

KIAER, A.W. (1895). The Representative Method of Statistical Surveys, English translation, 1976, Oslo: Statistik Sentralbyro.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

KISH, L. (1981). *Using Cumulated Rolling Samples*. Washington DC: Library of Congress.

KISH, L. (1985). Chance, statistics, sampling. *Journal of Official Statistics*, 1, 35-47.

KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-79.

KISH, L. (1994). Multipopulation survey designs. *International Statistical Review*, 62, 167-186.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KISH, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25, 129-138.

KISH, L., LOVEJOY, W. and RACKOW, P. (1961). A multistage probability sample for continuous traffic surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 227-230.

KRUSKAL, W.H., and MOSTELLER, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review*, 47, 13-24.

KRUSKAL, W.H., and MOSTELLER, F. (1979b). Representative sampling, II: Non-scientific literature. *International Statistical Review*, 47, 111-127.

KRUSKAL, W.H., and MOSTELLER, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review*, 47, 245-265.

KRUSKAL, W.H., and MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics. *International Statistical Review*, 48, 169-195.

MOONEY, H.W. (1956). Methodology of Two California Health Surveys, US Public Health Monograph 70, Washington DC: US Government Printing Office.

NEYMAN, J. (1934). On the different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

O'MUIRCHEARTAIGH, C., and Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: a review. *Bulletin of International Statistical Institute, 43rd Session*, 1, 465-493.

PORTER, T.M. (1987). The Rise of Statistical Thinking: 1820-1900, Princeton, NJ: Princeton University Press.

STIGLER, S.M. (1986): History of Statistics, Cambridge: Harvard University Press.

UNITED NATIONS STATISTICAL OFFICE (1950). The Preparation of Sample Survey Reports, New York: UN Series C No 1; also Revision 2 in 1964.

VON MISES, R. (1939). *Probability, Statistics, and Truth*, London: Wm. Hodge and Co.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 35-41
Statistics Canada

Still Rolling: Leslie Kish's "Rolling Samples" and the American Community Survey

CHARLES H. ALEXANDER[1]

ABSTRACT

Leslie Kish long advocated a "rolling sample" design, with non-overlapping monthly panels which can be cumulated over different lengths of time for domains of different sizes. This enables a single survey to serve multiple purposes. The Census Bureau's new American Community Survey (ACS) uses such a rolling sample design, with annual averages to measure change at the state level, and three-year or five-year moving averages to describe progressively smaller domains. This paper traces Kish's influence on the development of the American Community Survey, and discusses some practical methodological issues that had to be addressed in implementing the design.

KEY WORDS: Rolling sample; Multi-year averages; Asymmetrical cumulations.

1. INTRODUCTION

A "rolling sample design", defined below, gives a single survey the flexibility to serve multiple purposes. The concept was developed by Leslie Kish in a series of papers (including Kish 1979a, 1979b, 1981, 1983, 1986, 1990, 1997, 1998 and Kish and Verma 1983, 1986) in which he elaborated the principles of cumulating information over space and time from a rolling sample. Kish advocated its use for a variety of purposes (Kish 1998), especially in developing countries (Kish 1979b), but also in the context of the U.S. census (Kish 1981). His personal use of rolling samples goes back at least to 1958, under the name "continuous sampling" (Kish, Lovejoy and Rackow 1961); a still earlier project (Mooney 1956) is cited in Kish (1998).

The American Community Survey (ACS), which is being developed as a replacement for the traditional "long form" survey conducted as part of the census, will use a form of the rolling sample design. This paper describes how the rolling sample concept is being implemented for the ACS, influenced by its specific objectives and operational considerations. The design decisions made for the ACS illustrate some issues that may arise for rolling samples in general. They also illustrate how Leslie Kish influenced survey development on multiple levels: philosophical, personal, and practical.

2. ROLLING SAMPLES

A "rolling sample" design jointly selects k non-overlapping probability samples (panels), each of which constitutes $1/F$ of the entire population. One panel is interviewed each time period until all the sample has been interviewed after k periods. Depending on the precision requirements, a single panel of $1/F$ may be sufficient to provide good estimates for the population as a whole, and possibly for some large domains. For smaller domains or for greater precision for large domains, cumulations of different numbers of consecutive panels can be used, up to k/F of the population. A rolling sample design with $k=F$ is called a "rolling census". For a monthly rolling sample, it is natural to have F be a multiple of twelve, and natural cumulations are quarterly, semi-annual, annual, and multiple years.

"Domains" include both geographic areas and demographic subgroups. Kish (1987, section 2.3) presents a framework for the tradeoff between geographic and demographic detail, for a given required level of precision. Even more central to the idea of rolling samples was the idea of "asymmetrical cumulation" of data, over different lengths of time for different sizes of domain (Kish 1990, 1998), which was later broadened into a view of the basic similarities of averaging over space and averaging over time (Kish 1998), as well as averaging over different demographic domains. The flexibility of the rolling sample design comes from the opportunities it provides to make different tradeoffs between spatial, temporal, and demographic detail.

Leslie Kish left his colleagues with a challenge to extend these ideas into a "theory of combining populations" (Kish 1999, 2001). He organized a contributed paper session on "combining surveys" at the 1999 meetings of the International Statistical Institute, explaining to the presenters that we were all working on different aspects of the same problem, whether we knew it or not. The scope of this problem includes various forms of cumulation of data from rolling samples, as well as the question of how to combine data from different countries into statistics for larger entities such as the European Union. Kish (2001) suggests that these problems have fundamental features in common with the problem of combining information from different experiments (Cochran 1937, 1954).

[1] Charles H. Alexander, U.S. Bureau of the Census, Suitland, Maryland, U.S.A. 20233.

3. THE CENSUS LONG FORM AND INTERCENSAL ALTERNATIVES

The decennial census "long form" survey is the main source of subnational data about the *characteristics* of the U.S. population and housing. Estimates of the *number* of people and housing units come from the "short form" part of census administered to all households. With an overall sampling rate of one-in-six, the long form survey provides precise, detailed ("Precise" refers to the sampling error, and "detailed" means that estimates are given for many demographic domains within the geographic domain.) estimates of a variety of demographic and economic characteristics for individual states, large cities, and large counties or groups of counties. It provides useful, though less precise and less detailed, estimates for even very small areas such as small towns and Indian Reservations, as well as census tracts, which average about 4,000 population. For the smallest governmental units, higher sampling rates are used, as high as one-in-two for the smallest places, so that there are usable estimates for these areas. To compensate for the higher sampling rates in these areas, the rate is one-in-eight in the largest tracts.

Between the censuses, the federal government's statistical programs provide relatively little information about the characteristics of the population below the national level. The basic census counts are updated by an intercensal demographic estimates program, but other demographic and economic characteristics are available mainly from national surveys. The Current Population Survey (CPS), the U.S. monthly labor force survey, has about a one-in-1000 sampling rate with substantial overlap in the sample units from one month to the next so that the sample cannot be profitably cumulated over time as a rolling sample can. A March Supplement to the CPS collects additional information once a year, providing estimates for income and poverty at the state level, but with limited precision and demographic detail. There are programs which use modeling methods based on administrative records to make small-area estimates for unemployment, and for income and poverty, but not for a variety of characteristics.

The need for more frequent information for smaller domains (or "communities") has long been recognized (Hauser 1942; Eckler 1972, page 212; Bounpane 1986). Leslie gave credit to his friend, Philip Hauser, for proposing an "annual sample census" in 1941. Kish (1981) proposed a rolling sample as a way to meet this need, presenting several options including a rolling sample for the CPS. Instead a mid-decade census was authorized for 1985, but it was never funded. Nor was a proposal to double the size of the CPS (Tupek, Waite and Cahoon 1990).

Interest at the Census Bureau in intercensal information about population characteristics was revived by a proposal for a "Decade Census Program" advanced by Herriot, Bateman and McCarthy (1989). This program would have collected data in different states in different years; ultimately this proposal did not gain acceptance. However, Roger Herriot's energetic and eloquent advocacy of the importance and potential value of intercensal subnational data created awareness in federal statistical agencies of the possibility of a "new paradigm" for the decennial cycle of data collection. Awareness of Kish's rolling sample proposal was definitely a factor during this period, as the Bureau considered new approaches for the 2000 census (see Bounpane 1986).

There was renewed Congressional interest in intercensal characteristics data (Melnick 1991; Sawyer 1993), and a "continuous measurement" alternative to the census long form was considered as part of the research for Census 2000, starting in 1992. Kish's rolling sample design was eventually proposed for this purpose because it provided flexibility in making estimates, as well as the potential for efficient data collection (Alexander 1993, 1997; National Academy of Sciences 1994, 1995). My recollection is that the most influential articles were Kish (1981, 1990), and that Kish and Verma (1983, 1986) were also consulted. "Continuous Measurement" was later renamed the "American Community Survey (ACS)".

The proposed ACS was not adopted for Census 2000, but after limited testing during 1996-1998, the ACS methodology was implemented in 36 counties for the years 1999-2001, so that ACS results could be compared to the 2000 census long form data. There was also a large-scale test in 2000, for a state-representative annual sample of about 700,000 addresses called the Census 2000 Supplementary Survey, of collecting long-form data separately from the census, using the ACS questionnaire. In 2001 and 2002, the Supplementary Survey is being continued, as part of the transition to the ACS.

4. THE PLANNED AMERICAN COMMUNITY SURVEY

The ACS will start in 2003, if funded by Congress, with a monthly sample of about 250,000 addresses, a new panel of addresses starting each month. This corresponds to a monthly rolling sample with an average rate of approximately $F = 480$ or an annual sample with $F = 40$. The survey will use $k = 60$, with the shortest published cumulation being calendar year estimates. The ACS will be conducted by mail, with nonresponse followup by telephone. A random sample of one-third of the remaining nonrespondents will be selected for followup in person.

For domains with average response rates, with a monthly $F = 480$, the standard errors for a 5-year average estimate from the ACS will be somewhat larger than for a corresponding estimate from the census long form, typically on the order of 1.33 times as large. This was judged to be "sufficiently close" for most purposes, given the advantage of timeliness and the expected lower missing data rates due to having a permanent staff of interviewers. In areas with

lower-than-average mail response rates, the subsampling for nonresponse follow-up will reduce the effective sample size. This happens not only because the number of interviews is reduced, but also because the unequal weights typically lead to a higher design effect (Kish 1965, pages 429-431). To compensate for this, the ACS will probably use a higher nonresponse subsampling rate in low-response areas, balanced by a lower sampling rate in areas with higher-than-average mail response. The details of this are still being determined. There also will be an oversample of addresses in small governmental units, as with the census long form sample.

An important development in the last decade, that made the ACS possible, (Kish (1981) suggests an alternative approach of "cumulative rolling listings", but this would be quite expensive for making regular estimates for all of the smallest domains, such as census tracts.) is the Census Bureau's program to maintain an ongoing Master Address File (MAF), linked to our TIGER geographic database. The main source of address updates throughout the decade is the Postal Service's Delivery Sequence File (DSF). The Bureau is implementing a MAF/TIGER modernization program that will augment the DSF updates with new addresses from data files provided by local governments, and from other administrative sources. This will be supplemented by new addresses encountered by interviewers from the ACS and other surveys in more rural areas. The monthly samples are actually generated by selecting an annual sample from the MAF in the previous September, and dividing it into 12 monthly panels. In February, there is a supplemental sample of new units from the DSF, spread across the remaining months of the year.

Replacing the 2010 census long form, by the ACS, is one component of a program to re-engineer the 2010 census. This also includes the modernization of MAF/TIGER, as well as a program of early research and testing to automate, streamline, and improve the census operations for 2010. This combination of improvements is expected to have a budgetary cost for the full 10-year cycle that is less than the cost of repeating the Census 2000 methods in 2010. This is a quite different plan than the vision of ACS described in National Academy of Sciences (1994, Chapter 6; 1995, Chapter 6), where I expressed hopes that eliminating the long form by itself, without other fundamental improvements, might save enough to pay for the ACS.

5. SOME VARIATIONS ON THE BASIC DESIGN, AND SOME ISSUES

5.1 Multi-stage Cluster Samples

The ACS uses an unclustered one-stage systematic sample, because the goals include providing data for all small geographic domains, such as tracts or block groups, each year. From discussions in Kish (1981, 1998), it is clear that rolling samples can also use cluster samples and multiple stages of selection, as well as varying probabilities of selection. However, to qualify as a "rolling sample", the primary sampling units themselves must be a rolling sample. A design with a fixed set of primary sampling units (PSUs), with a rolling sample within each PSU, is a "cumulated representative sample" (Kish 1998).

Leslie was emphatic that the proposal by Herriot *et al.* (1989), was not what he meant by "rolling sample". However, it would seem to fit the definition as stated in Section 2, if the states are considered as PSUs. I think this demonstrates that there is an implicit requirement that a rolling sample must yield a useful representative probability sample in each time period, for each geographic domain of interest; this additional requirement does not hold if the PSUs are states. This requirement means that the clusters or PSUs need to be substantially smaller than the smallest domain of interest. (See Kish 1998, page 38.)

5.2 Differential Sampling Rates

Kish (1998, section 4) notes that a rolling sample can use different sampling fractions in different strata. This can get complicated, especially if the sampling fractions change over time, because the conditional probability of selecting a unit (without replacement) for the j^{th} panel in the h^{th} stratum depends on the sampling rates used in the previous panels in that stratum. This is even more complicated if the strata change over time, for example as the boundaries of governmental units change.

To simplify this for the ACS, we select the sample in two stages. The first stage selects a rolling "super sample" using a constant sampling rate for each panel and each year, equal to the largest sampling rate required in any stratum. The second stage subsamples the initial sample, to give the desired sampling rate for each stratum for that year. The selection of subsequent samples, which avoids overlap with the entire previous supersamples, needs only to keep track of the sampling rate for the first stage.

5.3 Updates to the Frame

In practice, the population is a little different for each panel. New addresses are added to the frame. Some old addresses cease to exist; they may be removed from the address list, or they may stay on the list and be deleted only after attempts to contact them. This presents no fundamental conceptual problem. It does mean that a "rolling census" would not necessarily contact every population unit that ever exists, since some units may go in and out of existence too quickly to fall into sample.

To avoid record-keeping of different conditional sampling rates for different "cohorts" of addresses which were added during Master Address File updates at different times, we have found it convenient to assign artificial "back samples" by selecting addresses from each set of new addresses not only for the current panel, but for past panels. These units are not interviewed, since the times for their assigned panels are past, but they are avoided during the without-replacement selection of future panels.

5.4 What Happens After Panel k?

One question Leslie did not address explicitly, as far as I know, is how to draw the sample for panel k+1. I think he assumed that panel k + 1 would be the same as panel 1, panel k + 2 repeats panel 2, and so forth. This works fine for a simple random sample, but not so well for a systematic sample intended to spread the sample over a geographically sorted list, because as the frame changes over time, panel 1 doesn't keep its even spacing.

Our plan is to select panel k + 1, and future panels, as a fresh systematic sample. Each one will avoid overlap with the previous k - 1 panels, so there will always be k consecutive non-overlapping panels, but we won't worry about overlapping with panels before that.

5.5 Questionnaire Reference Date, Given an Extended Interview Period

The interviews from each monthly ACS panel take place over a three-month period, allowing two months for mail returns and telephone followup before starting the more expensive personal visits in the third month. Thus, the data actually collected in June consist of early mail returns from the June panel, late mail returns and telephone interviews from the May panel, and personal-visit followup cases from the April panel. This raises the issue of whether to ask the survey questions as of the time the survey was mailed out – the best choice as far as sampling bias – or as of the time the questions are asked – the best choice as far as response error and other nonsampling errors, especially for people who have moved from the address.

Taking these quality tradeoffs into account, we chose to use a "current" reference date, collecting the characteristics of the household members at the time of interview. One reason for this decision is that we think the nonsampling errors will be harder to evaluate than the sampling bias. Also the sampling biases in the monthly estimates will tend to cancel over the course of the year. This is one reason for limiting the ACS to annual and multiple-year estimates.

5.6 Use of Intercensal Population Estimates as Survey Weighting Controls

The Census Bureau has a program of "intercensal" (Leslie would call these "post-censal" estimates, reserving "intercensal" for estimates between two censuses that have been completed.) demographic estimates, based on demographic models. These models update the previous census, using vital records and other administrative records information. These estimates are used as independent weighting controls, or "post-stratification factors", for most national household surveys (see Kish 1965, pages 90-92). Adjusting the survey weights to agree with controls can reduce the variances of survey estimates, adjust for differences in coverage by age, sex, race, or Hispanic origin, and improve consistency across surveys. The census long form similarly uses the census counts as controls in its weighting.

The weighting controls have traditionally not been available for the smallest geographic domains, at least not with the demographic detail available for larger areas. Plans to produce more detailed controls for use in ACS weighting are described in Alexander and Wetrogan (2000). Some improvements will come from improved sources of administrative data, but in addition the ACS itself will provide information on changes in the population, which can be incorporated into the demographic models. The problem is complicated by the differences between the "current resident rule" used in the ACS and the "usual resident rule" used in the census; the ACS includes a question about part-year residents to help in adjusting for this difference. To facilitate this integration of survey data and demographic models, and especially to develop error measures for the resulting estimates, the Census Bureau is trying to develop "statistical" versions of the demographic models used in producing the intercensal population estimates. The inspiration for this effort to blend the statistical and demographic approaches is Purcell and Kish (1979).

6. DIFFERENT CUMULATIONS FOR DIFFERENT PURPOSES

For the main ACS objective, to replace the census long form as a source of detailed descriptive statistics, we plan to use 5-year ACS cumulations, for a data product similar to traditional long form "summary files". This is the shortest time period for which the ACS sampling error is judged to be reasonably close to that of the census long form. All sizes and types of geographic areas would be included on these 5-year data files. For allocating government funds based on an assessment of current need for the funds, simulations suggest that 3-year cumulations may be preferable to the 5-year, sacrificing precision for greater recency (Alexander 1998).

For individual areas, the most prominently published data will be one-year averages for areas greater than 65,000 population, and 3-year averages for areas greater than 20,000, in addition to the 5-year averages for all areas. Annual average estimates for areas below these thresholds will be available for more "sophisticated" uses to use in time series models, and to indicate large variations within the multi-year averages, but will not be as prominently displayed in our publications or on our websites.

These planned published ACS data products are designed to encourage analysts to use the same length of cumulation when comparing areas of different sizes, on the grounds that to do otherwise may be perceived as unfair to smaller jurisdictions. In doing this, we have accepted the notion of "asymmetrical cumulations" as far as levels of geography, but not necessarily within the same level of geography. For example, we would use one year for comparing states, but would recommend 5-years for all the

counties in a table comparing large and small counties. In this latter recommendation, we differ somewhat from Kish (1998, pages 42-43) which would let us use tables of counties with one-year estimates for large counties, 3-year averages for medium-sized ones, and 5-year averages for small ones. It will be interesting to see what practices data users will adopt in this regard.

7. WEIGHTING THE YEARS IN MULTI-YEAR CUMULATIONS

Kish (1998) points out that there are a number of choices for weighting multi-year cumulations. If there are ten yearly means $\bar{y}_i$, then there are many choices of $\bar{y} = \sum w_i \bar{y}_i$, with $\sum w_i = 1$, to use as the ten-year cumulations.

For the ACS 5-year and other multi-year cumulations, discussed in section 6, our plans are to give the years equal weights in the standard published data products, *e.g.*, $w_i = 0.2$ for the 5-year average. This was an area of disagreement with Kish (1998), which gently urges us to consider of alternatives, in particular weights of the form $w_{i+1} = C w_i$, with $C > 1$.

An underlying issue in thinking about unequal weights is what statistical problem we are trying to solve. Using the 2003 – 2007 cumulation as an example, is the goal:

- to provide a "direct design based" estimate for the 2003 – 2007 historical average;
- to provide a "model-based" estimate for the 2007 value; or
- to provide a "direct, design-based" estimate for a weighted 2003 – 2007 historical average, with more weight on recent years?

To interpret the 2003 – 2007 estimate as an estimate for 2007 requires a model or assumptions about the time series for the area. The problem may be viewed as combining a direct estimate for 2007 with a forecast for 2007 based on the years 2003, ..., 2006, with the requirement that the same formula be used for all areas and all characteristics to preserve additivity in the tables and comparability across tables.

I have previously interpreted the decision as a choice between the first two goals, and have shied away from the second approach for the ACS, ultimately because of the concerns expressed in Hansen, Madow and Tepping (1983, sections 3 and 5.5) about using model-based estimates for general-purpose "official statistics". With the variety of statistics and geographic areas covered by the ACS, there inevitably will be some where the compromise model fails badly; a data user may be unaware of this failure, or may be very aware. In what sense can the compromise average be viewed as a valid estimate for 2007 when the compromise model clearly fails, and what measure of error would be associated with it? With this view of the issue, we have recommended using the unweighted multi-year averages as the standard general-purpose data product, with the time series of annual estimates being available for use in time series models for specific applications, and for interpreting the multi-year averages when there is variation within the 5-year period.

However, upon rereading Kish (1998), I now interpret his view of the weighted average to be the third formulation, a design-based estimator of a more up-to-date population parameter. This avoids the concerns about model fit for general-purpose uses, although there is still the question of how to justify and achieve a consensus solution. Also, the unequal weights tend to increase the standard errors of the multi-year averages. But Kish (1998, page 40) will get the last word on the subject:

> "*Important questions remain for further discussions and research. Perhaps forever, and this can become a 'growth industry.'* "

8. NOT COMBINING THE CPS AND THE ACS

Leslie often said he was pleased to see his idea being implemented in the ACS, but I think he was disappointed that we did not try to replace both the census long form and the CPS with one survey. By contrast with some other issues where we had lively discussions, Leslie took a "hands off" stance on this issue. I think he viewed this as a decision about quality tradeoffs, which the government agencies had to work out for ourselves. There were two main reasons for our decision:

We cannot adequately measure the monthly unemployment rate with a mail survey. Correct measurement of the unemployment rate requires complex questions that would not be feasible to ask by mail, for example, to probe to be sure that someone who is "looking for work" did conduct an active job search. (See Butani, Alexander and Esposito 1999). The Census 2000 Supplementary Survey, using the ACS procedures, dramatically overestimated the 2000 national unemployment rate (5.3 percent versus 4.0 percent in the CPS). A similar difference was seen in the 1990 census.

A mail survey would lag substantially in producing monthly rates, compared to the CPS. In addition, the impossibility of completing all the mail interviews for a panel in the designated month introduces biases in monthly estimates (see section 5.5 above). These problems would be reduced somewhat for quarterly moving averages instead of monthly estimates, which Leslie frequently suggested (for example Kish 1999), but the monthly unemployment report is an indispensable economic indicator in the U.S.

It is too expensive to replace the long form without using mail. A rolling sample survey, conducted in person with a large enough sample to replace the long form, would have to be 3 or 4 times as large as the CPS. This is a function of

the size of the U.S. population, and the number of tract-sized domains for which estimates are required from the long form. Such a survey would be much more expensive per case than the CPS, because it could not use a cluster sample or telephone interviews for repeated interviews of the same households, as does the CPS. The total cost of such a survey would be several times as great as the combined cost of the proposed ACS and the CPS.

Because it is designed so narrowly as a long form replacement, the ACS does not illustrate the full range of flexibility that Leslie envisioned from a rolling sample. Under different circumstances, for a smaller population, with less need for very small domains from the "long form survey", or less strict requirements for timing and questions for the labor force survey, it might be possible for a labor force survey with a rolling sample to meet the demands for small domain data. With the further addition of a split panel or other components (Kish 1998, pages 40-41) an even wider range of objectives could be met.

9. CONTRIBUTIONS: PHILOSOPHICAL, PERSONAL, AND PRACTICAL

The long list of articles by Leslie Kish on the subject of rolling samples clearly demonstrates the intensity and tenacity of his campaign for what he understood as an important idea. The evolution of the idea over the course of these papers also illustrates the depth of his attention to "philosophical" questions about the fundamental quality objectives for a survey: What are we trying to do? How does the choice of survey design relate to what we are trying to do, and why? This kind of guidance is crucial at the start of a survey program, when the "big questions" are being addressed, and makes the difference between ideas that quickly fall by the wayside and those that are "still rolling".

Leslie's personal support of other statisticians went far beyond his papers. Though I was by no means one of his closest colleagues, he regularly provided personal advice or encouragement when he sensed it was needed. The "still rolling" in this paper's title was the title I used in e-mail messages to him when I had news about the ACS's perilous passage through the annual budget cycle, which was most of the time. He would respond briefly by e-mail, but important messages always came in the form of handwritten letters.

Finally, based on these papers, it is clear that Leslie was always a practical person, even at his most philosophical, and that his papers cannot be fully appreciated without knowing what was going on in the survey world when he wrote them. Looking back over his rolling sample papers, I can see many comments, about both details and general principles, that were aimed at enlightening specific decisions that the Census Bureau needed to make at the time. I would guess that throughout his work, there are specific messages to help out someone somewhere in the world who faced a practical design decision at the time.

REFERENCES

ALEXANDER, C.H. (1993). A continuous measurement alternative for the U.S. Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 486-491.

ALEXANDER, C.H. (1997). The american community survey: Design issues and initial test results. *Proceedings of Symposium 97, New Directions in Surveys and Censuses*, 187-192.

ALEXANDER, C.H. (1998). Recent developments in the american community survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 92-100.

ALEXANDER, C.H., and WETROGAN, S. (2000). Integrating the american community survey and the intercensal demographic estimates program. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 295-300.

BOUNPANE, P. (1986). How increased automation will improve the 1990 census. *Journal of Official Statistics*, 4, 545-553.

BUTANI, S., ALEXANDER, C. and ESPOSITO, J. (1999). Using the american community survey to enhance the current population survey: Opportunities and issues. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Statistical Policy Working Paper 29, 3, 102-111.

COCHRAN, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (B)*, 4, 102-118.

COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

ECKLER, A.R. (1972). *The Bureau of the Census*. New York: Praeger Publishers.

HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 384, 776-793.

HAUSER, P.M. (1942). Proposed annual census of the population. *Journal of the American Statistical Association*, 37, 81-88.

HERRIOT, R.A., BATEMAN, D.B. and MCCARTHY, W. F. (1989). The Decade Census Program - A new approach for meeting the nation's needs for sub-national data. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-355.

KISH, L., LOVEJOY, W. and RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.

KISH, L. (1979b). Rolling samples instead of censuses. *Asian and Pacific Census Forum*, G(1), August 1979, 1-2, 12-13.

KISH, L. (1981). *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*. Washington, D.C., U.S. Government Printing Office.

KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright). New York: Academic, 72-84.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 1-12.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-79.

KISH, L. (1997). Periodic and rolling samples and censuses. Chapter 7 in *Statistics and Public Policy*, (Ed. Bruce D. Spencer). Clarendon Press, Oxford.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 1, 1998, 31-46.

KISH, L. (1999). Combining/cumulating population surveys. *Survey Methodology*, 25, 2, 129-138.

KISH, L. (2001). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, to appear in 2001.

KISH, L., and VERMA, V. (1983). Census plus samples: Combined uses and designs. *Bulletin of the International Statistical Institute*, 50(1), 66-82.

KISH, L., and VERMA, V. (1986). Complete Censuses and Samples. *Journal of Official Statistics*, 2, 381-93.

MELNICK, D. (1991). The census of 2000 A. D. and beyond. *Reviews of Major Alternatives for the Census in the Year 2000*. U.S. Government Printing Office, Washington, D.C., August 1, 1991, 60-74.

MOONEY, H.W. (1956). Methodology in two California Health Surveys. *U.S. Public Health Monograph*, 70.

NATIONAL ACADEMY OF SCIENCES (1994). *Country People in the Information Age*. (Eds. D.L. Steffey and N.M. Bradburn). National Academy Press, Washington, D.C.

NATIONAL ACADEMY OF SCIENCES (1995). *Modernizing the U.S. Census*. (Eds. B. Edmonston and C. Schultze). National Academy Press, Washington, D.C.

PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

SAWYER, T. C. (1993). Rethinking the census: Reconciling the demands for accuracy and precision in the 21st century. Presented at the research conference on undercounted ethnic populations, May 7, 1993.

TUPEK, A. R., WAITE, P. J. and CAHOON, L. S. (1990). Sample expansion plans for the current population survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 72-77.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 43-49
Statistics Canada

Redesign of the French Census of Population

JEAN-MICHEL DURR and JEAN DUMAIS[1]

ABSTRACT

Census-taking by traditional methods is becoming more difficult. The possibility of cross-linking administrative files provides an attractive alternative to conducting periodic censuses (Laihonen 2000; Borchsenius 2000). This was proposed in a recent article by Nathan (2001). INSEE's redesign is based on the idea of a "continuous census," originally suggested by Kish (1981, 1990) and Horvitz (1986). A first approach that would be feasible in France can be found in Deville and Jacod (1996). This article reviews methodological developments since INSEE started its population census redesign program.

KEY WORDS: Balanced sampling; Census; Continuous census; Calibration.

1. INTRODUCTION

1.1 Reasons for the Redesign

France has been conducting censuses for many years to measure the de jure population of its administrative districts and to describe the socio-demographic characteristics of its territory at all levels of geography, from districts of communes to the country as a whole. The 1999 census was conducted in the usual manner: delivering and retrieving questionnaires by census interviewers, organisation, technical assistance and control by INSEE, execution by the Mayor as the state representative. For various reasons, however, we decided to re-examine the census.

First, the interval between censuses has a tendency to increase in length. Indeed, the periodicity of censuses is not covered by laws, and each census date is determined by a statutory order. Before the war, censuses were taken every five years; then the gap grew to seven years, then eight, the last census, originally planned for 1997, was postponed until 1999 for budgetary reasons, that is, 9 years after the previous census. Moreover, the public does not always understand the need for such a massive operation at a time when the number of administrative files is increasing, even though that same public has expressed serious concerns about the cross-referencing of such files. In addition, the decentralization that has been going on in France for over 20 years has generated numerous requirements for statistics in support of local policy-making. As the supreme source of local information, the census had to adapt to these changes and provide fresher yet still highly detailed data.

As a result, a population census redesign program was established at INSEE in the late 1990s. Since France has no population register and, in view of the circumstances, is unlikely to institute one, the decision was made to consider a compromise solution that would combine annual sample surveys with the use of non-nominative administrative files that INSEE is authorized to use solely for statistical purposes. Communes whose population is below a certain threshold (10,000 for the moment) will be covered by annual take-all surveys with a rotation period of five years. For the other communes, a sample survey will be conducted each year, with the entirety of the commune being covered within the same five-year rotation period. To carry out this redesign, a new legal framework was needed. The project was submitted to the Conseil d'État, which recommended on July 2, 1998, that the government table draft legislation in Parliament.

Aside from the need to strenghten the census legal basis, the Conseil was of the view that since population counts were referred to in over 200 statutes or regulations, making a major change in the way they were produced would require legislative approval. Within this framework, the purpose of the legislation was essentially to set out the principles and rules governing the organization of the census.

The operation was placed under State responsibility and control: INSEE was to establish the collection framework (concepts, protocols), select the samples, ensure the quality of the information collected, and process and disseminate the data. The communes as local organisations, were to prepare and conduct the census surveys. The State would provide financial assistance to cover the costs. These arrangements clarity the role and responsabilities of each of the partners.

1.2 Quality Goals

The program has the following quality goals:

1.2.1 Data Quality

Timeliness: The goal is to be able to disseminate by the end of year A the de jure population of all administrative districts as at January 1 of year A-2; a statistical description of all geographic units (communes and commune groups, districts of major cities, lands, *etc.*) as of January 1 of year

[1] Jean-Michel Durr, Programme de rénovation du recensement de la population, INSEE, Direction générale, 18 boul. Adolphe Pinard, 75675 Paris CEDEX 14, France; Jean Dumais Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. This paper was prepared while the author was on secondment at the Programme de rénovation du recensement de la population, INSEE.

A-2; and a statistical description of France and its major geographic units (regions, *etc.*) as of January 1 of year A. In comparison with the general census, the redesigned census will produce similar population and housing data an average of three to four years earlier.

Relevance: The data produced must be relevant to local needs. In particular, data that are worth studying only at levels of geography far above the commune will be set aside in favour of data that are more useful for local purposes. What data will be collected will be determined by the Conseil national de l'information statistique (CNIS), whose membership includes representatives of various categories of producers and users of public statistics. A CNIS working group has proposed changes while at the same time preserving the necessary continuity with previous censuses and limiting the response burden.

Precision: The census must provide data that are meaningful for all levels of geography in France. The data produced must be sufficiently precise, even at the sub-communal levels, for the most useful cross-tabulations at those levels. This means, in particular, distributions by sex and age, by type of activity and socio-professional category, and by type of housing. It must be possible to estimate the precision of the data, and users must be informed of that precision.

User-friendliness: To avoid annoying users, the data produced must be easy to understand and comparable in use to data produced by a general census.

1.2.2 Process Quality

Response burden: To limit the response burden for the public, the amount of information collected must be kept to a bare minimum. In particular, information available for the same level of geography from other sources will not be collected in the census unless it can be used to produce useful cross-tabulations with other variables. As in previous censuses, the personal questionnaire will be confined to one double-sided sheet of paper.

Questionnaire: Since collection is by the drop-off/pick-up method, the questionnaires must be universally accessible. To ensure that the questions will be understood, qualitative testing was conducted using focus groups. In addition, a collection test was carried out on 4,000 dwellings in the first half of 2001.

Confidentiality: Data gathered in the census are protected by law. Personal information collected in the census can be accessed only by authorized persons. The data are for INSEE and can be used only for statistical purposes. Only data essential to the preparation and conduct of census surveys are shared with communes or commune groups, on a need-to-know basis.

Quality of coverage: The coverage of general censuses was not systematically evaluated. Following the 1990 census, a postcensal survey indicated that the rate of under-coverage was about 1.8% and the rate of overcoverage was about 0.9%, for an overall precision of roughly 0.9%. The largest undercounts were in large agglomerations. By conducting an annual sample survey in communes with a population of more than 10,000 and thereby reducing the number of people to be covered in the census, we will be able to focus our efforts on obtaining answers from respondents. The coverage of the redesigned census will be evaluated on a regular basis through comparison with administrative data and through special surveys.

Technical and organizational robustness: Because of the volume of data processed and the importance of the census, the program must be based on tried and true technical innovations. Furthermore, the robustness of the census apparatus must be evident in the launch of the operation. Technical or functional innovations can be introduced at any time in the census cycle as part of evolving maintenance or specific projects. The annual surveys can be used to test the effectiveness of such projects before they are applied to the entire process. However, major changes such as questionnaire updates will generally be made only for the beginning of a five-year cycle. The organization of the census will depend on a balanced partnership between INSEE and the communes. INSEE must be capable of building the proposed structure within its budget and its work program by reorganizing its operations. Similarly, the communes and intercommunal cooperation bodies must be able to support the census organization. The yearly cycle of surveying large communes and the option that small and medium communes will have of delegating collection to an intercommunal body are likely to promote the professionalization of collection workers.

With the integration of census operations into the annual work program of the regional offices, and the fact that the operation is one-seventh the size of the general census, INSEE will have tighter control of the census. Instead of having 110,000 census agents collecting data from 60 million people in 36,700 communes in a particular year, it will have only 18,000 agents visiting roughly 9 million residents in about 8,000 communes.

The division of responsibilities between INSEE and the communes, the resources that the communes will require, and the validation processes for the various stages will be set out in a decree.

Cost control: With the five-year collection cycle, the financial burden of conducting the census can be spread over a longer period. For communes with a population of more than 10,000, the cost of the redesigned census will be lower than the cost of the current census of population. On the other hand, for communes with fewer than 10,000 residents, the cost should be equal to that of a general census, but it would be every five years instead of the roughly eight-year cycle of the general census. The cost of the redesigned continuous census will be equivalent to one seventh of that of a general census. This will contribute to archive the reform without budget increase. However, a slightly larger budget in the first few years would help to iron the kinks out of the collection process.

2. SAMPLING STRATEGY

The commune is the linchpin of the redesign effort. The set of "small and medium-sized communes" (those with a population of less than 10,000) will be sampled at an average rate of 20% a year, and all their dwellings will be visited; all "large communes" will be visited annually, but only a fraction of their dwellings will be surveyed.

2.1 Small and Medium-sized Communes

Let's start with "small and medium-sized communes". In each region, five rotation groups of communes will be formed using data from the 1999 population census. They will consist of balanced samples (Deville and Tillé 1999, 2000) of the age-sex distribution of the communes' population. This approach should help minimize year-to-year variation due to sampling.

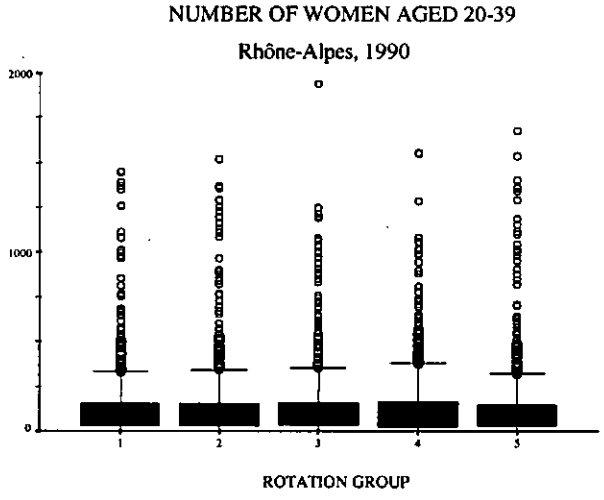


Figure 1

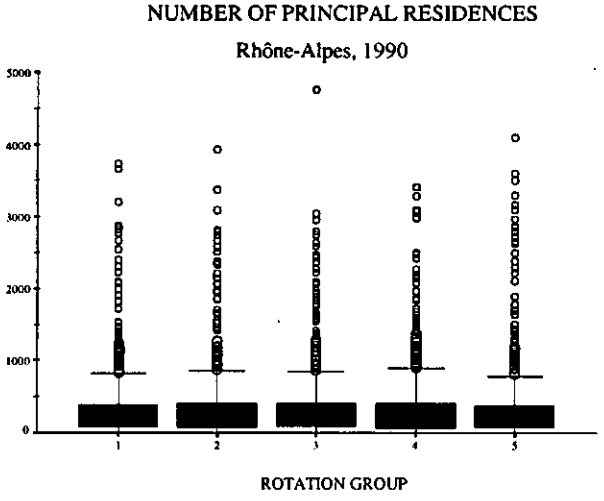


Figure 2

Figures 1 and 2 show how balanced the five rotation groups will be. They contain box-and-whisker diagrams of two variables measured in the 2,811 small and medium communes in Rhône-Alpes in the 1990 population census. For each rotation group, both the quartiles and the range of the distribution are shown. It is interesting to note how similar the charts are. The "number of women aged 20 to 39" variable was used to form the groups. Neither the number of principal residences nor any of the household or dwelling variables plays a part in the balancing.

Each year, the population and housing in all the communes in one rotation group will be fully enumerated. Hence, each "small and medium commune" will be completely enumerated once every five years, and a fifth of all the "small communes" will be covered each year.

2.2 Large Communes

The "large commune" sample will be based on the "répertoire d'immeubles localisés" (RIL) (inventory of located buildings). The RIL is a list of buildings (residential, institutional or commercial) identified individually so as to generate a digitized map. Initially, the RIL will be populated with data from the 1999 census, which will provide a statistical portrait of each residential building. (In the 1999 census, a building is defined as the set of dwellings served by the same staircase; thus, a single physical building can consist of more than one "census building".)

The RIL will be continually updated using building permits, demolition permits, utility records (water, gas, hydro, *etc.*), information supplied by local governments, and field observations. Thus, the RIL may be used to create a building sample frame for "large communes".

In each IRIS2000 (an IRIS2000 is a set of "îlots regroupés selon des indicateurs statistiques" (blocks grouped by statistical indicators), a homogeneous area with a population of about 2,000) of each "large commune", five rotation groups of addresses will be formed using the same sampling model as in "small and medium communes". Three additional strata will be created in each IRIS2000: one for industrial buildings (plants, warehouses, *etc.*), another for collective dwellings (institutions, group homes, communal groups, boarding schools, *etc.*) and a third for new addresses.

One fifth of the industrial buildings will be visited each year to verify that they contain no dwellings (custodian's quarters or space converted for habitation); any dwellings found in such buildings will be considered self-representing because of their special nature. All collective dwellings will be covered each year; 20% of them will be visited, and the population counts of the remaining 80% may be updated by telephone. Finally, all new residential buildings will be inserted in the rotation groups.

As noted above, each address rotation group will be visited once in each five-year period. A sub-sample of addresses, which corresponds to 40% of the dwellings of the group, will be selected. In each selected address, the complete dwelling content will be surveyed.

In summary, the annual sample will consist of some 8 million individual forms, 6 million from "small and medium communes" and 2 million from "large communes".

3. OVERALL AND DETAILED ESTIMATES

In the continuous census system, three sets of estimates will be produced and published each year: a set of de jure population estimates, a set of detailed estimates (from which the de jure population estimates will be derived) and a set of overall estimates that will be used to calibrate the detailed and de jure population estimates.

3.1 Overall Estimates

According to current dissemination plans, the national and regional results of the survey conducted at the beginning of year A will be published on December 31 of year A. These estimates will be the overall estimates for year A. In addition, the results for each "small and medium commune" visited during the year A collection campaign will be published on the same date.

3.2. Detailed Estimates

Administrative files will supply additional information at a sufficient level of detail. It will then be possible to measure the systematic error between what has been observed and what is in the files for similar objects (buildings, blocks, *etc.*). This systematic error in carefully chosen aggregates can be used to produce an adjustment factor which will then be applied to the administrative data to ensure that their adjusted totals match the census estimates.

Current plans are to use administrative files at a level of geographic aggregation (building, block, census agent district, *etc.*) that will provide information about individuals (age and sex according to health insurance files) or their dwellings (property tax files).

Detailed results for year A-2 will be released on December 31 of year A. (Aquisition and processing of administrative files are expected to take about two years.) These detailed results will be a blend of survey data (large communes) or census data (small and medium communes) with synthetic data.

The synthetic data will be obtained from the relationship between observed data and administrative data for the same point in time and space. For example, for commune C of Group II enumerated in year A-3 (census count denoted $R_{C,II}^{A-3}$), the imputed census count for target year A-2 will be given by

$$\tilde{R}_{C,II}^{A-2} = R_{C,II}^{A-3} \times \frac{\mathrm{Adm}_{II}^{A-2}}{\mathrm{Adm}_{II}^{A-3}} = R_{C,II}^{A-3} \times \frac{\sum_{c \in II} \mathrm{Adm}_c^{A-2}}{\sum_{c \in II} \mathrm{Adm}_c^{A-3}},$$

where Adm_c^a is the value derived from administrative sources for commune c and year a.

In the continuous census, for a "small and medium commune" surveyed in years A-5 and A (see the table below), person variables (age, sex, labour force activity, occupation, *etc.*) and dwelling variables (household size, number of rooms, tenure, conveniences, *etc.*) will be measured at two points in time.

In addition, for Groups IV and V, the synthetic estimates for year A-2 could benefit from the information collected in the campaigns for years A-1 and A respectively. Adjustment factors could be computed in relation to the most recent census and used to produce backward projections for the intercensal period. For example, for commune D in Group IV, we can compute the following:

$$\Theta_1 = R_{D,IV}^{A-6} \times \frac{\sum_{c \in IV} \mathrm{Adm}_c^{A-2}}{\sum_{c \in IV} \mathrm{Adm}_c^{A-6}} \text{ et } \Theta_2 = R_{D,IV}^{A-1} \times \frac{\sum_{c \in IV} \mathrm{Adm}_c^{A-2}}{\sum_{c \in IV} \mathrm{Adm}_c^{A-1}}.$$

It is virtually certain that these two series, extrapolations and backward projections, will not match. Nevertheless, it is best to publish just one set of estimates for any area and any point in time. It makes sense to produce a "composite" series whose end points are tied to census data. The following linear combination may accomplish just that while giving more weight to the more recent survey data:

	A-6		A-5		A-4		A-3		A-2		A-1		A	
Gr I		Adm		Adm	R	Adm		Adm	?	Adm		Adm		Adm
Gr II		Adm		Adm		Adm	R_{II}^{A-3}	Adm	$\tilde{R}_{C,II}^{A-2}$	Adm		Adm		Adm
Gr III		Adm		Adm		Adm		Adm	R_{III}^{A-2}	Adm		Adm		Adm
Gr IV	R	Adm		Adm		Adm		Adm	?	Adm	R	Adm		Adm
Gr V		Adm	R	Adm		Adm		Adm	?	Adm		Adm	R	
Total	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm

$$\tilde{R}^{A-2}_{D,IV} = 0.2 \times \Theta_1 + 0.8 \times \Theta_2.$$

Similarly, for commune E in Group V, with Q1 and Q2 appropriately defined, we would have:

$$\tilde{R}^{A-2}_{E,V} = 0.4 \times \Theta_1 + 0.6 \times \Theta_2.$$

Adjustment factors Θ will have to be calculated for relatively detailed population strata, such as age-sex classes, so as to keep as much demographic and geographic flexibility as possible in the census adjustment. The quality of the administrative files and local disparities will dictate the level at which the adjustment can be made most conveniently (for départements, metropolitan areas, ...). The same process can be applied to large communes if we replace "small commune" with "address".

Finally, when every commune in every group has been imputed, the estimated total for a variable of interest from the imputed file (detailed estimates) is unlikely to match the total estimated from observations alone (overall estimates published two years earlier). It has therefore been decided that the detailed estimates will be calibrated on the overall estimates. Once again, the level of calibration will depend on local trends and the quality of the overall estimates.

3.3. De Jure Population Estimates

The de jure population estimates are the third set of estimates derived from the census. They are the population figures that are used, by law, to determine commune funding, electoral boundaries, the composition of municipal councils, *etc.*

The "total de jure population" of a commune includes persons

- whose principal residence is within the commune,
- who live in an institution or a collective dwelling located within the commune,
- who have a residence in the commune and live in an institution or a collective dwelling located in another commune but have kept a dwelling in their commune of origin,
- who live in a collective dwelling in another commune for work or live in another commune for education,
- who are attached to the commune for administrative purposes (itinerant workers, sailors and so on).

Clearly, these populations cannot be estimated until the entire territory of the commune has been covered, that is, until the detailed estimates have been produced.

3.4. Estimation of Sampling Variance

The global and detailed estimates will be accompanied by a measure of their statistical quality. Work on this project began in the fall of 2001. The preferred option at this time is to use reference tables, as is done in the Canadian Labour Force Survey, for example. The sampling variances will probably be obtained by resampling the frame.

3.5. Imprecision Due to Synthesis

In the section 3.2, we showed how collected data will be used to produce synthetic estimates: first, an extrapolation for an "old" census, for two rotation groups (I and II, say); then directly using the census results for a third rotation group (III, say); and finally, combining extrapolations and backward projections to calibrate the last two groups (IV and V, say).

This synthesis can be formalized using a non-response model (Särndal 1990). The annual campaign is similar to a take-all survey that has an 80% non-response rate, which is dealt with using ratio imputation. If we let s represent the whole sample, r the respondents and s-r the non-respondents, we have

$$y_{.k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{\beta} x_k & \text{if } k \in s-r \end{cases} \quad \text{with} \quad \hat{\beta} = \frac{\bar{y}_r}{\bar{x}_r}.$$

Thus, the imputation model is

$$\xi: \begin{cases} y_k = \beta x_k + \varepsilon_k. \\ E(\varepsilon_k) = 0 \\ V(\varepsilon_k) = \sigma^2 x_k \end{cases}$$

where the errors ε_k are not correlated. With such a model, under simple random sampling,

$$\hat{Y}_. = \frac{N}{n} \sum y_{.k} = \frac{N}{n} \left\{ \sum_r y_k + \sum_{s-r} \hat{\beta} x_k \right\} = \ldots$$
$$= N \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s$$

The uncertainty around estimation with imputation depends on the sampling errors and the quality of imputation model ξ:

$(\hat{Y}_. - Y)$	=	$(\hat{Y} - Y)$	+	$(\hat{Y}_. - \hat{Y})$
Total uncertainty	=	sampling uncertainty	+	incertainty of model

If we assume that the imputation is unbiased:

$$E_\xi E_s E_r (\hat{Y}_. - Y) = 0$$

we have,

$$V_{\text{total}} = E_\xi E_s E_r (\hat{Y}_. - Y)^2 = \ldots$$
$$= E_\xi E_s E_r (\hat{Y} - Y)^2 + E_\xi E_s E_r (\hat{Y}_. - \hat{Y})^2$$
$$= E_\xi V_s + E_s E_r V_\xi$$
$$V_{\text{total}} = V_{\text{sample}} + V_{\text{imputation}}$$

assuming that the design and response mechanism are independent from imputation. Using imputed data as if they were observed data to compute the estimate of V_s results in an underestimate of V_{sample}. In terms of expectation,

$$E_\xi(\hat{V}_s - \hat{V}_{.s}) = V_{dif}.$$

For the estimators of these variances, Särndal shows that we get

$$\hat{V}_{\text{sampling}} = N^2 \left(\frac{1}{n}\frac{1}{N}\right)\left\{S^2 + C_0\,\hat{\theta}^2\right\}$$

with C_0 close to $\left(1 - \frac{m}{n}\right)\bar{x}_{s-r}$ and $\hat{\theta}^2$ close to $\frac{\sum_r e_k^2}{\sum_r x_k}$ and

$$\hat{V}_{\text{imputation}} = N_2\left(\frac{1}{m} - \frac{1}{n}\right) A\,\bar{x}_s\,\hat{\theta}^2,$$

with $A = \bar{x}_{s-r}/\bar{x}_r$, which we can take as a respondent selection effect. We note that if $x_k = 1$, then we obtain a two-phase sample of size m in n and n in N. In addition, if $s = r$, $V_{\text{total}} = V_{\text{sampling}}$.

In Särndal's model, the x (administrative data) and y (census data) are contemporaneous; at the very least, we will have observed some of the y. Using the structure developed in the previous section, we would have:

Year A-2	
y_k	x_k
$y_{.k}$	x_k

m respondents (Group III)

n-m imputations (other groups)

In the continuous census system, not everything is synchronous:

... A-4		A-3		A-2		A-1	A
Y_I^{A-4}	X_I^{A-4}		X_I^{A-3}		X_I^{A-2}		
	X_{II}^{A-4}	Y_{II}^{A-3}	X_{II}^{A-3}	$Y_{.II}^{A-2}$	X_{II}^{A-2}		
	X_{III}^{A-4}		X_{III}^{A-3}	Y_{III}^{A-2}	X_{III}^{A-2}		
	X_{IV}^{A-4}		X_{IV}^{A-3}		X_{IV}^{A-2}	... ...	
	X_V^{A-4}		X_V^{A-3}		X_V^{A-2}	... ...	

That is, Y_{II}^{A-3}, X_{II}^{A-3}, $Y_{.II}^{A-2}$, and X_{II}^{A-2} are not all measured or observed in the same year. In fact, if we look at Group III on its own, for example, we have a sample of size n in year A-2 and an identical but totally non-respondent sample in year A-3. Consequently, some parameters in the estimate of V_{total} cannot be calculated.

On the other hand, if we take the problem over a specific period, we have a sample of size n and $4n$ non-respondents. We could approximate the uncertainty of the asynchronous imputation process (the process we have in the redesigned census) with the uncertainty of the synchronous imputation process (similar to Särndal's model).

This approach was tested on the small and medium communes of Rhône-Alpes, for which the rotation groups, 1990 property tax data and 1990 population census data are available (Kauffmann 2000). The method gives good results for variables that are highly correlated with property tax; the results also indicate that a source of administrative data that are similar to variables describing people will be necessary to maintain the model errors at an acceptable level.

4. WORK IN PROGRESS

The methodological work involved in redesigning the census is far from complete. The following projects are still under way:

- establishment of rules for crossing the size threshold, problems of oscillation around the 10,000 population threshold, and calculation of the de jure population;
- the sensitivity of stratum boundaries in large communes and their robustness over time;
- the updating and maintenance of sampling frames and samples, especially adjustments that may be required when a commune crosses the size threshold and the incorporation of new objects into rotation groups;
- massive imputation and synthesis, both models and their precision;
- estimation of the precision of estimators; and
- collecting data from mobile population groups.

REFERENCES

BORCHSENIUS, L. (2000). From a Conventional to a Register-based Census of Population. Les Recensements après 2001, Séminaire Eurostat-INSEE, Paris.

DEVILLE, J.C., and JACOD, M. (1996). Replacing the Traditional French Census by a Large Scale Continuous Population Survey. *Annual Research Conference Proceedings*, USBC, Washington.

DEVILLE, J.C., and TILLÉ, Y. (1999). *Balanced Sampling by Means of the Cube Method.* CREST-ENSAI, working paper submitted for publication.

DEVILLE, J.C., and TILLÉ, Y. (2000). Echantillonnage équilibré par la méthode du cube et estimation de variance. *Journées de Méthodologie*, December 2000, INSEE, Paris.

HORVITZ, D.G. (1986). Statement to the Subcommittee on Census and Population. Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.

KAUFFMANN, B. (2000). *Estimations annuelles dans la rénovation du recensement de la population*. Working paper, Département de la démographie, INSEE.

KISH, L. (1981). Population Counts from Cumulated Samples. Congressional Research Service. *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.

KISH, L. (1990). Rolling Samples and Censuses. *Survey Methodology* 16, 1, 63-71, Statistics Canada, Ottawa.

LAIHONEN, A. (2000). 2001 Round Population Censuses in Europe. *Les Recensements après 2001*, Séminaire Eurostat-INSEE, Paris.

NATHAN, G. (2001). Models for combining longitudinal data from administrative sources and panel surveys. Invited paper, *ISI*, Seoul, August 2001.

SÄRNDAL, C.-E.(1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Ottawa, October 1990, 337-350.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 51-61
Statistics Canada

Benchmarking Parameter Estimates in Logit Models of Binary Choice and Semiparametric Survival Models

IAN CAHILL and EDWARD J. CHEN[1]

ABSTRACT

An approach to exploiting the data from multiple surveys and epochs by benchmarking the parameter estimates of logit models of binary choice and semiparametric survival models is developed. The goal is to exploit the relatively rich source of socio-economic covariates offered by Statistics Canada's Survey of Labour and Income Dynamics (SLID), and also the historical time-span of the Labour Force Survey (LFS), enhanced by following individuals through each interview in their six-month rotation. A demonstration of how the method can be applied is given, using the maternity leave module of the LifePaths dynamic microsimulation project at Statistics Canada. The choice of maternity leave over job separation is specified as a binary logit model, while the duration of leave is specified as a semiparametric proportional hazards survival model with covariates together with a baseline hazard permitted to change each month. Both models are initially estimated by maximum likelihood from pooled SLID data on maternity leaves beginning in the period 1993-1996, then benchmarked to annual estimates from the LFS 1976-1992. In the case of the logit model, the linear predictor is adjusted by a log-odds estimate from the LFS. For the survival model, a Kaplan-Meier estimator of the hazard function from the LFS is used to adjust the predicted hazard in the semiparametric model.

KEY WORDS: Microsimulation; Benchmarking; Semiparametric survival models; Binary logit.

1. INTRODUCTION

Researchers often base econometric models on a survey conducted over a short period of time. In this case it may be desirable to incorporate information from a supplementary data source covering a longer period, even if measurements are only available for the dependent variable. For a broad class of non-linear models, we develop a simple method of benchmarking the parameter estimates obtained from a survey rich in explanatory variables to information from a survey with significant historical depth. A primary objective is that model predictions accord with information from the secondary data source. We demonstrate application of the method first to a simple logit model of binary choice, and secondly to a semiparametric survival model. Since the survival model can be viewed as a sequence of binary choices, while retaining an interpretation as an incompletely observed continuous time model, it provides a natural generalization of the first application.

The illustration we provide is a study of maternity leave. The Statistics Canada Survey of Labour and Income Dynamics (SLID) provides data on both the incidence of choosing a maternity leave over withdrawing from the labour force, and on the duration of maternity leave, as well as a rich set of explanatory variables. Because of this we use SLID to estimate base parameters, including those determining the effects of the explanatory variables on the incidence (the logit model) and hazard of returning to work (the survival model). The Canadian Labour Force Survey (LFS) conducted by Statistics Canada provides reasonable proxies for both the incidence and duration extending back to 1976. The SLID parameter estimates are therefore benchmarked to LFS estimates of incidence and the hazard of returning to work during the period 1976-1992, which is prior to the availability of SLID data.

The work was carried out while developing the maternity leave module of the LifePaths microsimulation model at Statistics Canada. The goal of the LifePaths project is to construct a dynamic microsimulation model encapsulating as much detail as possible on socio-economic processes in Canada, as well as the historical patterns of change in those processes. LifePaths has been employed in a broad range of policy analysis and research activities. Examples include Canada Student Loan policy (under contract to Human Resources Development Canada and the Government of Ontario), returns to education (Appleby, Boothby, Rouleau and Rowe 1999), time use (Wolfson and Rowe 1996; Wolfson 1997; Wolfson and Rowe 1998a), tax-transfer and pensions (Wolfson, Rowe, Gribble and Lin 1998; Wolfson and Rowe 1998b), and labour force careers (Rowe and Lin 1999). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1997; Rowe and Chen 1998; Plager and Chen 1999) and earnings correlation (Chen and Rowe 1999).

LifePaths is intended to incorporate socio-economic information from all relevant sources available to Statistics Canada. Consequently the construction of the model has motivated research into application of methodologies for exploiting multiple data sources. Embedding an estimated model in LifePaths is a powerful tool for deriving implications of the model that can be compared to information

[1] Ian Cahill, Partnership and Continuous Evaluations, HRDC, 140 Promenade du Portage, Phase IV 3rd floor, Room 3D475, Gatineau, Québec K1A 0J9, and Edward J. Chen, Household Survey Methods Division, Statistics Canada, R.H. Coats Building 16th floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

from other sources. For example, Rowe and Lin (1999) derived job tenures by simulation from a model estimated using short-period longitudinal data, then compared the results with data from a cross-sectional survey. We report on one aspect of the continuing effort to build a tool providing the maximum information that can be extracted from Statistics Canada's data sources.

The paper is organized to illustrate the way in which technical problems are often encountered in the course of building LifePaths, and how their solution is integrated with the model development process. To do this, a fair amount of background detail on associated issues is provided. Section 2 outlines the context of the benchmarking problem, and section 3 presents the theory behind our solution, with some possible extensions for further work. Section 4 describes the models to which it will be applied, including some details concerning the estimation of their parameters in the base period, then section 5 describes the application of the benchmarking method to these models. We display and discuss our empirical results in section 6, then close with some overall conclusions in section 7.

2. CONTEXT OF THE PROBLEM

We provide context in this section by presenting an overview of the LifePaths model structure, a brief description of data sources involved, and a discussion of how the benchmarking problem arose.

2.1 Structure of the LifePaths Model

The LifePaths model simulates individual lifetimes as a series of events which modify the set of "state variables" describing the demographic, social, and economic circumstances of the individual. Waiting times to every possible event are associated with an individual, although they may be infinite. The waiting times may be conditioned on the values of state variables. The event type with the shortest waiting time occurs (its associated functions are called). Modification of any state variable at the occurrence of an event may lead to the generation of new waiting times for other events.

LifePaths initialises a case by randomly generating a "dominant" individual's sex, province of residence, age at immigration and year of birth. The year of birth can range from 1892 to 2051. Mortality and immigration assumptions are designed to reproduce provincial age-sex structures. When a dominant individual marries, enters a common-law union, or has a child, a non-dominant individual of suitable characteristics is created and is linked to the dominant individual, forming part of the case. Once created, non-dominant individuals undergo the same possible events as dominant individuals. However, since their purpose is to complete the profile of the dominant actor, they are usually filtered from all tabular reports.

LifePaths presently includes models of fertility, mortality, marriage (including common-law unions), educational careers, labour force careers, maternity leave, hours of work, earnings, taxes, and transfers. The model of the labour force careers describes transitions between the states "paid employee," "self-employed," and "not employed." It also includes a model of retirement and student work. The model of secondary and post-secondary educational careers at the provincial level is mature and highly developed.

2.2 The Data Sources

The estimation of base parameters for the model of maternity leave was carried out using data from SLID covering maternity leaves beginning in the period 1993-1996. Using data from 1997 allowed us to follow most maternity leaves to completion rather than using extensively censored data. This is a household survey designed to permit both longitudinal and cross-sectional analysis of people's financial and work situations. Starting in 1993, SLID follows the same respondents for six years, with new rotation groups introduced every three years. Each rotation groups includes about 15,000 households with 30,000 adults. From this survey we obtain the month of child birth, monthly data on labour force status, and a rich set of explanatory variables including job tenure, an indicator of self-employment, birth order of the child, presence of an employed spouse, province of residence, education level, and age. We can also determine if a mother who left a job within 4 months of birth has returned to the same job within 16 months. This is used as a practical definition of maternity leave and becomes our unit of analysis, with a slight expansion to include the 1% of cases where a mother returned to a different job from a labour market state of absence in the previous month. Using this unit of analysis we get a sample size of 835 births. As we show in section 6, this sample size is adequate to reveal some key explanatory factors. More precisely, several factors are found to be significant at the 95% confidence level. This sample contains about 730 unique mothers, representing over 87% of the sample of births. This means that there will be some correlation between observations as a result of those mothers who have two or more maternity leaves within the observation period, but we did not feel that it is of sufficient magnitude to warrant any special statistical tools.

The LFS is a monthly household survey focussing on labour force status, and also reporting a number of demographic characteristics. The survey is normally used exclusively for cross-sectional analysis. For the LifePaths project, however, a file covering the period from 1976 to 1995 was constructed that follows individuals as they rotate through the six monthly rotation groups of the survey, providing a six-month window on each individual's labour market activity. Since the number and ages of children are recorded each month, it is possible to observe the

appearance of a new child. Since all surveys throughout the period are used, the sample size is very large, and about 26,000 births are observed.

In the LFS window we note the labour force status of a new mother when the child is first reported. This is the key to estimating the probability of choosing a maternity leave, rather than leaving the labour force. We begin by considering P(E), the proportion of such mothers who are employed. If the mother is "employed, at work," we suppose that they took a brief absence from their job – less than a month. If they are "employed, absent from work," it may be that they have chosen to take a maternity leave absence from their job and then return to it. However this may not always be the case. A new mother who we observe as employed and absent (*EA*) may later make a transition out of employment (to *NE*). To correct for this, considering mothers with a child of age less than a year observed in a window, we calculate the proportion $P(EA \sim NE)$ of transitions out of the "employed, absent from work" state that are to a not-employed state. We also estimate the proportion $P(NE \sim OJ)$ of mothers who return to an old job (*OJ*) after having left employment. The estimate is obtained by using observations on mothers with a young child who make transitions from a not-employed state to a job with a start date earlier than the previous month. Our estimate of the probability of choosing a maternity leave is now $P(E) - P(EA \sim NE) + P(NE \sim OJ)$.

It is also possible to observe mothers with a child of age less than a year making a transition from the status "employed, absent from work for personal or family responsibilities" to the status "employed, at work." We use this transition as a proxy for the return to work after a maternity leave. Since the duration of absence is reported in the previous month, this is the key to benchmarking the survival model.

The preceding discussion illustrates the weakness of the LFS data for a study of maternity leave, relative to SLID data. In addition to having fewer explanatory variables available than in SLID, we must accept proxies for the dependent variables. Nevertheless, we require the historical depth of the LFS. This relationship between the data sets is the context of the benchmarking problem described in the next section.

Both the SLID and the LFS have complex sample designs involving detailed stratification, and complex methods for calculating observation weights. We always make use of observation weights, both in estimation and in the calculation of frequencies. The methods used are fairly simple, and are discussed in sections 4 and 5.

2.3 The Benchmarking Problem

The context of our benchmarking problem is a model of women choosing between leaving the labour force or taking a maternity leave, and if they choose a leave, deciding how long that leave should be. The first decision is represented by a binary logit model, and the second by a semiparametric survival model, both including a vector of explanatory variables and associated parameters. In LifePaths, the decisions are made as part of the maternity leave choices event, which always occurs in the middle of a pregnancy. SLID is quite adequate for estimation of the base parameters of both these models. However, since a major goal of the LifePaths project is to incorporate historical patterns of change in socio-economic processes, it was necessary to benchmark the SLID parameter estimates to annual estimates of dependent variable means obtained from the LFS.

In this problem, we assume stable observed characteristics of the population. There are two reasons for this. First, LifePaths is a work in progress, and the benchmarking exercise we report on was carried out at a stage when other parts of the model that predict these characteristics were being extensively revised. In section 3.3, we touch on the consequences of evolving population characteristics. Second, we suppose that the primary reason for systematic change in observed outcomes between time periods is change in some factors not included in the measured characteristics of individuals. In the case of our application we observed a trend towards choice of maternity leave over leaving the labour force which seems to be due to social change rather than changes in the composition the population. We also observed a change in the distribution of maternity leave durations that appears to be due to changes in the Unemployment Insurance (UI) program implemented in Bill C-21 in 1990. At that time Parental Benefits were introduced, which extended the period during which many mothers could receive benefits from 15 to 25 weeks. Many mothers return to work at a time close to when they have exhausted UI benefits.

3. BENCHMARKING METHODOLOGY

In this section we present the method in an abstract form in order to clarify the assumptions, develop notation, and to reveal the similarity between the application to binary choice and to survival analysis.

3.1 Application to Binary Choice

The basic model for the benchmarking methodology relates to binary choice. Since we are not primarily interested in changes in the population, we simplify the analysis by assuming that the explanatory variables or individual characteristics in period τ are represented by a series of independent identically distributed random vectors X^τ. We recognise that this is quite a strong assumption. Nevertheless, for the reasons discussed in section 2.3, we use it our empirical work. Section 3.3 shows that it is a fairly simple matter to extend the theory to incorporate trends in the independent variables.

Consider a linear predictor given by

$$\eta^\tau(x) = \beta' x + \gamma^\tau \qquad (3.1)$$

where β is a vector of coefficients constant over time, x is a possible outcome of X^τ, and γ^τ represents a parameter specific to period τ. Notice that x contains no "constant term." Let Y^τ be a random variable, jointly distributed with X^τ, that takes the values 1 if an event occurs and 0 if it does not. Suppose that the probability of the event, conditional on characteristics x, is given by

$$E(Y^\tau \mid X^\tau = x) = \pi^\tau(x) = F(\eta^\tau(x)) \tag{3.2}$$

where we require F to be a continuous distribution function. The values of the function will then be bounded by zero and one, and it will have an inverse g, so that

$$\eta^\tau(x) = g(\pi^\tau(x)). \tag{3.3}$$

In the context of generalised linear models, g is called a link function. We begin by finding maximum likelihood estimates of the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using data for the time period τ_0 (in our case this is the period when SLID data are available). Of course these data must include variables corresponding to outcomes of both X^τ and Y^τ. It remains to estimate γ^τ for each period τ. Equations (3.1) and (3.3) imply that

$$E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} = \gamma^\tau - \gamma^{\tau_0} = E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\}. \tag{3.4}$$

Since we have observations only on the outcomes of Y^τ from the LFS for every period, we estimate the terms γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) \tag{3.5}$$

where $\hat{\pi}^\tau$ is an estimate of $E(Y^\tau)$. Using the LFS, this estimate is the weighted frequency of the event in the time period τ (taking each weight from the month where a child is first observed). To justify this procedure we use equation (3.4) and assume an approximation

$$E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \approx g(E\{\pi^\tau(X^\tau)\}) - g(E\{\pi^{\tau_0}(X^{\tau_0})\}). \tag{3.6}$$

Inaccuracy will arise due to Jensen's inequality in regions where g is convex or concave. Nevertheless, if g can be locally approximated by a linear function in the regions where $\pi^\tau(X^\tau)$ and $\pi^{\tau_0}(X^{\tau_0})$ are concentrated, then (3.6) may be quite accurate. The fact that g has an inflection point at 0.5 may aid the approximation when probabilities are dispersed around this value.

Fortunately we are able to test the adequacy of the estimator by simulating the estimated model in LifePaths and comparing the predicted frequencies of the event with corresponding weighted frequencies observed in the data. The results indicate that it is quite adequate for our application.

3.2 Application to Survival Analysis

We will show in section 5.2 that the approach outlined above can also be extended for use with a semiparametric survival model by adding an index t representing the duration in the current state, so that (3.5) becomes

$$\hat{\gamma}^\tau(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) \tag{3.7}$$

where $\hat{\pi}^\tau(t)$ represents the empirical hazard function.

3.3 Trends in the Independent Variables

The benchmarking method may be improved by taking the changes in observed characteristics into account. As we noted in section 2.3, this would be considered when other parts of LifePaths are in a more mature form. To do this we relax the assumption that the random vectors X^τ are identically distributed. Equation (3.4) then becomes

$$E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} = \gamma^\tau - \gamma^{\tau_0} + \beta'\{E(X^\tau) - E(X^{\tau_0})\} = E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \tag{3.8}$$

Based on this, we might estimate γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) - \hat{\beta}'(\bar{x}^\tau - \bar{x}^{\tau_0}) \tag{3.9}$$

where $\bar{x}^\tau$ is the vector of mean values of the characteristics in period τ. Of course it may not be possible to obtain all of the mean values from the same data source. The method would extend to the survival model case in the same manner as (3.7) to give

$$\hat{\gamma}^\tau(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) - \hat{\beta}'(\bar{x}^\tau(t) - (\bar{x}^{\tau_0}(t)). \tag{3.10}$$

4. MODELS AND THE ESTIMATION OF BASE PARAMETERS

As explained in section 3.1, the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ are estimated by maximum likelihood using data from the period τ_0. We use data from SLID on all maternity leaves beginning in the period 1993-1996 (our base period τ_0). We do not attempt to estimate annual changes in the constant term γ throughout this period.

4.1 The Binary Logit Model

We adopt the logit model to represent a mother's choice between taking a maternity leave and withdrawing from the labour force. From now on we adopt a more conventional econometrics notation and use a subscript i to index a random variable or outcome associated with an individual i. We suppose that a random variable Y_i^τ takes values 0 or 1, with $Y_i^\tau = 1$ indicating that new mother i with vector of characteristics x_i in period τ chooses to take a maternity leave, conditional on her having been employed, and that

$$\pi_i^\tau = P(Y_i^\tau = 1) = F(\eta_i^\tau) = \frac{\exp(\eta_i^\tau)}{1 + \exp(\eta_i^\tau)}. \qquad (4.1)$$

where $\eta_i^\tau = \beta' x_i + \gamma^\tau$ is the linear predictor of equation (3.1) and F is the logistic distribution function. We estimate the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\beta, \gamma^{\tau_0})$ where

$$\begin{aligned} L(\beta, \gamma^\tau) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \\ &= \prod_{y_i = 0} [1 - F(\eta_i^\tau)] \prod_{y_i = 1} F(\eta_i^\tau) \\ &= \prod_i [F(\eta_i^\tau)]^{y_i} [1 - F(\eta_i^\tau)]^{1 - y_i} \qquad (4.2) \end{aligned}$$

and

$$\ln L(\beta, \gamma^\tau) = \sum_i \{ y_i \ln F(\eta_i^\tau) + (1 - y_i) \ln [1 - F(\eta_i^\tau)] \}. \qquad (4.3)$$

Longitudinal SLID weights in the year of the child's birth are scaled to sum to the sample size, and are then used to weight the terms of the log-likelihood and its derivatives. The weighted score equations are

$$\begin{aligned} \frac{\partial L(\beta, \gamma^\tau)}{\partial \beta} &= \sum_i w_i x_i y_i - \sum_i w_i x_i F(\eta_i^\tau) = 0 \\ \frac{\partial L(\beta, \gamma^\tau)}{\partial \gamma^\tau} &= \sum_i w_i y_i - \sum_i w_i F(\eta_i^\tau) = 0. \qquad (4.4) \end{aligned}$$

The solution, which maximises the log-likelihood, was found by Newton-Raphson iteration. The logit model has been used often by statisticians and econometricians, and there is an extensive literature. For example, see Chambless and Boyle (1985), Roberts, Rao, and Kumar (1987), and Morel (1989).

4.2 The Semiparametric Survival Model: Basic Form

For mothers who have chosen to take a maternity leave from their job, we use a survival model to describe the duration of their leave. The probability density function (pdf) of the distribution has a complex shape, as can be seen from the graphs in section 6.4. There is spike at durations of less than a month and a mode which appears to represent the maximum Unemployment Insurance special benefits entitlement available to mothers after 1990 (15 weeks of Maternity Benefits, plus 10 weeks of Parental Benefits, plus a two-week waiting period). We began the study by estimating various fully parametric models, including a log-logistic survival model combined with a logit model to predict durations of less than a month, but were unable to obtain an adequate fit. To solve this problem, we follow Prentice and Gloeckler (1978), Han and Hausman (1986) and Meyer (1990), by nonparametrically estimating the effect of time on the hazard of returning to work. The hazard of returning to work is specified in a proportional hazards form:

$$\lambda_i^\tau(t) = \lambda_0^\tau(t) \exp \{\beta' x_i(t)\} \qquad (4.5)$$

where $\lambda_0^\tau(t)$ is the unknown baseline hazard at leave duration t and time period τ, $x_i(t)$ is a vector of explanatory variables for mother i, and β is a vector of coefficients. The data tell us which of the intervals [0,1), [1,2), [2,3), ... contains the spell duration (in our case the units are months), and the model can be interpreted as an incompletely observed continuous time hazard model with no restriction on the form of the baseline hazard. If T_i^τ is the duration of leave for mother i during period τ, then for $t = 1, 2, 3, \ldots$, the probability that the spell lasts until time t, given that it has lasted until $t - 1$, can be written as

$$\begin{aligned} P(T_i^\tau > t \mid T_i^\tau \geq t - 1) &= \exp\left[-\int_{t-1}^{t} \lambda_i^\tau(u)\, du \right] \\ &= \exp\left[-\exp\{\beta' x_i(t)\} \int_{t-1}^{t} \lambda_0^\tau(u)\, du \right] \qquad (4.6) \end{aligned}$$

if we assume that $x_i(t)$ is constant on the interval between $t - 1$ and t. In order to apply the theory of section 3, we can rewrite equation (4.6) as

$$\begin{aligned} 1 - \pi_i^\tau(t) &= P(T_i^\tau \geq t \mid T_i^\tau \geq t - 1) \\ &= \exp[-\exp\{ \beta' x_i(t) + \gamma^\tau(t)\}] \\ &= \exp[-\exp\{\eta_i^\tau(t)\}] \qquad (4.7) \end{aligned}$$

where

$$\gamma^\tau(t) = \ln \left[\int_{t-1}^{t} \lambda_0^\tau(u)\, du \right]. \qquad (4.8)$$

One may censor any ongoing observations at some large duration T. Again we can estimate the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\gamma^{\tau_0}, \beta)$. Since we will always be referring to data from the base period for the remainder of section 4, we drop superscripts τ_0.

The likelihood function is given by

$$L(\gamma, \beta) = \prod_{i=1}^{N} [[1 - \exp\{ -\exp(\eta_i(k_i))\}]^{\delta_i} \prod_{t=1}^{k_i} \exp\{ -\exp(\eta_i(t))\}] \qquad (4.9)$$

where $\gamma = [\gamma(1), \gamma(2), \ldots, \gamma(T)]'$, C_i is a censoring time, $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise, $k_i = \min(\text{int}(T_i), C_i)$. The log-likelihood is therefore

$$\ln L(\gamma, \beta) = \sum_{i=1}^{N} [\delta_i \ln[1 - \exp\{-\exp(\eta_i(k_i))\}] - \sum_{t=1}^{k_i} \exp(\eta_i(t))]. \quad (4.10)$$

Weights from the months that a child is first observed are scaled to sum to the sample size, and then used to weight the terms of the log-likelihood function and its derivatives. The weighted log-likelihood function is maximised by the quasi-Newton algorithm of Broyden, Fletcher, Goldfarb, and Shanno (BFGS), using an implementation based on Dennis and Schnabel (1983).

4.3 The Semiparametric Survival Model: with Work-to-Birth Gap Decision

The situation in our application is complicated somewhat by our desire to model the duration from leaving the job until the birth (the work-to-birth gap), as well as the hazard of returning to work from a maternity leave. The model of work-to-birth gap is estimated separately, based on SLID data. Examination of the mean gap duration for each year in the LFS data indicates that this duration has been fairly stable over time, so the model is not benchmarked. Nevertheless, a modification of the semiparametric survival model is necessary to incorporate the separate model of work-to-birth gap. This can be accomplished by assuming that the work-to-birth gap decision, possibly involving health considerations, acts to constrain the desired total duration. This means that the above model would apply to the desired total duration, which is unobservable, and might be labelled T^*.

In cases where the desired duration was shorter than the work-to-birth gap, the mother might return to work as soon as possible after the birth. This means that in cases where we observe a significant work-to-birth gap (greater than a month), and the mother returns soon after birth (within a month), all that is known about desired duration is that

$$T^* \leq T$$

where T is the total duration of leave. This is equivalent to a situation labelled "left censoring" by Cox and Oaks (1984, page 178), where observation does not start immediately and some individuals have already failed before it does.

From such an observation we get a contribution to the likelihood function and its logarithm given by

$$L_i = 1 - \prod_{t=1}^{k_i} P(T^* \geq t \mid T^* \geq t-1) = 1 - \prod_{t=1}^{k_i} \exp[-\exp(\eta_i(t))] \quad (4.11)$$

and

$$\ln(L_i) = \ln\{1 - \exp[-\sum_{t=1}^{k_i} \exp(\eta_i(t))]\}. \quad (4.12)$$

Unfortunately the log-likelihood expression does not simplify like the corresponding expression for "right-censored" observations. In spite of this, Monte Carlo experiments indicate that estimation is not a problem even in heavily censored data sets.

Longitudinal SLID weights in year of the child's birth are used in same manner as for the basic form of the survival model.

5. BENCHMARKING THE MODELS

To begin the benchmarking procedure we must invert the distribution function F given in equation (3.2) to find the link function g. We then apply equation (3.5) in the case of the logit model, and equation (3.7) in the case of the survival model.

5.1 Application to the Binary Logit Model

To benchmark the logit model we first invert the logistic distribution function in equation (4.1) to obtain

$$\eta_i^\tau = g(\pi_i^\tau) = \ln\left(\frac{\pi_i^\tau}{1 - \pi_i^\tau}\right) \quad (5.1)$$

where g is the well-known logit function. We can then apply equation (3.5) and (5.1) to obtain

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) = \hat{\gamma}^{\tau_0} + \ln\left(\frac{\hat{\pi}^\tau/(1-\hat{\pi}^\tau)}{\hat{\pi}^{\tau_0}/(1-\hat{\pi}^{\tau_0})}\right) \quad (5.2)$$

where for $\tau < \tau_0$, each $\hat{\pi}^\tau$ is the frequency of choosing maternity leave calculated from LFS data for maternity leaves beginning in year τ, and $\hat{\pi}^{\tau_0}$ is the frequency from SLID data.

5.2 Extension to the Survival Model

From equation (4.7) we get

$$\pi_i^\tau(t) = 1 - \exp[-\exp\{\eta_i^\tau(t)\}] = F\{\eta_i^\tau(t)\} \quad (5.3)$$

where

$$\eta_i^\tau(t) = \beta' x_i(t) + \gamma^\tau(t). \quad (5.4)$$

In this case F is an extreme value distribution that is easily inverted to obtain

$$\eta_i^\tau(t) = \ln[-\ln(1 - \pi_i^\tau(t))] = g(\pi_i^\tau(t)). \quad (5.5)$$

For benchmarking we can use equation (3.7) with the observed frequencies in period τ represented by the empirical hazard or occurrence/exposure ratio given by

$$\hat{\pi}^{\tau}(t) = d^{\tau}(t) \Big/ r^{\tau}(t) \qquad (5.6)$$

where, for spells beginning in period τ, $d^{\tau}(t)$ is the number of mothers who fail in the interval $(t-1, t]$ and $r^{\tau}(t)$ is the number of mothers in view at duration t, including those censored at time t (censoring can only occur at the end of intervals). Numbers of mothers were calculated from sample counts by applying the LFS weight from the month that a new mother returns to work. The empirical hazard and the corresponding estimator for the survivor function implied by the product law of probabilities were studied by Kaplan and Meier (1958). The use of the empirical hazard in equation (3.7) together with equation (5.5) yields

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + \ln\left(\frac{\ln[1 - \hat{\pi}^{\tau}(t)]}{\ln[1 - \hat{\pi}^{\tau_0}(t)]}\right). \qquad (5.7)$$

6. EMPIRICAL RESULTS

The results of estimation in the base period, and the results of simulation with benchmarked parameter estimates are presented for both models. The simulation results are compared with annual survey sample frequencies of choosing a maternity leave in the case of the logit model, and with annual survey frequency distributions of maternity leave duration in the case of the survival model.

6.1 Estimation Results for the Binary Logit Model

The estimation results obtained from estimating the logit model from SLID data are presented in Table 1. Omitted dummy variable categories, which form the reference categories for the variables used in the model, were province of residence Ontario and highest education level "some post secondary." Individual and family income variables were tested, but were found not to be significant, and so were not included in the regression.

There may be some bias in the estimates, particularly those of the standard errors, due to the fact that the complex SLID sample design was accounted for only through the weights applied to the log-likelihood.

The significant positive effect of job tenure seems reasonable for a number of reasons. A lengthy tenure might indicate that the woman has acquired firm-specific human capital and has achieved some seniority. It would also be an indicator of strong attachment to the labour force generally. On the firm side, the longer the woman's job tenure, the longer the leave that the firm is likely to grant with a guarantee that she can return to her job. Also, provincial government guarantees of job security also depend on job tenure. Finally, a lengthy job tenure means that the woman will likely meet the Unemployment Insurance eligibility requirements (20 weeks of insured employment). A dummy variable indicating that UI entrance requirements were met was tested and found to be just significant at the 5% level. However, because we are not able at this stage to model changes in the UI program through the influence of covariates, because of uncertainty in interpretation, and because of high correlation with job tenure, it was not included. In the LFS, self-employed workers are reported as having a transition out of employment only when they terminate their business. Since taking a leave simply means not terminating the business, a significant positive effect for the indicator of self-employment is to be expected. Having been self-employed before the birth increases the odds of taking a maternity leave by 333%, the strongest effect that we see for an indicator variable.

Table 1
Binary Logit Parameter Estimation Results

Parameter	Estimate of Coefficient	Contribution to Odds Ratio*	Std Error of Coefficient	Prob-Value
Constant	-6.432	0.002	2.995	0.0318
NFLD	-0.829	0.436	0.741	0.2636
PEI	0.931	2.537	1.612	0.5633
NS	-0.456	0.634	0.541	0.3992
NB	0.207	1.230	0.675	0.7596
QUE	-0.361	0.697	0.247	0.1437
MAN	-0.490	0.613	0.503	0.3306
SASK	-0.163	0.850	0.458	0.7218
ALTA	-0.200	0.819	0.325	0.5379
BC	-0.120	0.887	0.300	0.6899
Job Tenure (mths)/10	0.094	1.099	0.026	0.0003
Self-employed?	1.203	3.330	0.418	0.0040
Age (Years)	0.479	1.614	0.199	0.0160
(Age^2)/10	-0.071	0.931	0.033	0.0296
< High School Grad	-0.702	0.496	0.357	0.0490
High School Grad	-0.148	0.862	0.276	0.5913
University Grad	-0.292	0.747	0.229	0.2027
First Child?	-0.525	0.592	0.192	0.0063

log-likelihood = -381.553
Number of Observations = 835
Observations are given the SLID longitudinal weight from the year of birth, scaled to sum to the sample size

* This is the exponential of the coefficient. It may be interpreted as the proportional change in the odds ratio due to a unit change in the corresponding independent variable.

The effect of the first child indicator also seems reasonable. The odds for maternity leave for a first-time mother is only 59% of the odds for maternity leave for a mother of more than one child, given that all other characteristics are the same – *i.e.* first-time mothers are more inclined to job separation than the mothers who already have children. This may be partly a consequence of the fact that our sample consists of mothers who have been employed within 4 months of the birth. Mothers who have more than one child tend to space them within a few years at most. If they are employed just before a second or subsequent births, they will have already demonstrated that they returned to work after an absence that must have been less than the gap between births. This at least rules out some common patterns of withdraw from the labour force – for example staying at home until all children are in school.

The effect of age is more difficult to interpret since the effect on the log-odds ratio is non-linear. By drawing a graph of the term -.479* *age* - .0071* age^2 one can see that, as age increases, the log-odds of taking a maternity leave first increases, but that the rate of increase declines until a level point at the maximum log-odds is reached by the age of 34. Since the number of mothers declines considerably after this age, the subsequent decline may not be meaningful. One might hazard a conjecture that, among young mothers, being relatively older indicates more attachment to the labour force and thus a stronger tendency to take a maternity leave, while among older mothers, who are past the stage of first entering the labour force, this effect is reduced. However, the results are probably not precise enough to draw any firm conclusion about this.

6.2 Simulation Results for the Benchmarked Binary Logit Model

The benchmarking exercise consists of adjusting the constant term of the model in the manner described by (5.2) for each year in the period 1975-1992. The constant term is not adjusted after 1992, partly because the LFS data do not indicate a strong trend after 1992. The model is then incorporated in LifePaths and a simulation is run. For each year from 1976 to 1995, Figure 1 shows both the frequency of choosing a leave in the LifePaths simulation, and the frequency estimated from the LFS. For the period 1993-1995, estimates from SLID are also presented.

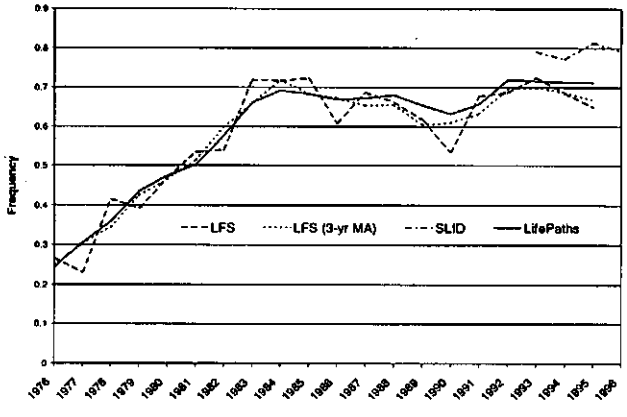


Figure 1. Frequency of Choosing a Maternity Leave 1976-1996

The simulation captures the change over time revealed by the LFS data during the period 1976-1992. There is no benchmark adjustment implemented in the LifePaths simulation after 1992, so that the base parameters estimated from pooled SLID data 1993-1996 are effective. The simulated frequency is slightly lower than the observed SLID frequency during this period. Two possible sources of error are an insufficiently flexible specification of the binary choice model, and differences between the SLID estimates of explanatory variables and those provided by LifePaths.

6.3 Estimation Results for the Survival Model

The results obtained from estimating the semiparametric survival model from SLID data are presented in Table 2. As in the binary logit model estimation, omitted dummy variable categories were province of residence Ontario and highest education level "some post secondary." Since the dependent variable is the hazard of returning to work, a positive coefficient for a covariate indicates an influence that tends to shorten the duration of maternity leave.

The estimates of the constant terms in the duration-dependent linear predictor given by (4.7) are denoted in Table 2 by GAMMA*i*, $i = 1, 2, \ldots, 15$. This represents the influence of the baseline hazard incorporating the influence of duration.

Table 2
Survival Model Parameter Estimation Results

Parameter	Estimate	Std Error	Prob-Value
Job Tenure (mths) /10	-0.030	0.010	0.0024
NFLD	0.195	0.426	0.6470
PEI	0.307	0.490	0.5313
NS	0.173	0.253	0.4940
NB	0.109	0.293	0.7091
QUE	0.111	0.117	0.3411
MAN	-0.402	0.253	0.1116
SASK	-0.303	0.213	0.1539
ALTA	0.270	0.154	0.0798
BC	-0.440	0.148	0.0030
Self-Employed?	1.665	0.157	0.0000
Age	-0.253	0.041	0.0000
Age** 2 / 10	0.043	0.007	0.0000
First Child?	-0.301	0.090	0.0009
< High School Grad	0.508	0.206	0.0135
High School Grad	-0.124	0.125	0.3212
University Grad	-0.374	0.108	0.0006
Employed Spouse?	0.109	0.151	0.4703
Gamma1	2.570	0.609	0.0000
Gamma2	-1.136	0.816	0.1636
Gamma3	-0.466	0.719	0.5176
Gamma4	0.780	0.640	0.2232
Gamme5	1.425	0.627	0.0231
Gamma6	2.755	0.613	0.0000
Gamma7	3.640	0.612	0.0000
Gamma8	3.413	0.620	0.0000
Gamma9	3.465	0.630	0.0000
Gamma10	3.387	0.649	0.0000
Gamma11	4.579	0.655	0.0000
Gamma12	4.285	0.785	0.0000
Gamma13	3.645	1.110	0.0010
Gamma14	3.746	1.281	0.0034
Gamma15	6.215	2.415	0.0101
log-likelihood = -1165.06			
Number of Observations 3411			
Observations are given the SLID longitudinal weight from the year of birth, scale to sum to the sample size			

Again, individual and family income variables were tested and found not to be significant. Both this finding and the importance of a self-employment indicator as a predictor of early return to work accord with the findings of Marshall (1999). Marshall found that education variables were not significant in determining whether a mother would return to work within a month. We find however, that university graduation has a significant negative effect on the hazard (positive effect on duration). Job tenure has a significant negative effect on the hazard, possibly reflecting its relationship with Unemployment Insurance entitlement and job security.

6.4 Simulation Results for the Benchmarked Survival Model

In the case of the semiparametric survival model, benchmarking consists of adjusting all of the terms GAMMAi, $i = 1, 2, \ldots, 15$ of the previous section according to (5.8) for each of the years in the period 1975-1992. The model is then simulated as part of LifePaths.

The frequency distribution of simulated maternity leave durations is presented and compared to the corresponding observed frequency distribution from LFS data. In order to present the results, the frequencies in 3-year periods were averaged. A key feature of the frequency distribution is an abrupt change apparently due to the introduction of parental benefits with Bill C-21 at the end of 1990. Since mothers with maternity claims in progress at the time of implementation were entitled to parental benefits, the claims beginning in 1990 represent a mixture of regimes. For this reason the year 1990 is not included in any of the 3-year averages. In Figures 2 and 3 we use disjoint 3-year periods covering 1976-1984. To balance periods before and after 1990 using available data, in Figures 4 and 5 we use the overlapping periods 1985-1987, 1987-1989, 1991-1993, and 1993-1995.

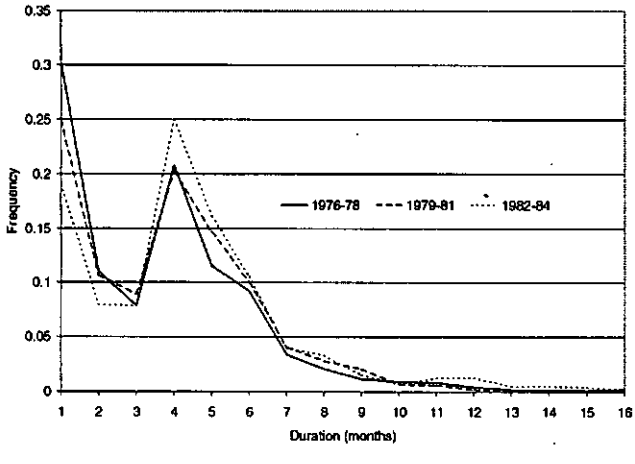


Figure 2. LifePaths: Distribution of Leave Durations for 1976-1984

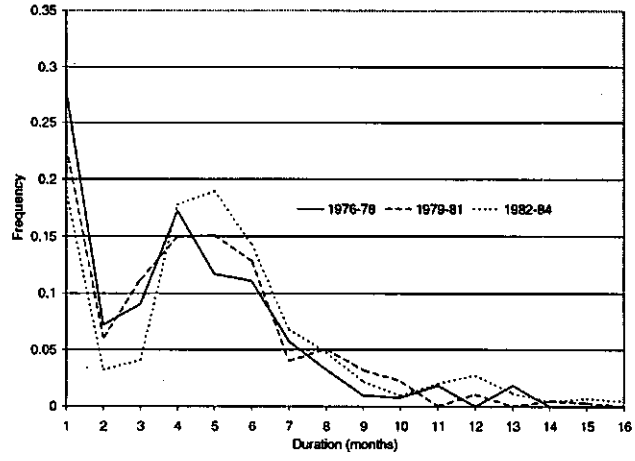


Figure 3. LFS Data: Distribution of Leave Durations for 1976-1984

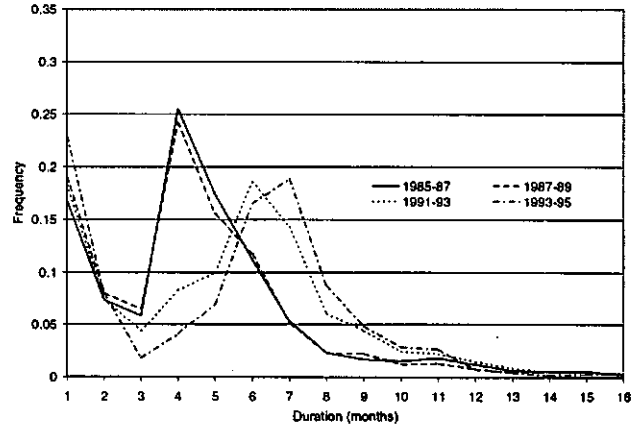


Figure 4. LifePaths: Distribution of Leave Durations for 1985-1989 and 1991-1995

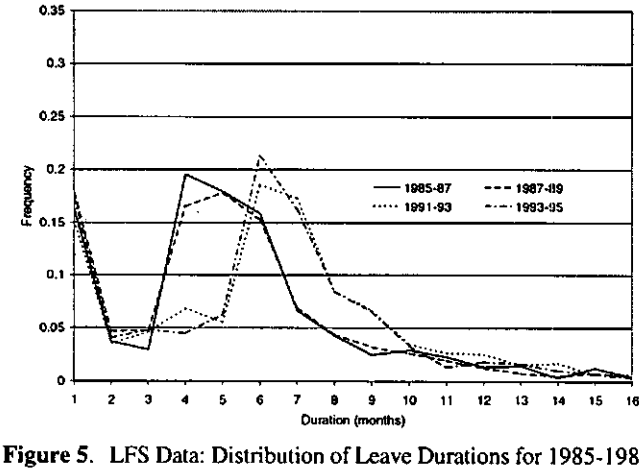


Figure 5. LFS Data: Distribution of Leave Durations for 1985-1989 and 1991-1995

The distribution of durations derived from SLID data 1993-1996 is presented in Figure 6. This may be compared with the simulated data shown in Figure 4 for the period 1993-1995, since no benchmarking is applied after 1992.

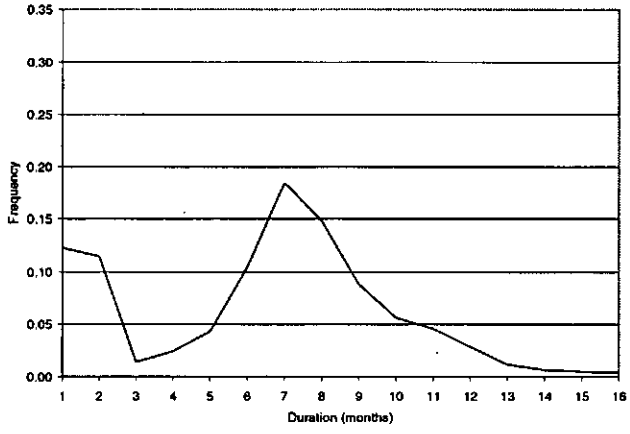


Figure 6. SLID Data: Distribution of Leave Durations for 1993-1996

In Figure 7 we present the average duration of maternity leaves beginning in each year of the observed period. The average of simulated durations are compared with those from the surveys.

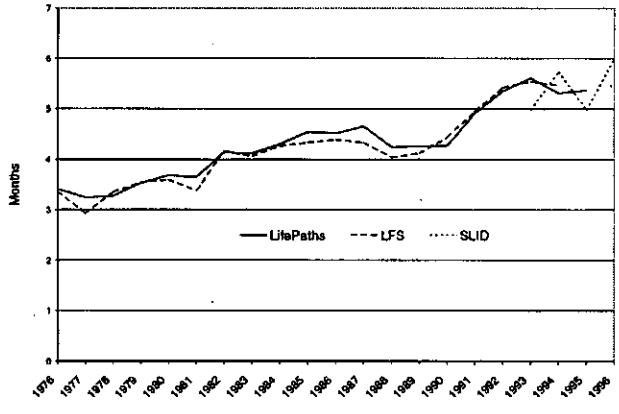


Figure 7. Average Duration of Maternity Leave 1976-1996

6.5 Evaluation of Benchmarking Performance

The benchmarking method appears to be very effective in the case of the binary logit model. The trend of the LFS data is well reflected in the LifePaths simulation. In the case of the survival model, the key feature of the LFS data is the abrupt shift of the mode of the frequency distribution after 1990, apparently due to the introduction of parental benefits. This shift has been captured by the simulated data. Also the average duration of maternity leave in the simulation fits the LFS data very closely.

A noticeable divergence between the simulation and the LFS data is the height of the mode at the interval (3, 4] months in the frequency distribution of the durations from LifePaths from 1982-1989. This may be due to the effect of trends in the values of explanatory variables, which we have assumed to be stable. Further work is necessary to establish this. A possible extension to the model was discussed in section 3.3.

7. CONCLUSIONS

The technique that we have developed appears to be quite successful in benchmarking of the logit and survival model parameters so that the essential features of the LFS data are captured in LifePaths predictions. The key to benchmarking the logit model is the adjustment of the parameter corresponding to the "constant term" in the linear predictor that is imbedded it the logistic distribution function in order to predict the conditional expectation of the dependent variable. Section 3.1 develops the technique in a general framework that includes other models of binary choice. Particularly, it would extend to the popular probit model where a linear predictor is embedded in the standard normal distribution function. Benchmarking of the semi-parametric survival model hinges on the adjustment of all the parameters representing the baseline hazard. Our results illustrate how the entire shape of the distribution of durations predicted by the model can be made to evolve through time according to a pattern revealed by supplementary data.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to Steve Gribble and members of the Socio-economic Modelling Group at Statistics Canada for useful comments throughout the development of the maternity leave module, to Geoff Rowe and Huan Nguyen for use of their computer program to follow individuals through rotations in the LFS, to Katherine Marshall for advice on the use of SLID and for sharing computer programs, to Adrienne ten Cate for fruitful discussions, and to an anonymous referee for several improvements. This work was performed when both authors worked in the Socio-Economic Modeling Group, Statistics Canada, R.H. Coats Building 24th floor, Tunney's Pasture

REFERENCES

APPLEBY, J., BOOTHBY, D., ROULEAU, M. and ROWE, G. (1999). Level and Distribution of Individual Returns to Post-Secondary Education: Simulation Results from the LifePaths Model. Presented at the 1999 meetings of the Canadian Economics Association.

CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum Likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, A: Theory and Methods*, 14, 177-192.

CHEN, E.J., and ODERKIRK, J. (1997). Varied Pathways: The Undergraduate Experience in Ontario, Feature article. *Education Quarterly Review*, Statistics Canada, 4, 3, 47-62.

CHEN, E.J., and ROWE, G. (1999). Trend Correlation of Labour Market Earnings in Canada: 1982 to 1995. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 173-179.

COX, D.R., and OAKS, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

DENNIS, J.E. Jr, and SCHNABEL, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

HAN, A., and HAUSMAN, J. A. (1986). Semiparametric Estimation of Duration and Competing Risk Models. M.I.T. Working Paper No. 450.

KAPLAN, E.L., and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-81.

MARSHALL, K. (1999). Employment after childbirth. *Perspectives on labour and income*. Statistics Canada, Autumn 1999, 18-25.

MEYER, B.D. (1990). Unemployment Insurance and Unemployment Spells. *Econometrica*, 58, 757-782.

MOREL, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology* 15, 205-223.

PLAGER, L., and CHEN, E.J. (1999). Student Debt from 1990-91 to 1995-96: An Analysis of Canada Student Loans Data. MAJOR RELEASES, *THE DAILY* and *Education Quarterly Review*, Statistics Canada, 5, 4, 10-35.

PRENTICE, R., and GLOECKLER, L. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data, *Biometrics*, 34, 57-67.

ROBERTS, G.A., RAO, J.N.K. and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

ROWE, G., and CHEN, E.J. (1998). An Increment-Decrement Model of Secondary School Progression for Canadian Provinces. *Proceedings: Symposium on Longitudinal Analysis for Complex Surveys*, Statistics Canada, 167-178.

ROWE, G., and LIN, X. (1999). Modelling Labour Force Careers for the LifePaths Simulation Model, Proceedings: *Symposium 99 Combining Data from Different Sources*, Statistics Canada, 57-64.

WOLFSON, M.C. (1997). Sketching LifePaths: A New Framework for Socio-Economic Statistics. *Simulating Social Phenomena*, (Eds. Conte, R. Gegselmann and P. Terna), Lecture Notes in Economics and Mathematical Systems, 456, Springer.

WOLFSON, M.C., and ROWE, G. (1996). Perspectives on Working Time Over the Life Cycle, Canadian Employment Research Forum Conference on Changes to Working Time, Ottawa.

WOLFSON, M.C., and ROWE, G. (1998a). LifePaths - Toward an Integrated Microanalytic Framework for Socio-Economic Statistics. 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

WOLFSON, M.C., and ROWE, G. (1998b). Public Pension Reforms - Analyses Based on the LifePaths Generational Accounting Framework, 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

WOLFSON, M.C., ROWE, G., GRIBBLE, S. and LIN, X. (1998). Historical Generational Accounting with Heterogeneous Populations. *Government Finances and Generational Equity* (Ed. M. Corak), Statistics Canada, 107-127.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 63-76
Statistics Canada

Improved Ratio Estimation in Telephone Surveys Adjusting for Noncoverage

STEVEN T. GARREN and TED C. CHANG[1]

ABSTRACT

Since some individuals in a population may lack phones, telephone surveys using random digit dialing within strata may result in asymptotically biased estimators of ratios. The impact from not being able to sample the nonphone population is examined. We take into account the propensity that a household owns a phone, when proposing a post-stratified phone-weighted estimator, which seems to perform better than the typical post-stratified estimator in terms of mean squared error. Such coverage propensities are estimated using the Public Use Microdata Samples, as provided by the United States Census. Non-post-stratified estimators are considered when sample sizes are small. The asymptotic mean squared error, along with its estimate based on a sample, of each of the estimators is derived. Real examples are analyzed using the Public Use Microdata Samples. Other forms of nonresponse are not examined herein.

KEY WORDS: Asymptotics; Census Public Use Microdata Samples; Post-stratification; Telephone survey.

1. INTRODUCTION

Consider surveys where the telephone population is sampled. Major problems in telephone surveys include nonresponse (*i.e.*, refusal to participate in the survey) and noncoverage (*i.e.*, lacking telephone service). Nonresponse may cause larger bias than noncoverage, since nonresponse propensities are usually much higher than noncoverage propensities. However, nonresponse is reviewed rather briefly, because the focus of this article is noncoverage.

1.1 Literature Review

Khurshid and Sahai (1995) provided an extensive bibliography of papers on telephone surveys. Examples of nonresponse rates may be found in Steeh, Groves, Comment and Hansmire (1983, pages 189-197). Corrections for nonresponse, using weights and imputation, were discussed by Little (1986) and Rubin (1987). Rao (1997) provided an overview of sample surveys, including discussions on resampling methods, especially the jackknife, for variance estimation. His discussion includes techniques to estimate the variance in the presence of imputation.

Regarding noncoverage, Brick, Waksberg and Keeter (1994) found the 94% of the households in the United States have phones at any given time. They also found that the households with interrupted telephone service usually are indigent. Keeter (1995) discussed that in a survey conducted from 1992 to 1993 more than half of all households without continuous telephone service during that year were *transient*, *i.e.*, these transient households were both with and without telephone service at different times during that year. He also found that most socioeconomic factors (excluding home ownership) for transient telephone households are similar to those factors for households which are continuously without phones. These similarities between the transient and the nonphone populations suggest that valid inferences may be made on the entire (phone, nonphone, and transient) population, based on telephone surveys. Thornberry and Massey (1988) examined noncoverage for various socio-demographic groups from 1963 to 1986, and found income to be the most important factor in determining the likelihood that a household has a phone.

1.2 Our Approach

Given several various characteristics, such as home ownership and household language, the propensity of a household to have phone service is estimated in this article using the Virginia portion of the 1990 Census Public Use Microdata Samples (PUMS), which represent 5% of the population. Whether or not households have phones is included in the PUMS. The estimation of these propensities, or probabilities of phone service, is based on generalized linear regression with a log – log link, since the logit link provides a poor fit. We advocate using our fitted regression model, with the estimated parameters, for estimating these likelihoods in general whenever a random sample is taken from the Virginian phone population.

Because it is such a huge data set, the PUMS have another useful purpose in this article. The PUMS are used to compare and contrast estimators in terms of bias and variance, by examining the entire phone population and by taking repeated samples of the phone population. Categorical data consisting of 75 household and 75 personal variables are listed for all individuals in the households selected to be in the PUMS.

[1] Steven T. Garren, Department of Mathematics and Statistics, MSC 7803, James Madison University, Harrisonburg, Virginia, 22807, U.S.A. Research partially supported by NIMH grant MH53259-01A2; Ted C. Chang, Division of Statistics, 108 Halsey Hall, University of Virginia, Charlottesville, Virginia, 22903, U.S.A. Research partially supported by ONR grant N-00014-92-J-1009.

In the examples in section 6 high school graduation rate, mean number of cars per household, and mean household income are estimated using both post-stratified and non-post-stratified estimators for samples of size 500 from the PUMS. The post-stratification variables for high school graduation rate are gender, age, and race of the head of household. The post-stratification variable for mean number of cars per household is household income only. Estimators of the mean household income are analysed twice. For one analysis, post-stratification is on only the race of the head of household. For the other analysis, post-stratification is on gender, age, and race of the head of household. Each of these post-stratification variables is divided into two categories, except income, which is divided among three categories.

A serious drawback to estimators not taking into account the propensities of phone service is that these estimators are not asymptotically unbiased as the sample size gets large. A major focus of this article is to show that bias is reduced substantially when the estimators take into account the propensities of phone service, as estimated by the PUMS. Since both *post-stratified* and *non-post-stratified* estimators as well as both *using* and *not using* the propensities of phone service are considered, then four estimators are examined herein. In particular, these four estimators of a population mean are the sample mean, the usual post-stratified estimator, a phone-weighted estimator, and a proposed post-stratified phone-weighted estimator. The mean squared errors (MSE) of the phone-weighted estimator and the post-stratified phone-weighted estimator go to zero as the sample size gets large, unlike the other two estimators.

We adopt a two-phase model for our four estimators. The first phase involves selection from the entire population into the phone population. We treat the propensity of a household to have phone service as the probability that the household will be selected into the phone population, and we assume that this probability is positive (although possibly small) for each household. The second phase is a stratified (perhaps geographically stratified) simple random sample from the phone population. In the examples in section 6, we consider post-stratification by characteristics such as race and age of the head of household. Since our sample sizes are small, we do not geographically stratify the population of Virginia, although our formulas allow for both stratification and post-stratification.

Ideally, one would post-stratify using the same covariates used for estimating the propensities of phone service in the first phase of our model. In this case, the three estimators which use the propensities of phone service and/or post-stratification will be almost identical. However, the sample size for each post-stratified category should not be too small, so practical limitations restrict the number of categories which should be used for post-stratification. Nevertheless, many categories may be used for constructing the propensities of phone service from the PUMS, because the entire population is used.

Even if post-stratification by many covariates is feasible, the usual variance formulas for post-stratification require that a stratified random sample be taken from the entire population. In our situation, however, a stratified random sample is taken from the phone population, so the usual variance formulas are not applicable to our situation. The techniques by Politz and Simmons (*cf.* Cochran 1977, pages 374-377) require the sampling frame to be the entire population, not just the phone population, and hence are not applicable to our scenario, which allows noncoverage.

We derive the asymptotic variances of the four estimators of a population ratio, and we determine reasonable estimates of these variances. Since a population mean is a special case of a ratio, and a population total is a multiple of a ratio, then the results regarding estimators of means or totals follow from the results regarding estimators of ratios.

2. NOTATION

Consider N households in a population, U. For each household in U, let two variables of interest be denoted by y_{1k} and y_{2k}, for $k \in U$. At any given time, the event that the kth household does or does not have a phone is treated as random, while y_{ik} is treated as fixed.

Letting

$$\alpha_i = N^{-1} \sum_{k \in U} y_{ik},$$

for $i = 1, 2$, the goal is to estimate α_1, α_2, and the ratio

$$\mu = \alpha_1 / \alpha_2.$$

Without loss of generality we concentrate on estimating α_1 and μ.

An important special case of estimating a ratio μ arises when one desires to estimate the mean of a variable z_k for $k \in D$ for some subpopulation $D \subset U$ but one cannot sample directly from D. Examples include subpopulations defined by race. Let x_k be 1 if $k \in D$ and 0 otherwise. Let $y_{1k} = z_k x_k$ and $y_{2k} = x_k$. Then μ is the population mean of z_k over the subpopulation D.

Assume there are H strata, and h is used to index the strata. Assume there are G groups, and g is used to index the groups, which are used to construct the post-strata. The strata are known prior to sampling, but the groups are not observed until after the final sample is taken. Therefore, U_{gh} denotes all households in group g and stratum h; N_{gh} denotes the size of U_{gh}, and N_h denotes the size of U_h. Other variables are defined similarly in terms of g and h.

Let U_T denote the population of households in U which currently have telephones, and let N_T denote the size of U_T. The probability, or propensity, that the kth household in U is also in U_T is denoted by p_k, and we assume that $p_k > 0$ for all k. A simple random sample of size n_h is taken from U_{Th} for $h = 1, \ldots, H$. Let s_h denote this final sample in stratum h. The size of the final sample, s, is denoted by

n. For asymptotics herein, we assume that $n/N \to 0$ as $n \to \infty$ in the same spirit as Särndal, Swensson and Wretman (1992, pages 166-169).

3. THE ESTIMATORS

The sampling design is treated as a two-phase design with Poisson sampling at the first phase and stratified simple random sampling at the second phase. Each individual enters the telephone population with probability p_k, for $k \in U$, and then enters the final sample according to a simple random sample of size n_h, $h = 1, ..., H$. The p_k are assumed known or can be estimated accurately, as shown in section 5. The estimators of μ discussed in this section will be validated in the appendix.

3.1 The Post-stratified and Ratio Estimators

Post-stratified estimates of α_1 and α_2 are

$$\hat{\alpha}_{ps(i)} = N^{-1} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} n_{gh}^{-1} \sum_{k \in s_{gh}} y_{ik},$$

for $i = 1, 2$, and the post-stratified estimate of μ is $\hat{\mu}_{ps} = \hat{\alpha}_{ps(1)} / \hat{\alpha}_{ps(2)}$. A valid estimate of the variance, conditional on U_T, is known to be (*cf.* Särndal *et al.* 1992, pages 270-271)

$$\widehat{\text{var}}\,(\hat{\mu}_{ps} | U_T)$$

$$= (N \hat{\alpha}_{ps(2)})^{-2} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh}^2 \left[\frac{1 - (n_h / N_{Th})}{n_{gh}(n_{gh} - 1)} \right]$$

$$\sum_{k \in s_{gh}} \left[y_{1k} - \hat{\mu}_{ps}\, y_{2k} - n_{gh}^{-1} \sum_{j \in s_{gh}} (y_{1j} - \hat{\mu}_{ps}\, y_{2j}) \right]^2. \quad (3.1)$$

Although the bias cannot be estimated from the final sample, the theoretical bias of $\hat{\mu}_{ps}$ is well-known to be

$$\text{bias}\ \hat{\mu}_{ps} = \frac{\sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{1k}}{\sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{2k}}$$

$$- \frac{\sum_{k \in U} y_{1k}}{\sum_{k \in U} y_{2k}} + O(n^{-1}) \quad (3.2)$$

as $n \to \infty$. Noting (3.2), the MSE of $\hat{\mu}_{ps}$ does not go to zero in general as the sample size n gets large.

To determine the variance and bias of $\hat{\alpha}_{ps(1)}$, set $y_{2k} = 1$ for all k, so that $\hat{\mu}_{ps}$ and μ become $\hat{\alpha}_{ps(1)}$ and α_1, respectively. One may then apply (3.1) and (3.2) so that

$$\widehat{\text{var}}\ \hat{\alpha}_{ps(1)} = N^{-2} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh}^2 \left[\frac{1 - (n_h / N_{Th})}{n_{gh}(n_{gh} - 1)} \right]$$

$$\sum_{k \in s_{gh}} \left[y_{1k} - n_{gh}^{-1} \sum_{j \in s_{gh}} y_{1j} \right]^2, \quad (3.3)$$

and

$$\text{bias}\ \hat{\alpha}_{ps(1)}$$

$$= N^{-1} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} \left(\sum_{j \in U_{gh}} p_j \right)^{-1} \sum_{k \in U_{gh}} p_k y_{1k} - \alpha_1 + O(n^{-1})$$

$$= O(1)$$

as $n \to \infty$. Cochran (1977, pages 134-135) provided a correction factor, which is of order n^{-2}, to (3.3). This correction factor, however, is irrelevant to (3.1), since the error term due to estimation from the ratio is $O(n^{-2})$.

As usual, the *ratio estimator*, denoted by $\bar{y}_1 / \bar{y}_2$, is defined to be the ratio of the sample mean of y_1 to the sample mean of y_2. That is,

$$\bar{y}_1 / \bar{y}_2 = \sum_{k \in s} y_{1k} \Big/ \sum_{j \in s} y_{2j}.$$

The post-stratified and ratio estimators are identical when $G = H = 1$. Since we will be using only one stratum (*i.e.*, $H = 1$) in section 6, we need not reference separate theory for the ratio estimator.

3.2 The Phone-weighted Estimator

Since the post-stratified estimator, $\hat{\mu}_{ps}$, is biased, two alternative estimators are suggested. One is the phone-weighted estimator, which takes into account the probability that an individual has a phone. In this section we assume that the p_k are known for all $k \in s$ or can be estimated accurately. Estimation of p_k using the PUMS is discussed in section 5.

For a crude estimate of α_i for $i = 1, 2$, use

$$\tilde{\alpha}_{w(i)} = N^{-1} \sum_{h=1}^{H} N_{Th} n_h^{-1} \sum_{k \in s_h} p_k^{-1} y_{ik}. \quad (3.4)$$

Then, estimate μ by

$$\hat{\mu}_w = \tilde{\alpha}_{w(1)} / \tilde{\alpha}_{w(2)}, \quad (3.5)$$

which is asymptotically unbiased for μ, since $\tilde{\alpha}_{w(i)}$ is unbiased for $\alpha_{w(i)}$, for $i = 1, 2$. A valid estimate of the variance of $\hat{\mu}_w$ is shown to be

$$\widehat{\text{var}}\,\hat{\mu}_w = [N\tilde{\alpha}_{w(2)}]^{-2}\sum_{h=1}^{H}\frac{N_{Th}^2\,[1-(n_h/N_{Th})]}{n_h(n_h-1)}$$

$$\sum_{k\in s_h}\left[\frac{y_{1k}-\hat{\mu}_w y_{2k}}{p_k}-n_h^{-1}\sum_{j\in s_h}\frac{y_{1j}-\hat{\mu}_w y_{2j}}{p_j}\right]^2. \quad (3.6)$$

Since the estimator, $\hat{\mu}_w$, is asymptotically unbiased, then a valid estimate of the MSE of $\hat{\mu}_w$ is identical to the estimate of the variance.

Setting $y_{2j} = 1$ in (3.4) and (3.5) allows a valid estimate of $\alpha_{w(1)}$ to be

$$\tilde{\alpha}_{w(1)} = \left[\sum_{h=1}^{H} N_{Th}\,n_h^{-1}\sum_{k\in s_h} p_k^{-1}\right]^{-1}\sum_{h=1}^{H} N_{Th}n_h^{-1}\sum_{k\in s_h} p_k^{-1}y_{1k}.$$

The variance of $\hat{\alpha}_{w(1)}$ may be estimated by setting $y_{2j} = 1$ in (3.6).

3.3 The Post-stratified Phone-weighted Estimator

Another proposed estimator combines post-stratification with the phone-weighted estimator, and is perhaps the best among the four, when sample sizes are large enough to justify post-stratification. This new estimator requires, however, that all N_{gh} be large enough so that with high probability the n_{gh} are not too small. To estimate α_i we use

$$\tilde{\alpha}_{\text{psw}(i)} = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}\left[\sum_{j\in s_{gh}} p_j^{-1}\right]^{-1}\sum_{k\in s_{gh}} p_k^{-1}\,y_{ik},$$

for $i = 1, 2$. We then estimate μ by $\hat{\mu}_{\text{psw}} = \hat{\alpha}_{\text{psw}(1)}/\hat{\alpha}_{\text{psw}(2)}$. The estimate of the variance of $\hat{\mu}_{\text{psw}}$ is

$$\widehat{\text{var}}\,\hat{\mu}_{\text{psw}} = [N\tilde{\alpha}_{\text{psw}(2)}]^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}^2\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-2}$$

$$\left(\frac{n_{gh}}{n_{gh}-1}\right)\left(1-\frac{n_h}{N_{Th}}\right)\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k}-\hat{\mu}_{\text{psw}}y_{2k}-\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m}-\hat{\mu}_{\text{psw}}y_{2m})\right]^2. \quad (3.7)$$

If any of the n_{gh} terms are small, then one might instead prefer the estimator

$$\widehat{\widehat{\text{var}}}\,\hat{\mu}_{\text{psw}} = [N\tilde{\alpha}_{\text{psw}(2)}]^{-2}\sum_{h=1}^{H} N_h^2\left(\sum_{j\in s_h} p_j^{-1}\right)^{-2}$$

$$\left(\frac{n_h}{n_h-1}\right)\left(1-\frac{n_h}{N_{Th}}\right)\sum_{g=1}^{G}\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k}-\hat{\mu}_{\text{psw}}y_{2k}-\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m}-\hat{\mu}_{\text{psw}}y_{2m})\right]^2. \quad (3.8)$$

Notice that if N_{Tgh} were known, which is however unlikely, then a more familiar and intuitive estimator of var $\hat{\mu}_{\text{psw}}$ would be

$$\widetilde{\text{var}}\,\hat{\mu}_{\text{psw}} = [N\tilde{\alpha}_{\text{psw}(2)}]^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G}\frac{N_{Tgh}^2}{n_{gh}(n_{gh}-1)}$$

$$\left(1-\frac{n_{gh}}{N_{Tgh}}\right)\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k}-\hat{\mu}_{\text{psw}}y_{2k}-\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m}-\hat{\mu}_{\text{psw}}y_{2m})\right]^2. \quad (3.9)$$

Since N_{Tgh} typically is unknown, then (3.9) usually is not a practical estimator. However, (3.9) helps motivate (3.7) and (3.8), which are quite practical.

Since the estimator, $\hat{\mu}_{\text{psw}}$, is asymptotically unbiased, then a valid estimate of the MSE is identical to the estimate of the variance. Further, setting $y_{2j} = 1$ in (3.7) and (3.8) allows one to estimate the variance of $\hat{\alpha}_{\text{psw}(1)}$.

When $G = 1$, the estimator $\hat{\mu}_{\text{psw}}$ does not reduce to $\hat{\mu}_w$, as one might naively anticipate. The preferred estimator when $G = 1$ is $\hat{\mu}_w$, since $\hat{\mu}_w$ is based on only one ratio, whereas $\hat{\mu}_{\text{psw}}$ is based on a ratio of ratios. The estimator $\hat{\mu}_{\text{psw}}$ requires large sample sizes in each stratum-group category, but $\hat{\mu}_w$ requires only a large overall sample size. When $H = G = 1$, however, the estimators $\hat{\mu}_w$ and $\hat{\mu}_{\text{psw}}$ are identical; the variance estimators based on $\hat{\mu}_w$ are preferable to those based on $\hat{\mu}_{\text{psw}}$ because the estimates of the variance of $\hat{\mu}_{\text{psw}}$ are based on a ratio of ratios.

4. ASYMPTOTIC MEAN SQUARED ERRORS

The asymptotic mean squared errors of the estimators defined in section 3 now are stated. The proofs follow from Taylor linearization and are given in the appendix, along with the minor regularity conditions needed.

4.1 The Post-stratified Estimator

To find the asymptotic theoretical variance of the post-stratified estimator of μ, we first define

$$\alpha_i^* = \text{plim}_{n\to\infty}\hat{\alpha}_{ps(i)} = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}\left(\sum_{j\in U_{gh}} p_j\right)^{-1}\sum_{k\in U_{gh}} p_k y_{ik}, \quad (4.1)$$

for $i = 1, 2$, and also define

$$\mu^* = \alpha_1^*/\alpha_2^*. \quad (4.2)$$

Note that $\alpha_i^* \neq \alpha_i$ and $\mu^* \neq \mu$ in general. The asymptotic theoretical variance of $\hat{\mu}_{ps}$ is

$$\text{var}\,\hat{\mu}_{ps} = (N\alpha_2^*)^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{N_{gh}^2\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]}{n_h\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_{gh}} p_j\right) - 1\right]}\sum_{k\in U_{gh}} p_k \left[y_{1k} - \mu^* y_{2k} - \frac{\sum_{j\in U_{gh}} p_j(y_{1j} - \mu^* y_{2j})}{\sum_{j\in U_{gh}} p_j}\right]^2 + O\left(n^{-2} + N^{-1}\right) \quad (4.3)$$

as $n \to \infty$. The asymptotic bias of $\hat{\mu}_{ps}$ was shown in (3.2) to be $O(1)$ as $n \to \infty$. Therefore, the asymptotic MSE of $\hat{\mu}_{ps}$ is also $O(1)$ as $n \to \infty$.

4.2 The Phone-weighted Estimator

The asymptotic theoretical variance of the phone-weighted estimator of μ is

$$\text{var}\,\hat{\mu}_w = (N\alpha_2)^{-2}\sum_{h=1}^{H} \frac{\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]\left(\sum_{j\in U_h} p_j\right)}{n_h\left[\left(\sum_{j\in U_h} p_j\right) - 1\right]}\sum_{k\in U_h} p_k \left[\frac{y_{1k} - \mu y_{2k}}{p_k} - \frac{\sum_{j\in U_h}(y_{1j} - \mu y_{2j})}{\sum_{j\in U_h} p_j}\right]^2 + O\left(n^{-2} + N^{-1}\right). \quad (4.4)$$

Since $\hat{\mu}_w$ is asymptotically unbiased, then its MSE is the same as the right hand side of (4.4).

4.3 The Post-stratified Phone-weighted Estimator

The asymptotic theoretical variance of the post-stratified phone-weighted estimator of μ is

$$\text{var}\,\hat{\mu}_{psw} = (N\alpha_2)^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]}{n_h\left[\left(\sum_{j\in U_{gh}} p_j\right) - 1\right]}\sum_{k\in U_{gh}} p_k^{-1}\left[y_{1k} - \mu y_{2k} - N_{gh}^{-1}\sum_{j\in U_{gh}}(y_{1j} - \mu y_{2j})\right]^2 + O\left(n^{-2} + N^{-1}\right). \quad (4.5)$$

Since $\hat{\mu}_{psw}$ is asymptotically unbiased, then its MSE is the same as the right hand side of (4.5).

5. ESTIMATING THE p_k USING PUBLIC USE MICRODATA SAMPLES

The United States Bureau of the Census produced the Public Use Microdata Samples (PUMS), which include 1% and 5% samples of the population in each of the 50 states and Washington, D.C., for year 1990. For each person selected in the sample, 75 household variables and 75 personal variables are listed, where each household has a clearly defined head of household. We utilize the PUMS for two reasons. We estimate the p_k using the PUMS in this section, whereas in section 6 we run simulations on the PUMS to construct examples for comparing and contrasting the estimators.

In this article, we use the 5% sample from Virginia. Since 5% represents a huge number of households, we treat this sample as if it were the entire population of Virginia. Since we are interested in telephone surveys, then from this 5% sample we will sample households. Inferences may be made on personal variables, such as high school graduation rate, and household variables, such as the number of cars in a household or household income. Information pertaining to whether or not each household has a phone is included in the PUMS. We removed from our study all households whose telephone status is listed as "not applicable." Such households were either vacant or were group quarters (institutions and non-institutions). The number of households remaining in 1990 is 110,744, of which 104,606 have phones; hence, the proportion of these households which have phones is 94.5%

Using generalized linear regression with a log – log link on the 5% sample from Virginia along with the household weights assigned in the PUMS, we estimate p_k, which is the probability, or propensity, that the kth household has a phone. McCullagh and Nelder (1991, pages 107-110)

recommended the use of a log – log link when the probabilities are close to one, and we found that this link provided a good fit. We also found that the logit link function provided a poor fit.

The PUMS household weights are used when estimating the p_k but are not used elsewhere herein. In particular, in section 6 when constructing Monte Carlo samples of the PUMS population, the samples are simple random samples from the telephone population.

Examples of estimating the p_k

Six covariates, the number of persons in the household, tenure (home owner or renter), the date the head of household moved into the dwelling, household income, household language, and race of the head of household, are used to estimate the p_k. These six covariates were chosen, along with the categories for each covariate, based on a thorough analysis of the 1990 PUMS using generalized linear regression techniques in SAS. All of these covariates were found to be highly statistically significant. Estimates of the p_k are made by summing the appropriate estimates of the covariates in Table 1. The covariate for the number of persons should be multiplied by the number of persons in the household; however, if the number of persons exceed five, then, for computations, convert this number of persons to five. For example, if the household consists of three English-speaking Asian Americans with two cars in a house purchased in 1987, where the household income is $75,000, then Table 1 indicates that the estimate of p_k is the solution to

$$\log(-\log p_k) = 3\times 0.2747 - 0.5552 + 0.5920 + 0.1896 + 1.0004 + 0.6156 + 0.0000.$$

Notice that in Table 1 within each of the covariates *date moved in*, *number of cars*, and *income*, the values corresponding to the categories are monotonically decreasing, as anticipated, except when income is negative.

An adjustment which should be made when using random digit dialing is to ask each respondent the number of phone lines in the household, and multiply that number by the estimate of p_k from Table 1 to obtain a new estimate of p_k Consequently, p_k now is a weight, rather than a probability. For the simulations discussed in section 6 this adjustment is not necessary, since households are equally likely to be selected using simple random sampling from the PUMS, regardless of the number of phone lines.

Table 1

Values of covariates for estimating $p_k$using the Virginia 5% PUMS. Standard errors are in parentheses. If the number of persons exceeds five, then convert this number to five. The covariate "tenure" did not appear in the 1980 PUMS. The 1980 category "$40,000 to $49,999" actually includes "$40,000 or greater". The "other" category for the 1980 covariate "language" includes Spanish.

Covariate	Category	1990 Value		1980 Value	
Number of persons		0.2747	(0.0022)	0.1929	(0.0020)
tenure	home owner	-0.5552	(0.0079)	-0.7845	(0.0057)
	renter	0.0000	(0.0000)	0.0000	(0.0000)
date moved in	1989 or 1990	0.9742	(0.0121)	NA	
	1985 to 1988	0.5920	(0.0119)	NA	
	1980 to 1984	0.3489	(0.0138)	NA	
	1970 to 1979	0.2185	(0.0136)	NA	
	1969 or earlier	0.0000	(0.0000)	NA	
number of cars	0	1.2927	(0.0152)	0.8633	(0.0118)
	1	0.6842	(0.0143)	0.3981	(0.0109)
	2	0.1896	(0.0145)	0.0399	(0.0112)
	3 or more	0.0000	(0.0000)	0.0000	(0.0000)
income	less than $0	3.5325	(0.1294)	2.3639	(0.0830)
	$0 to $9,999	3.7929	(0.0539)	2.5238	(0.0260)
	$10,000 to $19,999	3.4878	(0.0538)	1.9763	(0.0258)
	$20,000 to $29,999	3.0299	(0.0539)	1.0220	(0.0269)
	$30,000 to $39,999	2.4297	(0.0543)	0.3889	(0.0317)
	$40,000 to $49,999	1.8899	(0.0556)	0.0000	(0.0000)
	$50,000 to $59,999	1.5992	(0.0578)	NA	
	$60,000 to $69,999	1.2144	(0.0631)	NA	
	$70,000 to $79,999	1.0004	(0.0704)	NA	
	$80,000 or greater	0.0000	(0.0000)	NA	
language	English	0.6156	(0.0164)	0.4232	(0.0153)
	Spanish	0.4889	(0.0216)	NA	
	other	0.0000	(0.0000)	0.000	(0.0000)
race	black	-0.4233	(0.0064)	-0.3837	(0.0058)
	other	0.0000	(0.0000)	0.0000	(0.0000)
intercept		-7.6707	(0.0588)	-4.9024	(0.0322)

Table 1 thus can be used for estimating p_k when conducting telephone surveys. When a generalized linear regression model calculated from a PUMS of an earlier date is used to analyse a later survey, rescaling should be performed to take into account changes in the distribution of household income across time. Table 1 also gives the coefficients of a model calculated from the 1980 PUMS. We discuss in section 6 an example when the 1980 PUMS model is used to calculate p_k for a sample from the 1990 PUMS population. We note that the 1980 PUMS did not include "date moved in" and that a better fitting model arose when the language categories "Spanish" and "other" were combined. In addition, median household income almost doubled between 1980 and 1990, so fewer income categories were used in 1980.

Although Table 1 is convenient when sampling from the PUMS and performing simulations, the covariates listed in Table 1 might not be available in actual surveys involving random digit dialing. One may reproduce Table 1 using different covariates, or one may estimate the p_k according to the following alternative method.

An alternative method for estimating the p_k

The participants in a telephone survey based on random digit dialing may be asked the following two questions: "(1) How many telephone lines have been in your household during the past twelve months? (2) During the past twelve months, how many months was each telephone line in service?" Now, let p_k be the sum of the answers to question (2). For example, in a household with two phone lines, where one of the lines was in service all twelve months and the other was in service only five months, the estimate of p_k would be 12+5=17. Again, p_k represents a weight rather than a probability here. Asking the respondent this second question is similar to an approach advocated by Brick *et al.* (1994), who also suggested weighting the data to take into account the probability that a household has phone service.

6. INFERENCES ON HOUSEHOLD AND PERSONAL VARIABLES

We will compare the four proposed estimators of μ as we make inferences on the high school graduation rate among people at least 21-years-old, the mean number of cars per household, and the mean household income, in the state of Virginia. We performed 100,000 simulations of simple random samples of 500 households with telephones from the 1990 Virginia 5% PUMS using one stratum (*i.e.*, H=1).

In section 6.1, two sets of p_k are used. One is based upon a GLIM regression fit to the 1990 PUMS, and the other is based upon a GLIM fit to the 1980 PUMS with the income categories inflated by the ratio of the 1990 median household income ($32,800) to the 1980 median household income ($17,510). Using the 1980 p_k to estimate a 1990 parameter demonstrates how well our method works when GLIM coefficients are used for future data sets, provided than an adjustment for inflation is made. Only the 1990 p_k are used in section 6.2 and section 6.3.

Post-stratification should be used when the sample sizes are sufficiently large. Non-post-stratified estimators may be compared to each other, and post-stratified estimators may be compared to each other. Comparing $\bar{y}_1/\bar{y}_2$ to $\hat{\mu}_w$ is appropriate, and comparing $\hat{\mu}_{ps}$ to $\hat{\mu}_{psw}$ is appropriate. These comparisons show the improvements when using the p_k in the estimators.

6.1 Estimating the High School Graduation Rate

Using the entire 1990 Virginia 5% PUMS, the mean high school graduation rate among all Virginians at least 21-years-old is μ=0.75118. When estimating the graduation rate using a simple random sample and $\hat{\mu}_{ps}$ or $\hat{\mu}_{psw}$, we post-stratify by gender (male, female), age (less than 45 years old, at least 45 years old), and race (black, other) of the head of household. The p_k are estimated using Table 1. The values of the biases and standard deviations discussed below are shown in Table 2, when 1990 p_k are used.

Table 2
Biases and Standard Deviations of Estimates of High School Graduation Rate

	Estimator			
	not post-stratified		post-stratified	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	0.01471	0.00722	0.01461	0.00874
telephone bias	0.01472	0.00720	0.01463	0.00850
second phase bias	0.00000	0.00002	-0.00002	0.00024
theoretical bias	0.00777	0	0.00663	0
simulated standard deviation	0.01683	0.01737	0.01605	0.01643
estimated standard deviation	0.01680	0.01734	0.01601	0.01635*
theoretical standard deviation	0.01700	0.01752	0.01617	0.01658
root mean squared error	0.02236	0.01881	0.02171	0.01861

The true high school graduation rate is 0.75118. Post-stratification is based on gender, age, and race. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is 0.01610.

The *aggregate biases* of the four estimators of μ are estimated by the average over 100,000 simulations of the difference between the estimate from a sample of size 500 and μ. These *aggregate biases* produced by $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are estimated to be 0.01471, 0.00722, 0.01461, and 0.00874, respectively, when 1990 p_k are used. Hence using the p_k reduces the bias of the non-post-stratified estimator by 51%, and reduces the bias of the post-stratified estimator by 40%.

When the 1980 p_k are used, similar results arise. These *aggregate biases* produced by $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are estimated to be 0.00578, and 0.00856, respectively, when the 1980 p_k are used. These results, however, are not summarized in the tables.

The *telephone bias*, listed in Table 2, is the bias obtained when the entire telephone population, U_T, is sampled when calculating $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$. This bias is caused by the fact that U_T is sampled rather than U. Throughout this example, we use the convention of listing the estimates based on the 1980 p_k in parentheses, when these estimates differ from those based on the 1990 p_k. The *telephone biases* are 0.01472, 0.00720 (0.00577), 0.01463, and 0.00850 (0.00838), and are relatively close to the *aggregate biases*.

The *second phase bias* is the difference between the *aggregate bias* and the *telephone bias*, and is caused by the fact that the estimator approximates a ratio. This *second phase bias*, modulus rounding error, for $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are estimated to be 0.00000, 0.00002 (0.00001), -0.00002, and 0.00024 (0.00018), respectively. Hence, the *second phase bias* is trivial compared to the *telephone bias* for this example.

The *theoretical biases*, based on (3.2), of $\bar{y}_1/\bar{y}_2$, and $\hat{\mu}_{ps}$ are 0.00777 (0.00905) and 0.00663 (0.00678), respectively. These biases differ from the *aggregate biases*, since (3.2) is based on all possible phone populations, whereas the *aggregate biases* are conditional on the one realization of the phone population. The theoretical bias is based upon the model that each household has a phone with probability p_k and hence is dependent upon the model used to fit p_k. Since $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are aymptotically unbiased, then their *theoretical biases* are defined to be zero.

The *simulated standard deviations* of the 100,000 simulated estimates of μ for $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are 0.01683, 0.01737 (0.01734), 0.01605, and 0.01643 (0.01634). These four numbers are fairly close to the *estimated standard deviations*, which are the squareroot of the average estimated variance of the estimator of μ, based on (3.1), (3.6), and (3.7). Specifically, *these estimated standard deviations* are 0.01680, 0.01734 (0.01732), 0.01601, and 0.01635 (0.01628), respectively. The estimated alternative standard deviation, based on (3.8), of $\hat{\mu}_{psw}$ is 0.01610 (0.01606), which again is fairly close to the value 0.01635 (0.01628). The *theoretical standard deviations* are 0.01700 (0.01697), 0.01752 (0.01749), 0.01617 (0.01621), and 0.01658 (0.01653), based on the entire 1990 Virginia 5% PUMS and (4.3), (4.4), and (4.5). These *theoretical standard deviations* also are close to the other standard deviations calculated.

Using the p_k reduces the *aggregate bias* in the non-post-stratified estimator by 51% (61%), and in the post-stratified estimator by 40% (41%). The standard deviation, however, increases slightly. Using the *aggregate biases* and the *simulated standard deviations*, the root mean squared errors of the estimators $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are 0.02236 (0.02236) 0.01881 (0.01828), 0.02171 (0.02171), and 0.01861 (0.01844), respectively. Hence, using the p_k reduces the MSE in the non-post-stratified estimator by 29% (33%), and reduces the MSE in the post-stratified estimator by 27% (28%). Notice that there is little difference between $\bar{y}_1/\bar{y}_2$ and $\hat{\mu}_{ps}$ and between $\hat{\mu}_w$ and $\hat{\mu}_{psw}$, in terms of MSE. Therefore, post-stratification offers little improvement.

6.2 Estimating the Mean Number of Cars per Household

The mean number of cars per household is 1.80397, as determined by the entire 1990 Virginia 5% PUMS. Post-stratification was based upon household income, using categories {less than \$20,000, at least \$20,000 but less than \$35,000, and at least \$35,000}. The p_k are again estimated, but this time the covariate "numbers of cars" was excluded from the GLIM fit to the 1990 PUMS, since mean number of cars per household is what is being estimated.

As shown in Table 3, the estimates of the *aggregate biases* using 100,000 simulations of 500 simple random samples are 0.04872, 0.01629, 0.02226, and 0.01471, and the *telephone biases* are 0.04872, 0.01623, 0.02220, and 0.01458, for estimators $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$, respectively. Therefore, the *second phase biases* are rather small. Using the p_k reduces the bias from the non-post-stratified estimator by 67%, and reduces the bias from the post-stratified estimator by 34%. Perhaps the reason why this latter amount of bias that can be removed is smaller than the former is that income is a strong predictor of whether or not a household has a phone (*cf.* Groves 1989, pages 116-119; Thornberry and Massey 1988), and the post-stratification groups for determining $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$ are based on income.

Table 3
Biases and Standard Deviations of Estimates of Mean Number of Cars per Household

	Estimator			
	not post-stratified		post-stratified	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	0.04872	0.01629	0.02226	0.01471
telephone bias	0.04872	0.01623	0.02220	0.01458
second phase bias	0.00000	0.00006	0.00006	0.00013
theoretical bias	0.03388	0	0.00859	0
simulated standard deviation	0.04694	0.04764	0.04162	0.04172
estimated standard deviation	0.04682	0.04753	0.04148	0.04158*
theoretical standard deviation	0.04715	0.04791	0.04152	0.04161
root mean squared error	0.06765	0.05035	0.04720	0.04424

The true mean number of cars per household is 1.80397. Post-stratification is based on income. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is 0.04142.

The standard deviations of the simulations are 0.04694, 0.04764, 0.04162, and 0.04172, respectively. The root mean squared errors for the four estimators are approximately 0.06765, 0.05035, 0.04720, and 0.04424, respectively, so using the p_k reduces the MSE by 45% and 12% for non-post-stratification and post-stratification, respectively.

We also performed simulations, not summarized in the tables, where "number of cars" was retained for the GLIM fit to the 1990 PUMS. These *aggregate biases* for the estimators $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are 0.00116 and 0.00006, respectively, which are much smaller than 0.01629 and 0.01471, the respective *aggregate biases* when "number of cars" was removed from the GLIM fit. Furthermore, we feel that appropriate analysis requires removing the variable being studied (*i.e.*, number of cars) from the GLIM fit to the PUMS.

6.3 Estimating the Mean Household Income

The mean household income is $40,187, as determined by the entire 1990 Virginia 5% PUMS. The p_k are again estimated, but this time the covariate "income" was excluded from the GLIM fit to the 1990 PUMS, since mean household income is what is being estimated.

In Table 4, when estimating household income using a simple random sample of size 500 and $\hat{\mu}_{ps}$ or $\hat{\mu}_{psw}$, we post-stratified only by the race (black, other) of the head of household. The estimates of the *aggregate biases* using 100,000 simulations are $1,412, $640, $1,192, and $633, and the *telephone biases* are $1,414, $640, $1,193, and $630, for estimators $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$, respectively. Thus, the *second phase biases* are small relative to the *telephone biases*. Overall, using the p_k reduces the bias from the non-post-stratified estimator by 55%, and reduces the bias from the post-stratified estimator by 47%.

The standard deviations of the simulations are $1,534, $1,518, $1,502, and $1,488, respectively. Hence the root mean squared errors for the four estimators are approximately $2,085, $1,647, $1,918, and $1,617, respectively, so using the p_k reduces the MSE by 38% and 29% for non-post-stratification and post-stratification, respectively. The improvements from using post-stratification are more minor, according to the MSE criterion.

In Table 5, we again are estimating household income, but this time we post-stratify by gender (male, female), age (less than 45 years old, at least 45 years old), and race (black, other) of the head of household. Note that the non-post-stratified estimators are not affected by this new post-stratification. The estimates of the *aggregate biases* using 100,000 simulations are $1,173 and $757, and the *telephone biases* are $1,177 and $747 for the post-stratified estimators, $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. Again, the *second phases biases* are small relative to the *telephone biases*. Using the p_k reduces the bias from merely post-stratification by 35%.

The theoretical bias for the post-stratified estimator is $463. The standard deviations of the simulations are $1,445 and $1,435, for estimators $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. The root mean squared errors are $1,861 and $1,622, for estimators $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. Hence, using the p_k reduces the MSE of the post-stratified estimator by 24%.

The MSE of $\hat{\mu}_{psw}$ is approximately the same in Table 4 and Table 5. However, the MSE of $\hat{\mu}_{ps}$ decreases somewhat from Table 4 to Table 5.

Table 4
Biases and Standard Deviations of Estimates of Household Income, Post-stratified by Race

	Estimator			
	not post-stratified		post-stratified	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	$1,412	$640	$1,192	$633
telephone bias	$1,414	$640	$1,193	$630
second phase bias	-$2	$0	-$2	$3
theoretical bias	$789	$0	$586	$0
simulated standard deviation	$1,534	$1,518	$1,502	$1,488
estimated standard deviation	$1,537	$1,521	$1,506	$1,491*
theoretical standard deviation	$1,535	$1,518	$1,503	$1,488
root mean squared error	$2,085	$1,647	$1,918	$1,617

The true mean household income is $40,187. Note that $\bar{y}_1/\bar{y}_2$ and $\hat{\mu}_w$ are independent of post-stratification, so their results are identical to those in Table 5. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is $1,490.

Table 5
Biases and standard deviations of estimates of household income, post-stratified by gender, age, and race

	Estimator			
	not post-stratified		post-stratified	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	$1,412	$640	$1,173	$757
telephone bias	$1,414	$640	$1,177	$747
second phase bias	-$2	$0	-$4	$10
theoretical bias	$789	$0	$463	$0
simulated standard deviation	$1,534	$1,518	$1,445	$1,435
estimated standard deviation	$1,537	$1,521	$1,448	$1,438*
theoretical standard deviation	$1,535	$1,518	$1,440	$1,430
root mean squared error	$2,085	$1,647	$1,861	$1,622

The true mean household income is $40,187. Note that $\bar{y}_1/\bar{y}_2$ and $\hat{\mu}_w$ are independent of post-stratification, so their results are identical to those in Table 4. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is $1,421.

7. DISCUSSION

We have proposed here to use publicly available large data bases (*e.g.*, the PUMS) to develop a model for the propensity p_k of a household to have a telephone. We have used, for Virginia in 1990, a GLIM model with a log – log link and predictor variables number of persons, tenure, date moved in, number of cars, household income, language, and race.

We have proposed to use the telephone weights p_k to reduce the bias of estimators due to noncoverage in telephone surveys. This bias can be expected to occur when the variable of interest is related to telephone ownership. The examples we have chosen are all variables of this type and hence the improvements using telephone weights are better than one would expect for variables with little relationship to telephone ownership.

The weights can be combined with post-stratification. We have found that the use of such telephone weights greatly reduces the bias of both non-post-stratified and post-stratified estimators.

Post-stratification requires a large enough sample size so that each post stratum has a negligible probability of being empty. Our experiments dealt with samples of size 500, and hence the number of post strata was relatively limited. Certainly, if one had a large enough sample so that one could post-stratify on the same predictor variables as used to develop the p_k, the use of telephone weights should offer negligible improvement over post-stratification. However, many nationwide telephone opinion polls use approximate sample sizes of 1,000, and we believe for these sample sizes, the use of telephone weighs would offer a genuine improvement.

We have also reported results from using telephone weights developed from the 1980 PUMS on 1990 data, with categories related to household income adjusted for inflation. The results are comparable to those for telephone weights developed from the 1990 PUMS. Therefore, although PUMS data are produced only every ten years and might be as much as twelve years out of date, substantial reductions in the biases of telephone sampling can be made using propensity models derived from older PUMS data sets, provided that the categories are suitably adjusted for inflation.

Finally, the PUMS are divided by state and major metropolitan areas. This allows separate telephone-weighted models to be developed for major geographical units, and this would seem appropriate for large surveys.

ACKNOWLEDGEMENT

The authors are very thankful to an anonymous referee for many helpful suggestions.

APPENDIX: DERIVATIONS OF EQUATIONS

Before deriving the equations in section 3 and section 4, some regularity conditions must be assumed for sequences $\{\alpha_{i1}, \alpha_{i2}, \ldots\}$, for $i = 1, 2$. Further, some lemmas must be proved. Then, the equations involving the estimators $\hat{\mu}_{ps}$, $\hat{\mu}_{w}$, and $\hat{\mu}_{psw}$ will be derived in the subsections below. Whenever the error variable ξ_k is introduced below, then $\xi_k = O_p(1)$ and $E(\xi_k)^2 = O(1)$ as $k \to \infty$. For simplicity (but slight abuse) of notation, the sequence $\{\xi_1, \xi_2, \ldots\}$ will be allowed to be different across different equations.

Condition A: Each α_{ik} represents a sample mean of observations such that $E\alpha_{ik} - \alpha_i = O(k^{-1})$, $E|\alpha_{ik} - \alpha_i|^3 = O(k^{-3/2})$, and $\alpha_{ik} - \alpha_i = O_p(k^{-1/2})$ as $k \to \infty$ for $i = 1, 2$. Let $\mu_k = \alpha_{1k}/\alpha_{2k}$ for $k = 1, 2, \ldots$.

LEMMA A.1 Condition A implies that $E\mu_k - \mu = O(k^{-1})$ as $k \to \infty$.

PROOF: Define the function $f(\gamma_1, \gamma_2) = \gamma_1/\gamma_2$. By a Taylor series linear expansion,

$$\mu_k - \mu = \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2$$

$$= (\alpha_{1k} - \alpha_1)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} + k^{-1}\xi_k$$

$$= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu(\alpha_2)^{-2} + k^{-1}\xi_k.$$

The result follows from Condition A.

Condition B: The sequence $\{\alpha_{i1}, \alpha_{i2}, \ldots\}$ for $i = 1, 2$ satisfies

$$k^{1/2}\left[\begin{pmatrix}\alpha_{1k}\\ \alpha_{2k}\end{pmatrix} - \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right] \stackrel{d}{=}$$

$$N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) + \begin{pmatrix}k^{-1/2}\xi_{1k}\\ k^{-1/2}\xi_{2k}\end{pmatrix}$$

for some constants σ_1^2, σ_2^2, and ρ.

LEMMA A.2 Under Conditions A and B,

$$\text{MSE}\,\mu_k = (\alpha_2)^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) + O(k^{-2}),$$

and

$$\text{var}\,\mu_k = (\alpha_2)^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) + O(k^{-2}),$$

as $k \to \infty$.

PROOF: By a Taylor series linear expansion,

$$\mu_k - \mu = \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2$$

$$= (\alpha_{1k} - \alpha_1)\frac{\partial f(\alpha_1,\alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2)\frac{\partial f(\alpha_1,\alpha_2)}{\partial \alpha_2}$$

$$+\frac{1}{2}\left[(\alpha_{1k} - \alpha_1)^2 \frac{\partial^2 f(\alpha_1,\alpha_2)}{\partial(\alpha_1)^2} + (\alpha_{2k} - \alpha_2)^2 \right.$$

$$\left. \frac{\partial^2 f(\alpha_1,\alpha_2)}{\partial(\alpha_2)^2} + 2\,(\alpha_{1k} - \alpha_1)(\alpha_{2k} - \alpha_2)\frac{\partial^2 f(\alpha_1,\alpha_2)}{\partial \alpha_1\, \partial \alpha_2}\right]$$

$$+ k^{-3/2}\xi_k$$

$$= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu\,(\alpha_2)^{-1}$$

$$+ (\alpha_{2k} - \alpha_2)^2 \mu\,(\alpha_2)^{-2} - (\alpha_{1k} - \alpha_1)\,(\alpha_{2k} - \alpha_2)\,(\alpha_2)^{-2}$$

$$+ k^{-3/2}\xi_k$$

$$= \alpha_2^{-1}(\alpha_{1k} - \mu\,\alpha_{2k})\left[1 - \alpha_2^{-1}(\alpha_{2k} - \alpha_2)\right] + k^{-3/2}\xi_k.$$

Therefore,

$$(\mu_k - \mu)^2 = \alpha_2^{-2}(\alpha_{1k} - \mu\,\alpha_{2k})^2\left[1 - 2\alpha_2^{-1}(\alpha_{2k} - \alpha_2)\right] + k^{-2}\xi_k,$$

which implies that

$$\text{MSE}\,\mu_k = \alpha_2^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) - 2\alpha_2^{-3}$$

$$\text{cov}\,\{(\alpha_{1k} - \mu\,\alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} + k^{-2}\xi_k. \qquad \text{(A.1)}$$

Now we will show that the covariance term in (A.1) is asymptotically negligible. Since

$$k^{1/2}(\alpha_{1k} - \mu\,\alpha_{2k}) \underline{d} N(0, \sigma_3^2) + k^{-1/2}\xi_k$$

for some constant σ_3^2, then

$$k(\alpha_{1k} - \mu\,\alpha_{2k})^2 \underline{d} \sigma_3^2\chi_1^2 + k^{-1/2}\xi_k,$$

where χ_1^2 denotes a chi squared random variable with one degree of freedom. Furthermore,

$$k^{1/2}(\alpha_{2k} - \alpha_2) \underline{d} N(0, \sigma_2^2) + k^{-1/2}\xi_k.$$

If the signs on α_{ik} are negated for $i = 1, 2$, then $k(\alpha_{1k} - \mu\,\alpha_{2k})^2$ does not change but $k^{1/2}(\alpha_{2k} - \alpha_2)$ is negated. Therefore, by symmetry,

$$\text{cov}\{k(\alpha_{1k} - \mu\,\alpha_{2k})^2,\ k^{1/2}(\alpha_{2k} - \alpha_2)\} = O(k^{-1/2})$$

as $k \to \infty$. Hence,

$$\text{cov}\{(\alpha_{1k} - \mu\,\alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} = O(k^2) \qquad \text{(A.2)}$$

as $k \to \infty$. Combining (A.1) and (A.2) the first part of the lemma follows. Since Lemma A.1 implies that

$$\text{bias}\ u_k = O(k^{-1})$$

as $k \to \infty$, then the second part of this lemma follows.

Condition C: Defining $\alpha_{Ti} = \text{plim}_{n\to\infty}\hat{\alpha}_i$ given U_T, the estimator, $\hat{\alpha}_i$, of α_i satisfies the following, for $i = 1, 2$:

$$E\,(\hat{\alpha}_i \mid U_T) - \alpha_{Ti} = O(n^{-1});$$

$$\text{Given}\ U_T,\ \hat{\alpha}_i - \alpha_{Ti} = O_p(n^{-1/2});$$

and

$$E(\{|\hat{\alpha}_i - \alpha_{Ti}|^3\}\mid U_T) = O(n^{-3/2})$$

as $k \to \infty$.

Condition D: Given U_T,

$$n^{1/2}\left[\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\end{pmatrix} - \begin{pmatrix}\alpha_{T1}\\ \alpha_{T2}\end{pmatrix}\right] \underline{d}$$

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right) + \begin{pmatrix}n^{-1/2}\,\xi_n\\ n^{-1/2}\,\xi_n\end{pmatrix}$$

for some positive definite matrix Σ, where $\alpha_{Ti} = \text{plim}_{n\to\infty}\hat{\alpha}_i$ given U_T. Also,

$$E(\{|\hat{\alpha}_i - \alpha_{Ti}|^3\}\mid U_T) = O(n^{-3/2})$$

as $n \to \infty$, for $i = 1, 2$.

THEOREM A.1 Under conditions C and D, we have that

$$\text{var}\,\hat{\mu} = \alpha_2^{-2}\,E\,\text{var}\,(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) + O(n^{-2} + N^{-1})$$

as $n \to \infty$, where $\mu_T = \alpha_{T1}/\alpha_{T2}$.

PROOF: First we determine $E\,\text{var}\,(\hat{\mu}\mid U_T)$. Under Condition D we apply Lemma A.2 to obtain

$$\text{var}\,(\hat{\mu}\mid U_T) = \alpha_{T2}^{-2}\,\text{var}\,(\hat{\alpha}_1 - \mu_{U_T}\hat{\alpha}_2 \mid U_T) + n^{-2}\xi_n. \qquad \text{(A.3)}$$

Since

$$\alpha_{T2}^{-2} = (\alpha_2)^{-2} + N^{-1/2}\xi_n$$

and

$$\text{var}\,(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) = n^{-1}\xi_n,$$

then (A.3) implies that

$$E\,\text{var}\,(\hat{\mu}\mid U_T) =$$

$$\alpha_2^{-2}E\,\text{var}(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) + O(n^{-2} + n^{-1}N^{-1/2}) \qquad \text{(A.4)}$$

as $n \to \infty$. Now we determine $\text{var}\,E(\hat{\mu}\mid U_T)$. Condition C and Lemma A.1 imply that

$$E(\hat{\mu}\mid U_T) = \mu_T + n^{-1}\xi_n = \mu + (n^{-1} + N^{-1/2})\xi_n.$$

Hence,

$$\operatorname{var} E(\hat{\mu}|U_T) = O(n^{-2} + N^{-1}) \quad \text{(A.5)}$$

as $n \to \infty$. Combining (A.4) with (A.5) the result follows.

A.1 The post-stratified estimator

Here we derive the equations related to the post-stratified estimator, $\hat{\mu}_{ps}$, where $\hat{\alpha}_{ps(1)}$ and $\hat{\alpha}_{ps(2)}$ satisfy Conditions C and D. Note that

$$E(\hat{\alpha}_{ps(i)}|U_T) = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh} N_{Tgh}^{-1} \sum_{k\in U_{Tgh}} y_{ik}$$

for $i = 1, 2$, and we define $\mu_{T,ps} = E(\hat{\alpha}_{ps(1)}|U_T)/E(\hat{\alpha}_{ps(2)}|U_T)$. Recall the definitions of α_i^* and μ^* in (4.1) and (4.2).

Derivation of (4.3), the asymptotic variance of $\hat{\mu}_{ps}$:

Since

$$\operatorname{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)}|U_T, n_{gh})$$

$$= N^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{N_{gh}^2[1 - n_{gh}/N_{Tgh}]}{n_{gh}(N_{Tgh}-1)} \sum_{k\in U_{Tgh}}$$

$$\left[y_{1k} - \mu_{T,ps}y_{2k} - N_{Tgh}^{-1}\sum_{j\in U_{Tgh}}(y_{1j} - \mu_{T,ps}y_{2j})\right]^2 \quad \text{(A.6)}$$

then

$$E\operatorname{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)}|U_T, n_{gh})$$

$$= N^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{N_{gh}^2\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]}{n_h\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_{gh}} p_j\right) - 1\right]}$$

$$\sum_{k\in U_{gh}} p_k\left[y_{1k} - \mu^* y_{2k} - \frac{\sum_{j\in U_{gh}} p_j(y_{1j} - \mu^* y_{2j})}{\sum_{j\in U_{gh}} p_j}\right]^2$$

$$+ O(n^{-2} + n^{-1}N^{-1/2}) \quad \text{(A.7)}$$

as $n \to \infty$. Also, since

$$E(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)}|U_T, n_{gh})$$

$$= N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}N_{Tgh}^{-1}\sum_{k\in U_{Tgh}}(y_{1k} - \mu_{T,ps}y_{2k}), \quad \text{(A.8)}$$

then

$$E[\operatorname{var}E(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)}|U_T, n_{gh})|U_T] = 0. \quad \text{(A.9)}$$

Since Theorem A.1 and (A.7) imply that

$$\operatorname{var}\hat{\mu}_{ps} = (\alpha_2^*)^{-2}E\operatorname{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)}|U_T, n_{gh}) + O(n^{-2} + N^{-1}) \quad \text{(A.10)}$$

as $n \to \infty$, then (A.9) implies (4.3).

Derivation of (3.1), the estimated variance of $\hat{\mu}_{ps}$:

In light of (A.6) we have the estimator

$$\widehat{\operatorname{var}}\left(n_{gh}^{-1}\sum_{k\in s_{gh}}\{y_{1k} - \hat{\mu}_{T,ps}\, y_{2k}\}|U_T, n_{gh}\right)$$

$$= \frac{1 - n_{gh}/N_{Tgh}}{n_{gh}(n_{gh}-1)}\sum_{k\in s_{gh}}$$

$$\left[y_{1k} - \mu_{T,ps}y_{2k} - n_{gh}^{-1}\sum_{j\in s_{gh}}(y_{1j} - \mu_{T,ps}y_{2j})\right]^2.$$

Using (A.10) the result follows.

Derivation of (3.2), the estimated bias of $\hat{\mu}_{ps}$:

Lemma A.1 implies that

$$\hat{\mu}_{ps} - \mu^* = O(n^{-1})$$

as $n \to \infty$. Since

$$E\hat{\alpha}_{ps(i)} = \alpha_i^* + O(N^{-1})$$

as $N \to \infty$ for $i = 1, 2$, the result follows.

A.2 The phone-weighted estimator

Here we derive the equations related to the phone-weighted estimator, $\hat{\mu}_w$, under Conditions C and D, where $\tilde{\alpha}_{w(1)}$ and $\tilde{\alpha}_{w(2)}$ satisfy Conditions C and D. Note that

$$E(\tilde{\alpha}_{w(i)}|U_T) = N^{-1}\sum_{k\in U_T} y_{ik}/p_k$$

for $i = 1, 2$, and we define $\mu_{T,w} = E(\tilde{\alpha}_{w(1)}|U_T)/E(\tilde{\alpha}_{w(2)})|U_T)$.

Derivation of (4.4), the asymptotic variance of $\hat{\mu}_{ps}$:

Since

$$\operatorname{var}(\tilde{\alpha}_{w(1)} - \mu_{T,w}\,\tilde{\alpha}_{w(2)}|U_T)$$

$$= N^{-2}\sum_{h=1}^{H}\frac{(N_{Th} - n_h)N_{Th}}{n_h(N_{Th}-1)}\sum_{k\in U_{Th}}$$

$$\left[\frac{y_{1k} - \mu_{T,w}y_{2k}}{p_k} - N_{Th}^{-1}\sum_{j\in U_{Th}}\frac{y_{1j} - \mu_{T,w}y_{2j}}{p_j}\right]^2,$$

then

$$E\,\mathrm{var}\left(\tilde{\alpha}_{w(1)} - \mu_{T,w}\,\tilde{\alpha}_{w(2)} \mid U_T\right)$$

$$= N^{-2} \sum_{h=1}^{H} \frac{\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]\left(\sum_{j\in U_h} p_j\right)}{n_h\left[\left(\sum_{j\in U_h} p_j\right) - 1\right]} \qquad \text{(A.11)}$$

$$\sum_{j\in U_h} p_k \left[\frac{y_{1k} - \mu\, y_{2k}}{p_k} - \frac{\sum_{j\in U_h}(y_{1j} - \mu\, y_{2j})}{\sum_{j\in U_h} p_j}\right]^2$$

$$+ O\,(n^{-1}\,N^{-1/2}) \qquad \text{(A.12)}$$

as $n \to \infty$. Applying Theorem A.1 to (A.11) the result follows.

Derivation of (3.6), the estimated variance of $\hat{\mu}_{ps}$:

In light of Theorem A.1 a valid estimate of $\mathrm{var}\,\hat{\mu}_w$ also estimates

$$(\alpha_2)^{-2} E\,\mathrm{var}\left(\hat{\alpha}_{w(1)} - \mu_{T,w}\hat{\alpha}_{w(2)} \mid U_T\right),$$

which is equivalent to

$$(\alpha_2)^{-2} E\,\mathrm{var}\left(N^{-1}\sum_{h=1}^{H} \frac{N_{Th}}{n_h}\sum_{k\in s_h} \frac{y_{1k} - \mu_{T,w}\, y_{2k}}{p_k} \mid U_T\right).$$

The result follows.

A.3 The post-stratified phone-weighted estimator

Here we derive the equations related to the post-stratified estimator, $\hat{\mu}_{psw}$, under Conditions C and D, where $\hat{\alpha}_{psw(1)}$ and $\hat{\alpha}_{psw(2)}$ satisfy Conditions C and D. Lemma A.1 implies that

$$E\left(\hat{\alpha}_{psw(i)} \mid U_T\right) = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh} \frac{\sum_{k\in U_{Tgh}} p_k^{-1} y_{ik}}{\sum_{k\in U_{Tgh}} p_k^{-1}} + n^{-1}\,\xi_n$$

for $i = 1, 2$, and we define $\mu_{T,psw} = E(\hat{\alpha}_{psw(1)} \mid U_T)/E(\hat{\alpha}_{psw(2)} \mid U_T)$.

Derivation of (4.5), the asymptotic variance of $\hat{\mu}_{psw}$:

Using Lemma A.2 it follows that

$$\mathrm{var}\left(\frac{\sum_{k\in s_{gh}} p_k^{-1}\{y_{1k} - \mu_{T,psw}\, y_{2k}\}}{\sum_{k\in s_{gh}} p_k^{-1}} \mid U_T, n_{gh}\right)$$

$$= \frac{1 - n_{gh}/N_{Tgh}}{n_{gh}(N_{Tgh} - 1)}\left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}} p_j^{-1}\right)^{-2}$$

$$\sum_{k\in U_{Tgh}} p_k^{-2}\left[y_{1k} - \mu_{psw,U_T} y_{2k} - \left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}} p_j^{-1}\right)^{-1}\right.$$

$$\left.\left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}}\left\{\frac{y_{1j} - \mu_{T,psw}\, y_{2j}}{p_j}\right\}\right)\right]^2 + n^{-2}\,\xi_n.$$

Therefore,

$$E\left[\mathrm{var}\left(\frac{\sum_{k\in s_{gh}} p_k^{-1}\{y_{1k} - \mu_{T,psw}\, y_{2k}\}}{\sum_{k\in s_{gh}} p_k^{-1}} \mid U_T, n_{gh}\right) \mid U_T\right]$$

$$= \frac{(N_{Th} - n_h)}{n_h N_{Tgh}(N_{Tgh} - 1)}\left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}} p_j^{-1}\right)^{-2}$$

$$\sum_{k\in U_{Tgh}} p_k^{-2}\left[y_{1k} - \mu_{T,psw}\, y_{2k} - \left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}} p_j^{-1}\right)^{-1}\right.$$

$$\left.\left(N_{Tgh}^{-1}\sum_{j\in U_{Tgh}}\left\{\frac{y_{1j} - \mu_{T,psw}\, y_{2j}}{p_j}\right\}\right)\right]^2 + n^{-2}\,\xi_n. \qquad \text{(A.13)}$$

Since

$$E\; N_{Tgh}^{-1}\sum_{j\in U_{Tgh}} p_j^{-1} = \left(\sum_{k\in U_{gh}} p_k\right)^{-1} N_{gh} + O(N^{-1})$$

as $N \to \infty$. then (A.13) implies the unconditional expectation

$$E\,\mathrm{var}\left(\frac{\sum_{k\in s_{gh}} p_k^{-1}\{y_{1k} - \mu_{T,psw}\, y_{2k}\}}{\sum_{k\in s_{gh}} p_k^{-1}} \mid U_T, n_{gh}\right)$$

$$= \frac{N_{gh}^{-2}\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]}{n_h\left[\left(\sum_{j\in U_{gh}} p_j\right) - 1\right]}$$

$$\sum_{k\in U_{gh}} p_k^{-1}\left[y_{1k} - \mu\, y_{2k} - N_{gh}^{-1}\sum_{j\in U_{gh}}(y_{1j} - \mu\, y_{2j})\right]^2$$

$$+ O\,(n^{-2} + N^{-1}) \qquad \text{(A.14)}$$

as $n \to \infty$. By Theorem A.1,

$$\mathrm{var}\,\hat{\mu}_{psw}$$

$$= \alpha_2^{-2}\, E\,\mathrm{var}\left(\hat{\alpha}_{psw(1)} - \mu_{T,psw}\,\hat{\alpha}_{psw(2)} \mid U_T, n_{gh}, \forall g, h\right)$$

$$+ \alpha_2^{-2}\, E\,\mathrm{var}\left[E(\hat{\alpha}_{psw(1)} - \mu_{T,psw}\,\hat{\alpha}_{psw(2)} \mid\right.$$

$$\left. U_T, n_{gh}, \forall g, h) \mid n_{gh}, \forall g, h\right] + O\,(n^{-2} + N^{-1}) \qquad \text{(A.15)}$$

as $n \to \infty$. Lemma A.1 implies that

$$\operatorname{var}\Big[E(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}}\,\hat{\alpha}_{\text{psw}(2)} \mid U_T, n_{gh}, \forall g,h) \mid n_{gh}, \forall g,h \Big] = n^{-2}\xi_n. \quad \text{(A.16)}$$

Since (4.1) implies that

$$\operatorname{var}(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}}\,\hat{\alpha}_{\text{psw}(2)} \mid U_T, n_{gh}) = N^{-2}\sum_{h=1}^{H} N_h^2 \operatorname{var}\left(N_h^{-1}\sum_{g=1}^{G} N_{gh} \frac{\sum_{k\in s_{gh}} p_k^{-1}\{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k\in s_{gh}} p_k^{-1}} \mid U_{T,n_{gh}} \right),$$

then (A.14), (A.15), and (A.16) imply the result.

Derivation of (3.7), the estimated variance of $\hat{\mu}_{\text{psw}}$: Observe that

$$\frac{N_{Tgh}\sum_{j\in U_h} p_j}{n_{gh}n_h} \approx \frac{N_{Tgh}N_{Th}}{n_{gh}n_h} \approx \left[\frac{N_{gh}}{\sum_{k\in s_{gh}} p_k^{-1}}\right]^2.$$

Noting (4.5) the result follows.

Derivation of (3.8), another estimated variance of $\hat{\mu}_{\text{psw}}$:

Since

$$\frac{N_{gh}}{\sum_{k\in s_{gh}} p_k^{-1}} \approx \frac{N_{Tgh}}{n_{gh}} \approx \frac{N_{Th}}{n_h} \approx \frac{N_h}{\sum_{k\in s_h} p_k^{-1}},$$

the result follows from (3.7).

REFERENCES

BRICK, J.M., WAKSBERG, J. and KEETER, S. (1994). Evaluating the use of data on interruptions in telephone service for nontelephone households. *Proceedings of the Survey Research Methods Section*, American Statistical Association 19-28.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.

KHURSHID, A., and SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

McCULLAGH, P., and NELDER, J.A. (1991). *Generalized Linear Models*. New York: Chapman and Hall.

RAO, J.N.K. (1997). Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics*, 25, 1-21.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

STEEH, C.G., GROVES, R.M., COMMENT, R. and HANSMIRE, E. (1983). Report on the survey research center's surveys of consumer attitudes. *Incomplete Data in Sample Surveys*, (Ed. W.G. Madow, H. Nisselson and I. Olkin), Academic Press, New York, 1.

THORNBERRY, JR., O.T., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Surveys*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg). New York: John Wiley & Sons, Inc., 25-49.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 77-85
Statistics Canada

Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement

YVES TILLÉ[1]

ABSTRACT

The post-stratified estimator sometimes has empty strata. To address this problem, we construct a post-stratified estimator with post-strata sizes set in the sample. The post-strata sizes are then random in the population. The next step is to construct a smoothed estimator by calculating a moving average of the post-stratified estimators. Using this technique it is possible to construct an exact theory of calibration on distribution. The estimator obtained is not only calibrated on distribution, it is linear and completely unbiased. We then compare the calibrated estimator with the regression estimator. Lastly, we propose an approximate variance estimator that we validate using simulations.

KEY WORDS: Unbiased estimation; Calibration on a distribution function; Conditional inclusion probabilities; Weighting.

1. INTRODUCTION

It is possible during a survey by sampling to identify the values of an auxiliary character for all population units. This information may be available when the units are selected in a database containing other variables of interest. The temptation is then to calibrate the results of a survey on this auxiliary information. The decision is made either to retain from this auxiliary variable only certain functions (moments, sizes) for the purpose of using a calibration method (see for example Deville and Särndal 1992 or Estevao, Hidiroglou and Särndal 1995), or this variable can be divided into classes with the view to using a post-stratified estimator.

If the decision is to opt for the post-stratified estimator, deciding on the strata divisions can be delicate. Theoretically, the strata must be defined prior to the selection of the sample. Where should the post-strata boundaries be placed? What size should the post-strata be? This latter question is the most embarrassing because the main problem with post-stratification is the possibility of obtaining empty post-strata. This means that the post-strata have to be large enough so that the probability of obtaining an empty post-stratum is negligible. These problems are not limited to post-stratified estimators. Indeed, it is also possible to have no regression or calibrated estimators for some samples.

Our objective is to define a new method of using auxiliary information in the population. This method is based on the definition of post-strata for which the number of units is set in the sample and not in the population. In this way, it is possible to import into the estimator complex auxiliary information resulting from knowledge of all of the values taken by the auxiliary variable, while avoiding both the problem of defining post-strata borders and the problem of empty post-strata.

This article is organized as follows. In section 2, the notation is defined and in section 3, we describe the principle of rank conditioning, which is used to define the unbiased estimators in section 4. In section 5, the smoothed estimator is defined, and a specific case is examined in detail in section 6. Section 7 contains an application of the estimation of a distribution function. In section 8, this new estimator is compared with the regression estimator and the estimator for a simple design without replacement. Computation of variance is discussed in section 9. As a result of the impossibility of providing an exact solution, an approximation is provided in section 10, which is tested by simulations in section 11. Lastly, general conclusions are presented in section 12.

2. NOTATION

We assume a population composed of N observation units, with the labelling being denoted as $\{1, ..., k, ..., N\}$. In this population, we are interested in a character of interest $Y_k, k \in U$. The objective is to estimate the total $Y = \sum_{k \in U} Y_k$. We select a random sample S of fixed size n by means of a simple random design without replacement. We denote I_k the random indicator variable, which takes the value 1 if the unit k is in the sample and 0 if not. The inclusion probabilities first order are therefore defined by $\Pr(k \in S) = \pi_k = n/N, k \in U$, and the second order inclusion probabilities by $\Pr(k, l \in S) = \pi_{kl} = n(n-1)/(N(N-1)), k \neq l \in U$.

We will be interested in the class of linear estimators of Y, which is written as

$$\hat{Y}_w = \sum_{k \in S} w_k Y_k,$$

[1] Yves Tillé, Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 827, 2002 Neuchâtel, Suisse. E-mail: yves.tille@unine.ch

where the weights w_k may depend on the sample S and therefore be random.

One of the possibilities is to take $w_k = 1/\pi_k = n/N$, which gives the Horvitz-Thompson estimator,

$$\hat{Y}_{HT} = \sum_{k \in S} \frac{Y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} Y_k,$$

which is unbiased.

We will be focussing instead on the more general class of conditionally weighted estimators (Tillé 1998, 1999a) where the units are weighted by inverses of conditional inclusion probabilities. If Z is some statistic, then the conditionally weighted estimator

$$\hat{Y}_Z = \sum_{k \in S} \frac{Y_k}{E(I_k|Z)} \tag{1}$$

is strictly unbiased if and only if $E(I_k|Z) > 0$, for all $k \in U$. In effect,

$$E(\hat{Y}|Z) = \sum_{k \in U} \frac{E(I_k|Z)Y_k}{E(I_k|Z)} = Y.$$

Since the estimator is conditionally unbiased, it is also unconditionally unbiased. Depending on which statistic Z is used, estimator (1) generalizes the stratified estimator as well as (a close approximation) the regression estimator (see Tillé 1998).

3. CONDITIONING ON RANKS

Let us now assume that the N values $X_1, ..., X_k, ..., X_N$ of an auxiliary character x are known for N units of the population. First, we assume that all of the X_k take separate values (this hypothesis will be removed in section 5). The rank R_k of unit k is

$$R_k = \#\{l \in U \,|\, X_l \le X_k\}.$$

Moreover, we denote r_j, $j = 1, ..., n$, the ordered population ranks of the n selected units in the sample, thus $r_1 < r_2 < ... < r_{n-1} < r_n$. The r_j are random variables with a negative hypergeometric distribution (see Tillé 1999b).

The statistic used to define the conditional probabilities of inclusion is a subset of $\{r_1, ..., r_j, ..., r_n\}$. First, we define

- an integer q such that $2 \le q \le n$, defining the period,
- an integer b such that $2 \le b$, defining the border,
- an integer l such that $b \le l \le b + q - 1$, defining the interval.

The quantities q, b, and l serve to define a subset of indices:

$$E_l = \{r_l, r_{l+q}, r_{l+2q}, ..., r_{l+hq}, ..., r_{l+Hq}\},$$

$$\text{for } l = b, ..., b + q - 1.$$

For example, if $n = 18$, $q = 4$, $b = 3$, then

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\},$$

$$E_4 = \{r_4, r_8, r_{12}, r_{16}\},$$

$$E_5 = \{r_5, r_9, r_{13}\},$$

$$E_6 = \{r_6, r_{10}, r_{14}\}.$$

The conditional inclusion probability is computed in relation to one of the E_l.

The value of H is defined in such a way that $l + Hq \le n - b + 1$ and thus H is the largest integer such that $H \le (n - b - l + 1)/q$. It is clear that H depends on l.

The next step is to compute the inclusion probabilities:

$$E(I_k|E_l) = \begin{cases} 1 & \text{if } k \in E_l \\ \dfrac{q-1}{r_{l+hq} - r_{l+(h-1)q} - 1} & \text{if } r_{l+(h-1)q} < k < r_{l+hq},\ h = 1, ..., H \\ \dfrac{l-1}{r_l - 1} & \text{if } k < r_l \\ \dfrac{n-(l+Hq)}{N - r_{l+Hq}} & \text{if } k > r_{l+Hq}. \end{cases}$$

These inclusion probabilities are thus relatively uneven. However, they are all positive, including the borders. It is important to use a border $b \ge 2$ so that the first and the last post-stratum are not empty.

4. CLASS OF UNBIASED ESTIMATORS

Since $E(I_k|E_l) > 0$, we can construct an estimator that is unbiased and even conditionally unbiased with respect to E_l. By denoting $y_1, ..., y_j, ..., y_n$ the n values taken by the units in the sample ordered according to the R_k, we obtain

$$\hat{Y}_l = \sum_{k \in S} \frac{Y_k}{E(I_k|E_l)}$$

$$= \frac{r_l - 1}{l-1} \sum_{j=1}^{l-1} y_j + y_l$$

$$+ \sum_{h=1}^{H} \left(\frac{r_{l+hq} - r_{l+(h-1)q} - 1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j} + y_{l+hq} \right)$$

$$+ \frac{N - r_{l+Hq}}{n - (l+Hq)} \sum_{j=l+Hq+1}^{n} y_j$$

$$= N_{0|l} \hat{\bar{y}}_{0|l} + y_l + \sum_{h=1}^{H} \left(N_{h|l} \hat{\bar{y}}_{h|l} + y_{l+hq} \right) + N_{H+1|l} \hat{\bar{y}}_{H+1|l}$$

where

$$N_{0|l} = r_l - 1,$$

$$N_{h|l} = r_{l+hq} - r_{l+(h-1)q} - 1, h = 1, ..., H,$$

$$N_{H+1|l} = N - r_{l+Hq},$$

$$\hat{\bar{y}}_{0|l} = \frac{1}{l-1} \sum_{j=1}^{l-1} y_j,$$

$$\hat{\bar{y}}_{h|l} = \frac{1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j}, h = 1, ..., H,$$

and

$$\hat{\bar{y}}_{H+1|l} = \frac{1}{n-(l+Hq)} \sum_{j=l+Hq+1}^{n} y_j.$$

This estimator is in reality a post-stratified estimator where the sizes of the post-strata are set in the sample. Since $E(I_k|E_l) > 0$, $\hat{Y}_l$ is strictly unbiased unconditionally and conditionally to E_l, which is clearly not the case for the traditional post-stratified estimator, because the latter has a non-zero probability of having an empty post-stratum. By setting the size of the post-strata in the sample, creating empty post-strata becomes impossible. The corresponding size of the post-stratum in the population is a random variable $N_{h|l}$.

The estimator $\hat{Y}_l$ has another interesting property. By using the definition of the $E(I_k|E_l)$, we can quite easily show that

$$\sum_{k \in S} \frac{1}{E(I_k|E_l)} = N.$$

The estimator is thus calibrated on the size of the population. This property, which is also held by the Horvitz-Thompson estimator in simple designs, is therefore not lost. Units where the ranks are in E_l are called pivot units, and are assigned a weight equal to 1, which makes the weights very unequal. A downside to $\hat{Y}_l$ is the use of widely dispersed weights. This problem can be resolved by smoothing the estimators.

5. SMOOTHING ESTIMATORS

To resolve the problem of the dispersion of the weights, we compute a moving average for the estimators as follows:

$$\hat{Y}_c = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{Y}_l.$$

$\hat{Y}_c$ retains all of the properties of the $\hat{Y}_l$. This means that it is unbiased, calibrated on N and linear and can therefore be written as

$$\hat{Y}_c = \sum_{j=1}^{n} w_j y_j,$$

where $w_j =$

$$\begin{cases} \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{r_l - 1}{l-1}, & j < b, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{m^-(j+l-b-q)} - 1}{j+l-b-m^-(j+l-b-q)-1} + 1 \right), & b \le j < b+q-1, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right), & b+q-1 \le j \le n-b+2-q, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{m^+(j+l-b)} - r_{j+l-b-q} - 1}{m^+(j+l-b)-j+l-b-q-1} + 1 \right), & n-b+2-q < j \le n-b+1, \\ \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N+1-r_{n+1-l}-1}{n+1-(n+1-l)-1} = \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N-r_{n+1-l}}{l-1}, & n-b+1 < j, \end{cases}$$

$$m^-(x) = \begin{cases} 0 & \text{if } x < b \\ x & \text{if not} \end{cases},$$

$$m^+(x) = \begin{cases} n+1 & \text{if } x > n-b+1 \\ x & \text{if not} \end{cases} \quad (2)$$

$r_0 = 0$, and $r_{n+1} = N + 1$.

Under the apparent complexity arising from the specific treatment of the borders, the weighting system is relatively simple. In the case where we are not too close to the borders, it takes the value

$$w_j = \frac{1}{q}\left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right)$$
$$= \frac{1}{q\,(q-1)} \sum_{\alpha=0}^{q-1} (r_{j+\alpha} - r_{j+\alpha-q}).$$

If all of the values of the auxiliary variable are not distinct, we can assign the unit ranks with common values randomly. For example, if $X_1 = 2, X_2 = 5, X_3 = 5, X_3 = 5, X_4 = 7, X_5 = 8$, we select with a probability ½, between, ranks $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5$, or $R_1 = 1, R_2 = 3, R_3 = 2, R_4 = 4, R_5 = 5$. We then compute the smoothed estimator for each permutation, and we calculate their mean. The advantage of this method is that it preserves an unbiased estimator. In effect, for each possible permutation, the estimator is unbiased. In practice, it is not necessary to compute estimators for all of the permutations. We can calculate the estimator for one permutation and then simply equalize the weights of the units having the same values for the variable x.

6. CASE WHERE $q = 2, b = 2$

When $q = 2$, and $b = 2$, we obtain after a few calculations

$$\hat{Y}_c = \frac{1}{2}\left\{ \sum_{j=3}^{n-2} y_j (r_{j+1} - r_{j-1}) + \frac{r_3 + 2r_2 - 3}{2} y_1 + \frac{r_3 + 1}{2} y_2 + \frac{r_{n+1} - r_{n-2} + 1}{2} y_{n-1} + \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n \right\}$$
$$= \frac{1}{2}\left\{ \sum_{j=1}^{n} y_j (r_{j+1} - r_{j-1}) + y_1 \frac{r_3 - 3}{2} + y_2 \frac{2r_1 + 1 - r_3}{2} + y_{n-1} \frac{r_{n+1} + r_{n-2} + 1 - 2r_n}{2} + y_n \frac{r_{n+1} - r_{n-2} - 3}{2} \right\},$$

where $r_0 = 0$ and $r_{n+1} = N + 1$. This brings us to an estimator proposed by Ren (2000, page 140) and obtained using a calibration argument. The way in which the borders are managed is the only slight difference.

Example 1: With a population of size $N = 20$. Let us assume that the values of the variable of interest are found in Table 1. We also assume that the sample of size $n = 7$ is composed of the units with ranks {3, 7, 8, 11, 12, 15, 17}. If we take $q = 2$, $l = 2$, $b = 2$ we obtain $E_2 = \{r_2, r_4, r_6\} = \{7, 11, 15\}$. We can then calculate $E(I_k \mid E_2 = \{7, 11, 15\})$. The conditional inclusion probabilities are as follows:

$$E(I_3 \mid E_2 = \{7, 11, 15\}) = 1/6,$$
$$E(I_7 \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_8 \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{11} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{12} \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{15} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{17} \mid E_2 = \{7, 11, 15\}) = 1/5.$$

Table 1
Example of a Population of Size $N = 20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	71	72	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	14	8	17	9	16	12	19	4	5	11	18	10

The estimator

$$\hat{Y}_0 = \sum \frac{y_k}{E(I_k \mid E_2 = \{7,\ 11,\ 15\})}$$

is therefore unbiased and conditionally unbiased. Further, it is linear and

$$\sum_{k \in S} \frac{1}{E(I_k \mid E_2 = \{7, 11, 15\})} = N.$$

However, if we take $q = 2$, $l = 3$, $b = 2$, we obtain $E_3 = \{r_3, r_5\} = \{8, 12\}$. Using the same example, we then compute $E(I_k \mid E_3 = \{8, 12\})$, and we obtain

$$E(I_3 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_7 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_8 \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{11} \mid E_3 = \{8, 12\}) = 1/3,$$
$$E(I_{12} \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{15} \mid E_3 = \{8, 12\}) = 2/8 = 1/4,$$
$$E(I_{17} \mid E_3 = \{8, 12\}) = 2/8 = 1/4.$$

The estimator

$$\hat{Y}_1 = \sum \frac{y_k}{E(I_k \mid E_3 = \{8,\ 12\})}$$

is also unbiased and linear.

Lastly, we compute the mean of the two estimators:

$$\hat{Y}_c = \frac{\hat{Y}_0 + \hat{Y}_1}{2}.$$

The weights are obtained simply by calculating the mean of the weights of estimators $\hat{Y}_0$ and $\hat{Y}_1$, and have the values

$$w_3 = (6 + 7/2)/2 = 19/4,$$
$$w_7 = (1 + 7/2)/2 = 9/4,$$
$$w_8 = (3 + 1)/2 = 2,$$
$$w_{11} = (1 + 3)/2 = 2,$$
$$w_{12} = (3 + 1)/2 = 2,$$
$$w_{15} = (1 + 4)/2 = 5/2,$$
$$w_{17} = (5 + 4)/2 = 9/2.$$

Estimator $\hat{Y}_c$ is linear and strictly unbiased.

7. APPLICATION TO THE ESTIMATION OF THE DISTRIBUTION

There are several ways to appropriately use auxiliary information to estimate a distribution function. A description of these techniques can be found in Ren (2000) and in Wu and Sitter (2001). The method that we suggest also makes it possible to estimate the distribution. The distribution in the population is defined by

$$F_1(y) = \frac{1}{N}\sum_{k\in U} I\{y_k \le y\},$$

and can be estimated by

$$\hat{F}_1(y) = \frac{\sum_{k\in S} w_k I\{y_k \le y\}}{\sum_{k\in S} w_k},$$

where $I\{y \le y_k\}$ is the indicator function, and the w_k are the weights allocated to the units k which have the value $1/\pi_k = N/n$ for the Horvitz-Thompson estimator, and which are given in (2) for the calibrated estimator.

Note that the two functions are discrete, but that there are far fewer jumps in S than in U. To offset the differences in the distributions between the sample and the population, we have smoothed the distribution functions by using, as Deville (1995) did, a linear interpolation of the centres of the risers, which involves defining $F_2(y)$ by linking the points

$$\frac{1}{2}\left\{F_1(y_k) - F_1(y_k - \varepsilon)\right\},$$

for $k \in U$, where ε is a strictly positive, arbitrarily small real number. We then define $\hat{F}_2(y)$ by linking the points

$$\frac{1}{2}\left\{\hat{F}_1(y_k) - \hat{F}_1(y_k - \varepsilon)\right\},$$

for the sample.

Example 2: A population of size $N = 1\ 000$ was generated using independent log-normal variables that are equally distributed. A sample of size $n = 16$ was then selected and we set $h = 5$. Figure 1 gives $F_2(x)$ in the population.

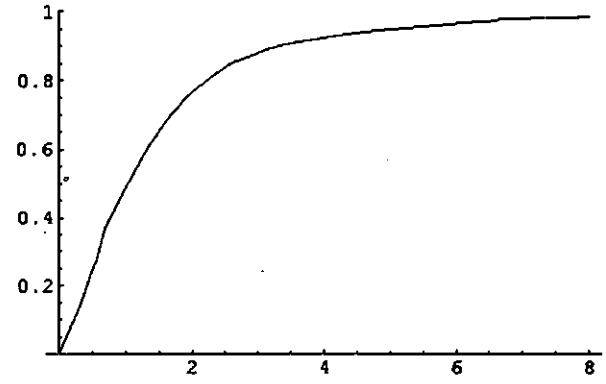


Figure 1. Population distribution function

Figure 2 shows $F_2(x)$ and the distribution estimated by the Horvitz-Thompson estimator. Lastly, Figure 3 shows $F_2(x)$ and the distribution estimated by the calibrated estimator.

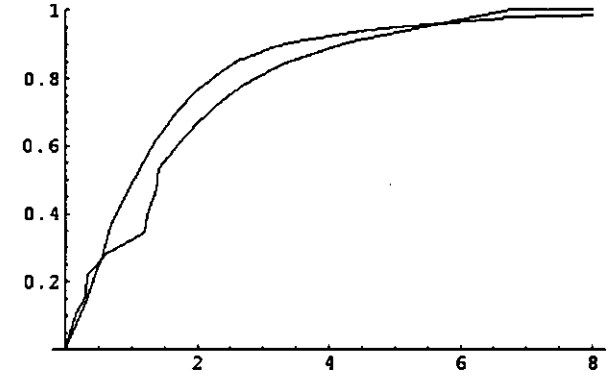


Figure 2. Population distribution function and Horvitz-Thomson distribution estimator

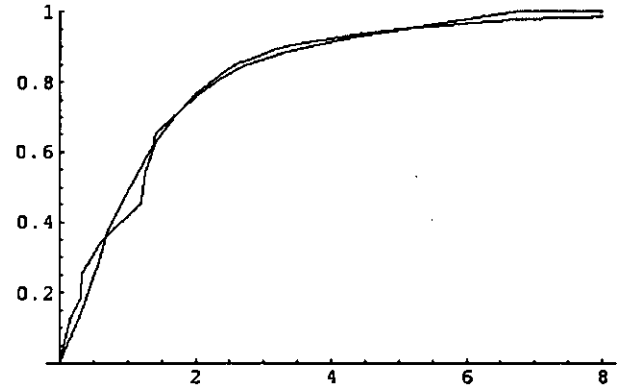


Figure 3. Population distribution function and calibrated distribution estimator

8. COMPARISON WITH THE REGRESSION ESTIMATOR

In order to compare the qualities of the proposed estimator, a series of simulations was conducted to compare the estimator calibrated on distribution with the Horvitz-Thompson estimator and the regression estimator. Three populations of size 1,000 were generated by means of the following models.

- *Model A Linear dependence*: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = X_k + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.616154$.
- *Model B Non-linear dependence* 1: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = (0.2 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.28975$.
- *Model C Non-linear dependence* 2: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = (0.4 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.476158$.

In each population, 100,000 samples of size 100 were selected. Three weighting systems were computed for each sample.

1. the weights associated with the simple design $w_k = N/n$,
2. the weights of the distribution calibrated estimator given in (2) using the window $q = 10$ and border $b = 6$,
3. the weights of the regression estimator given by

$$w_k = \frac{N}{n} + \left(X - \hat{X}_{\mathrm{HT}}\right) \frac{\left(X_k - \hat{\bar{X}}\right)}{\sum_{k \in S} \left(X_k - \hat{\bar{X}}\right)^2},$$

where X is the total of the X_k in the population, $\hat{X}_{\mathrm{HT}}$ is the Horvitz-Thompson estimator of X, and $\hat{\bar{X}} = \hat{X}_{\mathrm{HT}}/N$.

Using these weights, the estimator of the mean and of the nine deciles were calculated for each sample. We then estimate the variance of these estimators by means of the simulations.

The results are given in Tables 2, 3 and 4. The variances were brought to 1 for the simple design. For the linear model, the regression estimator is slightly preferable. However, in the non-linear case, the distribution calibrated estimator significantly increases the precision on the mean and on the quantiles. This means that our proposed estimator is robust when there is a non-linear relationship between the auxiliary variable and the variable of interest.

Table 2
Model A: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.674422	0.632608
1st decile	0.905273	0.893876
2nd decile	0.815403	0.802082
3rd decile	0.842681	0.815071
4th decile	0.806749	0.768283
5th decile	0.783731	0.740765
6th decile	0.818051	0.782549
7th decile	0.794411	0.773794
8th decile	0.857114	0.844874
9th decile	0.884424	0.884032

Table 3
Model B: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.429689	0.953025
1st decile	0.913598	0.958656
2nd decile	0.919394	1.009270
3rd decile	0.829860	0.987950
4th decile	0.792094	0.989114
5th decile	0.703908	0.992023
6th decile	0.622705	1.009830
7th decile	0.550028	0.981249
8th decile	0.443828	1.010340
9th decile	0.549615	1.029120

Table 4
Model C: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.30768	0.808114
1st decile	0.95560	0.983582
2nd decile	0.85920	0.970913
3rd decile	0.73854	0.930401
4th decile	0.65728	0.950651
5th decile	0.60500	0.956807
6th decile	0.52139	0.930514
7th decile	0.45709	0.907537
8th decile	0.40752	0.903593
9th decile	0.39820	0.860050

9. VARIANCE AND ESTIMATION OF VARIANCE

To compute the variance of $\hat{Y}_c$, we begin by computing the variance of $\hat{Y}_l$. Since $\hat{Y}_l$ is unbiased conditionally to E_l, we have

$$V(\hat{Y}_l) = E\,V(\hat{Y}_l \mid E_l).$$

As with each of the post-strata, conditionally to E_l the design is a fixed-size simple sampling without replacement, we have

$$V(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2\, V(\hat{\bar{y}}_{h|l})$$

$$= \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{n|l}} \frac{S_{h|l}^2}{n_{h|l}}, \qquad (3)$$

where

$$n_{0|l} = l - 1,$$

$$n_{h|l} = q - 1, h = 1, ..., H,$$

$$n_{H+1|l} = n - (l + Hq),$$

$$\bar{Y}_{0|l} = \frac{1}{N_{0|l}} \sum_{k=1}^{r_{l-1}} Y_{(k)},$$

$$\bar{Y}_{h|l} = \frac{1}{N_{h|l}} \sum_{k=r_{l+(h-1)q}+1}^{r_{l+hq-1}} Y_{(k)}, h = 1, ..., H,$$

$$\bar{Y}_{H+1|l} = \frac{1}{N_{H|l}} \sum_{k=N-r_{l+Hq}+1}^{N} Y_{(k)},$$

$$S_{0|l}^2 = \frac{1}{N_{0|l} - 1} \sum_{k=1}^{r_{l-1}} (Y_{(k)} - \bar{Y}_{0|l})^2,$$

$$S_{h|l}^2 = \frac{1}{N_{h|l} - 1} \sum_{k=r_l+(h-1)q+1}^{r_l+h_q-1} (Y_{(k)} - \bar{Y}_{h|l})^2, h = 1, ..., H,$$

and

$$S_{H+1|l}^2 = \frac{1}{N_{H+1|l} - 1} \sum_{k=N-r_{l+Hq}+1}^{N} (Y_{(k)} - \bar{Y}_{H+1|0})^2,$$

where the $Y_{(k)}$ represent the values of Y_k sorted by increasing order of the X_k.

Note that it is very difficult to calculate the unconditional variance of $\hat{Y}_l$, that is, the expectation of $V(\hat{Y}_l|E_l)$. In effect, $N_{h|l}$ and $S_{h|l}^2$ are random. However, we can estimate $V(\hat{Y}_l|E_l)$ simply and obtain an unbiased estimator of the conditional variance (and thus of the variance) by simply estimating (3), by

$$\hat{V}(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{h|l}\, n_{h|l}} s_{h|l}^2, \qquad (4)$$

where

$$s_{0|l}^2 = \frac{1}{n_{0|l} - 1} \sum_{j=1}^{l-1} (y_j - \hat{\bar{y}}_{0|l})^2,$$

$$s_{h|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=1}^{q-1} (y_{l+(h-1)q+j} - \hat{\bar{y}}_{h|l})^2, h = 1, ..., H,$$

and

$$s_{H+1|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=l+Hq+1}^{n} (y_j - \hat{\bar{y}}_{H+1|l})^2.$$

The estimator $\hat{V}(\hat{Y}_l|E_l)$ is not only unbiased for $V(\hat{Y}_l|E_l)$ but also for $V(\hat{Y}_l)$.

10. APPROXIMATIONS FOR COMPUTING THE VARIANCE

Unfortunately, computing the variance of $\hat{Y}_c$ becomes more complex because of covariances. In effect, we have

$$V(\hat{Y}_c) = \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \mathrm{Cov}(\hat{Y}_l, \hat{Y}_i).$$

When $l = i$, the problem is to estimate $V(\hat{Y}_i)$, which is done easily. When $l \neq i$, it is necessary to compute

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i|E_l)$$

$$+ \mathrm{Cov}(E(\hat{Y}_l|E_l), E(\hat{Y}_i|E_l)).$$

Since $E(\hat{Y}_l|E_l) = Y$, we obtain

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i|E_l)$$

$$= E\,E(\hat{Y}_l \hat{Y}_i|E_l) - Y^2.$$

Unfortunately, it does not appear possible to extricate the computation of $E(\hat{Y}_l, \hat{Y}_i|E_l)$ and therefore we must find an approximation.

One way is to find a value that is greater than the variance. Since

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) \leq \sqrt{V(\hat{Y}_l)\,V(\hat{Y}_i)},$$

we have a greater value given by

$$V(\hat{Y}_c) \leq \frac{1}{q^2}\sum_{l=b}^{b+q-1}\sum_{i=b}^{b+q-1}\sqrt{V(\hat{Y}_l)V(\hat{Y}_i)}$$

$$= \frac{1}{q^2}\left(\sum_{l=b}^{b+q-1}\sqrt{V(\hat{Y}_l)}\right)^2,$$

which can be estimated by

$$\hat{V}_1(\hat{Y}_c) = \frac{1}{q^2}\left(\sum_{l=b}^{b+q-1}\sqrt{\hat{V}(\hat{Y}_l \mid E_l)}\right)^2,$$

which comes back to estimating the standard deviation of the means by the mean of the standard deviations.

Lastly, a second possibility involves using a residuals technique. Generally, when an estimator is corrected using a calibration technique, the variance is estimated by means of a residuals technique (see Deville and Särndal 1992 and Deville 1999 on this topic). When computing the variance of $\hat{Y}_l$, it is possible to use a residuals technique to obtain the exact variance. Consider the variable

$$v_k(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)}\right)^{-1/2}\left(\dfrac{N_{h|l}^2(N_{h|l}-n_{h|l})}{N_{h|l}n_{h|l}(N_{h|l}-1)}\right)^{1/2}(Y_k - \bar{Y}_{h|l}) \\ \qquad \text{if } k = r_{l+(h-1)q+1}, \ldots, r_{l+hq-1} \\ 0 \qquad \text{if } k = r_{l+(h-1)q} \text{ or } k = r_{l+hq} \end{cases}$$

which can appear as a residual associated with the estimator $\hat{Y}_l$. The variable $v_k(l)$ inserted in the traditional expression of the fixed-size simple sampling design without replacement is exactly equal to the conditional variance $\hat{Y}_l$ given in (3). In effect,

$$N^2\frac{N-n}{nN}\frac{1}{N-1}\sum_{k\in U}\left(v_k - \frac{\sum_{k\in U} v_k}{N}\right)^2 = V(\hat{Y}_l|E_l).$$

This variable, however, depends on the $\bar{Y}_{h|l}$ which are unknown. We can estimate $v_k(l)$ by

$$\hat{v}_j(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)}\right)^{-1/2}\left(\dfrac{N_{h|l}^2(N_{h|l}-n_{h|l})}{N_{h|l}n_{h|l}(n_{h|l}-1)}\right)^{1/2}(y_j - \hat{\bar{y}}_{h|l}) \\ \qquad \text{if } j = l+(h-1)q+1, \ldots, l+hq-1 \\ 0 \qquad \text{if } j = l+(h-1)q \text{ or } j = l+hq \end{cases}$$

If we insert $\hat{v}_k(l)$ in the estimator of the variance for the simple design without replacement, we obtain an unbiased estimator of the conditional variance, and therefore of the variance.

$$N^2\frac{N-n}{nN}\frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{v}_j - \frac{\sum_{j=1}^{n}\hat{v}_j}{n}\right)^2 = \hat{V}(\hat{Y}_l|E_l).$$

Deville (1999) shows that the variance of a sum of estimators can be determined by adding the residuals associated with these estimators, the residuals having been computed by linearization. To estimate the variance of $\hat{Y}_c$, we could therefore simply take the mean of the residuals $\hat{v}_k(l)$, which is written

$$\hat{v}_k = \frac{1}{q}\sum_{l=b}^{b+q-1}\hat{v}_k(l).$$

In this way, it would be possible to estimate the variance by

$$\hat{V}_2(\hat{Y}_c) = \frac{N^2(N-n)}{nN}\frac{1}{n-1}\sum_{k\in S}\left(\hat{v}_k - \frac{\sum_{k\in S}\hat{v}_k}{n}\right)^2.$$

These two variance estimators need to be tested by simulations.

11. SIMULATIONS FOR VARIANCE ESTIMATORS

The simulations shown in Table (5) are based on populations of size $N = 100$, that are generated by means of normal independent random variables. For each case studied, we give the value of q and the coefficient of correlation between the variable of interest Y_k and the rank R_k of the auxiliary variable X_k. The border b is defined by taking the integer of $q/2+1$. Since our purpose is to validate the variance estimator, we use 3,000 samples of size $n = 20$ for each simulation and we compare the variance estimated by the simulations of the calibrated estimator $V_{si}(\hat{Y}_c)$ with the expectations under the simulations of the two variance estimators denoted $E_{si}(\hat{V}_\alpha(\hat{Y}_c))$, $\alpha = 1, 2$. The last two columns of the tables show the relative bias defined by

$$\mathrm{RB}_{si}\hat{V}_\alpha(\hat{Y}_c) = \frac{E_{si}\hat{V}_\alpha(\hat{Y}_c) - V_{si}(\hat{Y}_c)}{V_{si}(\hat{Y}_c)}, \quad \alpha = 1, 2.$$

The simulations show that the two proposed estimators overestimate the variance. The overestimation appears to diminish as q increases. The estimator $\hat{V}_2(\hat{Y}_c)$ definitely has the smallest bias. We will therefore prefer to use $\hat{V}_2(\hat{Y}_c)$.

Table 5
Simulation Results

	Correlation:	0.802			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.045	0.065	0.054	0.444	0.200
5	0.045	0.066	0.057	0.467	0.267
6	0.056	0.076	0.070	0.357	0.250
7	0.058	0.079	0.059	0.362	0.017
8	0.063	0.088	0.087	0.397	0.381
	Correlation:	0.481			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.048	0.066	0.059	0.375	0.229
5	0.045	0.060	0.054	0.333	0.200
6	0.044	0.056	0.051	0.273	0.159
7	0.044	0.,054	0.051	0.227	0.159
8	0.045	0.052	0.048	0.156	0.067
	Correlation:	0.111			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.281	0.471	0.363	0.676	0.292
5	0.297	0.420	0.356	0.414	0.199
6	0.279	0.363	0.316	0.301	0.133
7	0.267	0.342	0.324	0.281	0.213
8	0.282	0.327	0.281	0.160	-0.004

12. CONCLUSIONS

Our proposed estimator is one of the rare estimators that is both unbiased and linear, that uses auxiliary information and that is calibrated on the size of the population. It can be parameterized using the width of window q. This new estimator is robust compared with the regression estimator. It can take into account auxiliary information with a non-linear relationship with the variable of interest. Most simulations appear to show that the width of the window does not significantly impact the preciseness of the mean estimator. However, it also appears that a small window width is not penalizing, even if there is no dependence between the auxiliary variable and the variable of interest. The smaller q is, the tighter the calibration, and the variance estimator will then be significantly penalized because the degree of freedom is lost in each post-stratum. The choice of q must therefore reflect this problem.

There are many other methods that allow for the use of the information given by a distribution function (see Ren 2000) to improve an estimator. The results that we have presented are limited to simple sampling designs, but we think they are important just as post-stratification is important as a specific case of calibration techniques. Post-stratification is one of the few examples where it is possible to show with accuracy that calibration corresponds to a conditional approach. Further, our approach can be seen as a calibration on a distribution function providing an unbiased estimator. A good general distribution calibration technique should therefore include in simple sampling designs the method we have presented.

ACKNOWLEDGEMENTS

We would like to thank Jean-Claude Deville and Anne-Catherine Favre, two referees and an associate editor for their constructive comments, which considerably improved this article.

REFERENCES

DEVILLE, J.-C. (1995). *Estimation de la variance du coefficient de Gini mesuré par sondage*. INSEE Méthode, working paper, Methodology F9510.

DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-203.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ESTEVAO, V., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principle for a generalized estimation system in Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

REN, R. (2000). *Estimation par calage sur la répartition*. Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.

TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities : simple random sampling. *International Statistical Review*, 66, 303-322.

TILLÉ, Y. (1999a). Sur la détermination a posteriori des bornes des post-strates. In *Les Sondages* (Eds. G. Brossier and A.-M. Dussaix). Dunod, 202-208.

TILLÉ, Y. (1999b). Estimation in surveys using conditional inclusion probabilities: complex design. *Survey Methodoloogy*, 25, 57-66.

WU, C., and SITTER, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289-307.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 87-95
Statistics Canada

Variance Estimation for the Current Employment Survey

JUN SHAO and SHAIL BUTANI[1]

ABSTRACT

Like most other surveys, nonresponse often occurs in the Current Employment Survey conducted monthly by the U.S. Bureau of Labor Statistics (BLS). In a given month, imputation using reported data from previous months generally provides more efficient survey estimators than ignoring nonrespondents and adjusting survey weights. However, imputation also has an effect on variance estimation: treating imputed values as reported data and applying a standard variance estimation method leads to negatively biased variance estimators. In this article we propose some variance estimators using the grouped balanced half sample method and re-imputation to take imputation into account. Some simulation results for the finite sample performance of the imputed survey estimators and their variance estimators are presented.

KEY WORDS: Balanced half samples; Non-negligible sampling fractions; Ratio imputation; Stratified sampling.

1. INTRODUCTION

The Current Employment Survey (CES), commonly known as the payroll survey, is conducted monthly by the U.S. Bureau of Labor Statistics (BLS). The data are obtained from establishments on a monthly basis by various automated methods including computer assisted telephone interviews, touchtone data entry, FAX, electronic data interchange, mail, *etc.* The main variables are the employment, production or non-supervisory workers and their working hours and earnings on nonagricultural establishment payrolls. Population employment counts are obtained once a year from Unemployment Insurance administrative records.

Nonresponse often occurs in the CES. In any particular month, imputation using reported data from previous months generally provides more efficient survey estimators than using reported data in the current month only and adjusting survey weights. This is particularly true in the CES because the nonresponse rate is about 60-80% and about 60% of the nonrespondents in a given month may become available one or several months later so that these data can be used as "reported data from previous months" (historical data) in a future month.

However, it is well known that treating imputed values as reported data and applying a standard variance estimation method leads to biased (often negatively biased) variance estimators. Valid variance estimators can be derived under some assumptions on sampling designs, imputation methods, and response mechanisms (and, sometimes, models that generate data).

The purposes of this article is to study variance estimation for the CES. After describing the sampling design and the imputation procedure currently used for the CES in section 2, we derive valid (asymptotically unbiased and consistent) variance estimators for imputed survey estimators in section 3. To simplify the computation of variation estimators, we propose some approximations in section 4 and study their properties by simulation in section 5. Some conclusions are made in section 6. Although our study is motivated by the CES, we believe that our results are general and applicable to any survey that adopts a similar sampling design and a similar imputation method.

2. SAMPLING DESIGN AND IMPUTATION

The CES adopts the following stratified probability sampling design. Let P be a finite population containing a set of establishments $\{1, ..., N\}$, which is stratified by the type of industry and by the size of the establishment. Within the hth stratum, a sample of size $n_h \geq 2$ is taken without replacement from N_h population units, using probability sampling independently across strata. The sampling fractions n_h/N_h are not necessarily negligible; for some strata with large establishment sizes, $n_h = N_h$. Let S denote the sample. For $i \in S$, at month $t = 0, 1, ..., T$, values on the number of employees ($y_{t,i}^E$), the number of non-supervisory workers ($y_{t,i}^W$), the number of hours worked ($y_{t,i}^H$), and the weekly pay ($y_{t,i}^P$) are observed (if there is no nonresponse). Let $y_{t,i}$ denote any of $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$, or $y_{t,i}^P$. In CES, the main parameters of interest are population totals $Y_t = \sum_{i \in P} y_{t,i}$, $t = 1, ..., T$. Since population totals can be obtained once a year from administrative records, we assume without loss of generality that Y_0 is known. If there is no nonresponse, Y_t is estimated by a ratio estimator

$$\hat{Y}_t = Y_0 \sum_{i \in S} w_i y_{t,i} \Big/ \sum_{i \in S} w_i y_{0,i}, \quad t = 1, ..., T, \qquad (1)$$

where w_i is the survey weight for the ith unit in the sample and the hth stratum.

[1] Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706; Shail Butani, Statistical Methods Division, The Bureau of Labor Statistics, Washington, D.C. 20212.

In our research, starting from month 1, nonrespondents are imputed using the imputation method proposed in Butani, Harter and Wolter (1997), as described below. Imputation is carried out within an imputation cell, which is the same as stratum or a union of strata. Imputed values in months $1, ..., t-1$ are carried over to impute nonrespondents in month t, unless nonrespondents in months $1, ..., t-1$ become respondents prior to month t.

1. The number of employees. If $y_{t,i}^E$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^E = \hat{\alpha}_t \tilde{y}_{t-1,i}^E,$$

where $\tilde{y}_{t-1,i}^E = y_{t-1,i}^E$ (reported value) if $y_{t-1,i}^E$ is available at month t and otherwise $\tilde{y}_{t-1,i}^E$ is an imputed value,

$$\hat{\alpha}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^E}{\sum_{j \in R_t} w_j y_{t-1,j}^E},$$

and R_t is the set of all reporting units for months t and $t-1$.

2. The number of non-supervisory workers. If $y_{t,i}^W$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^W = \tilde{y}_{t-1,i}^W \tilde{y}_{t,i}^E / \tilde{y}_{t-1,i}^E,$$

where $\tilde{y}_{t-1,i}^W$ is defined similarly to $\tilde{y}_{t-1,i}^E$.

3. The number of hours worked. If $y_{t,i}^H$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^H = \hat{\gamma}_t \tilde{y}_{t-1,i}^H \tilde{y}_{t,i}^W / \tilde{y}_{t-1,i}^W,$$

where $\tilde{y}_{t-1,i}^H$ is defined similarly to $\tilde{y}_{t-1,i}^E$ and

$$\hat{\gamma}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^H / \sum_{j \in R_t} w_j y_{t,j}^W}{\sum_{j \in R_t} w_j y_{t-1,j}^H / \sum_{j \in R_t} w_j y_{t-1,j}^W}.$$

4. The weekly pay. If $y_{t,i}^P$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^P = \hat{\beta}_t \tilde{y}_{t-1,i}^P \tilde{y}_{t,i}^H / \tilde{y}_{t-1,i}^H,$$

where $\tilde{y}_{t-1,i}^P$ is defined similarly to $\tilde{y}_{t-1,i}^E$ and

$$\hat{\beta}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^P / \sum_{j \in R_t} w_j y_{t,j}^H}{\sum_{j \in R_t} w_j y_{t-1,j}^P / \sum_{j \in R_t} w_j y_{t-1,j}^H}.$$

Once nonrespondents are imputed, estimated monthly totals are calculated according to (1) by treating imputed values as reported data.

Assume that the population P is divided into K disjoint imputation cells $P_1, ..., P_K$ and for each k,

$$y_{t,i} = \alpha_{t,k} y_{t-1,i} + \sqrt{y_{t-1,i}}\, e_{t,i},$$

$$E_m(y_{t,i}) = \mu_{t,k}, \qquad E_m(e_{t,i}) = 0, \quad i \in P_k, \ t = 1, 2, ...,$$

$$V_m(y_{t,i}) = v_{t,k}, \qquad V_m(e_{t,i}) = \sigma_k^2, \tag{2}$$

where $y_{t,i}$ denotes any of $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$, or $y_{t,i}^P$, E_m and V_m are the model (marginal) expectation and variance, respectively, $\alpha_{t,k}$ and σ_k^2 are unknown parameters, $e_{t,i}$'s are iid and the two processes $\{y_{t,i}\}$ and $\{e_{t,i}\}$ are independent. Within each P_k, it is assumed that the response indicator $a_{h,i}$ (=1 if $y_{t,i}$ is a respondent and = 0 otherwise) and $y_{t,i}$ are independent, given $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, ..., t$. Under this response mechanism, which is called unconfounded response mechanism (Lee, Rancourt and Särndal 1994), $a_{t,i}$ and $y_{t,i}$ are dependent, but through $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, ..., t$. It is more general than the assumption that $(y_{1,i}, ..., y_{t,i})$ and $(a_{1,i}, ..., a_{t,i})$ are independent. Finally, response indicators from different units are assumed to be independent. Under these assumptions, the estimators $\hat{Y}_t$ based on imputed data as described in the previous section are asymptotically unbiased with respect to the joint expectation under model (2) and sampling from the finite population.

In the CES, the imputation cells are unions of strata so that

$$\sum_{i \in S \cap P_k} w_i = M_k, \quad k = 1, ..., K,$$

where M_k is the number of population units in the kth imputation cell P_k. Consequently, the $\hat{Y}_t$ are conditionally unbiased with respect to the model expectation (given S), *i.e.*,

$$E_m(\hat{Y}_t - Y_t) = 0.$$

3. VARIANCE ESTIMATION

Let E_s and V_s be the sampling expectation and variance, respectively, and V be the overall variance. Then

$$V(\hat{Y}_t - Y_t) = E_s[V_m(\hat{Y}_t - Y_t)] + V_s[E_m(\hat{Y}_t - Y_t)]$$

$$= E_s[V_m(\hat{Y}_t - Y_t)], \tag{3}$$

since $E_m(\hat{Y}_t - Y_t) = 0$. Furthermore, it is shown in the Appendix that

$$V_m(\hat{Y}_t - Y_t) = V_m(\hat{Y}_t) - V_m(Y_t). \tag{4}$$

Note that (4) is obvious in the case of no nonresponse.

Because of (3) the estimation of $V(\hat{Y}_t - Y_t)$ is the same as the estimation of $V_m(\hat{Y}_t - Y_t)$. Also, because of (4), we can first derive estimators v_{t1} and v_{t2} of $V_m(\hat{Y}_t)$ and $V_m(Y_t)$, respectively, and then take the difference $v_{t1} - v_{t2}$ as our variance estimator for $\hat{Y}_t$. Since $V_m(\hat{Y}_t)$ is a conditional variance, given S, we do not need to consider the sampling fractions n_h/N_h in the estimation of $V_m(\hat{Y}_t)$.

We first consider the estimation of $V_m(\hat{Y}_t)$. If an approximate formula of $V_m(\hat{Y}_t)$ can be derived, then we can directly estimate $V_m(\hat{Y}_t)$ by substitution. The explicit form of $\hat{Y}_t$, however, is very complex when t is not small so that the derivation of $V_m(\hat{Y}_t)$ is very difficult. Thus, in the CES we adopt a grouped half sample method that incorporates Rao and Shao's (1992) adjustment (or re-imputation) to take imputation into account. Specifically, sampled units in each stratum are randomly grouped into two groups. R half samples are created using a Hadamard matrix, where $H+1 \le R \le H+4$ and H is the number of strata. For the rth half sample and the ith sampled unit, define

$$w_i^{(r)} = \begin{cases} (1+0.5)\,w_i & \text{if the unit is in the } r\text{th half sample} \\ (1-0.5)\,w_i & \text{if the unit is not in the } r\text{th half sample,} \end{cases}$$

where w_i is the original survey weight. Let $\hat{Y}_t^{(r)}$ be the same as $\hat{Y}_t$ except that the weights w_i are replaced by the $w_i^{(r)}$, including the weights used in imputation (*i.e.*, $\hat{\alpha}_t$, $\hat{\gamma}_t$, and $\hat{\beta}_t$ are re-computed for every r, which is equivalent to Rao and Shao's adjustment). A grouped half sample variance estimator of $V_m(\hat{Y}_t)$ is

$$v_{t1} = \frac{4}{R}\sum_{r=1}^{R}\left(\hat{Y}_t^{(r)} - \frac{1}{R}\sum_{t=1}^{R}\hat{Y}_t^{(t)}\right)^2. \qquad (5)$$

Note that the use of 0.5, instead of 1, in the construction of $w_i^{(r)}$ is based on Fay's method (Dippo, Fay and Morganstein, 1984; Judkins 1990; Rao and Shao 1999). Asymptotically, v_{t1} is unbiased and consistent for $V_m(\hat{Y}_t)$ (Shao, Chen, and Chen 1998; Rao and Shao 1999; Shao and Chen 1999).

We now consider the estimation of $V_m(Y_t)$. Under model (2),

$$V_m(Y_t) = \sum_k M_k v_{t,k},$$

which is of the order $O(N)$, where N is the size of the population P. Usually $V_m(\hat{Y}_t)$ is of the order $O(N^2/n)$, where $n = \sum_h n_h$ is the sample size. Hence $V_m(Y_t)/V_m(\hat{Y}_t)$ is of the order $O(n/N)$ and the estimation of $V_m(Y_t)$ is not necessary if n/N is negligible (although some sampling fractions n_h/N_h are not negligible).

In the CES, however, n/N is around 15% and is not negligible. Hence, the estimation of $V_m(Y_t)$ is necessary. An asymptotically unbiased and consistent estimator of $V_m(Y_t)$ is

$$v_{t2} = \sum_k M_k s_{k,t}^2, \qquad (6)$$

where $s_{k,t}^2$ is the usual sample variance based on the respondents $y_{t,i}$ in the kth imputation cell.

4. APPROXIMATE VARIANCE ESTIMATORS

From section 3, a correct variance estimator for $\hat{Y}_t$ is $v_{t1} - v_{t2}$, where v_{t1} and v_{t2} are given by (5) and (6), respectively. Although v_{t1} can be easily extended to the case where $\hat{Y}_t$ is replaced by some nonlinear estimator such as $\hat{Y}_t^P/\hat{Y}_t^H$ (the ratio of weekly pay over hour), the extension of v_{t2} involves the derivation of Taylor expansion for each separate nonlinear estimator. Thus, for the CES, it is desired to derive an approximate variance estimator that is not exactly correct but does not require the computation of v_{t2}.

Note that if n/N is negligible, then we can simply use v_{t1} as an estimator of $V(\hat{Y}_t - Y_t)$. In the CES, however, using v_{t1} leads to overestimation, since n/N is not negligible (see also the simulation results in section 5). Since this overestimation is caused by the sampling fraction, a possible way to fix the problem is to incorporate sampling fractions in the half sample method. When there is no nonresponse, sampling fractions can be incorporated into the half sample method by using formula (2) with $w_i^{(r)}$ replaced by

$$\tilde{w}_i^{(r)} = \begin{cases} (1+0.5\sqrt{1-n_h/N_h}\,)\,w_i & \text{if the unit is in the } r\text{th half sample} \\ (1-0.5\sqrt{1-n_h/N_h}\,)\,w_i & \text{if the unit is not in the } r\text{th half sample,} \end{cases} \qquad (7)$$

when i is in stratum h.

Let $\tilde{v}_{t1}$ be the variance estimator obtained using (5) but with $w_i^{(r)}$ replaced by $\tilde{w}_i^{(r)}$. If we use $\tilde{v}_{t1}$ as an estimator of $V(\hat{Y}_t - Y_t)$, however, it has a negative bias, although it is better than the naive estimator that treats imputed values as observed data (see the simulation results in section 5).

While v_{t1} overestimates and $\tilde{v}_{t1}$ underestimates the true variance $V(\hat{Y}_t - Y_t)$, a compromise is to replace the sampling fraction n_h/N_h in (7) by the "estimated sampling fraction" $r_{h,t}/N_h$, where $r_{h,t}$ is the number of respondents in stratum h at month t. Let $\hat{v}_{t1}$ be the variance estimator obtained using (5) and (7) but with n_h/N_h in (7) replaced by $r_{h,t}/N_h$. Then

$$\tilde{v}_{t1} \le \hat{v}_{t1} \le v_{t1}.$$

All three variance estimators are asymptotically unbiased and are approximately equal when n/N is negligible. When n/N is not negligible, however, they are asymptotically biased.

To see the magnitude of the biases of $\tilde{v}_{t1}$, $\hat{v}_{t1}$, and v_{t1}, we consider the simplest case of no strata and $t = 1$. Let $y_i = y^E_{1,i}$, $x_i = y^E_{0,i}$ and

$$\hat{Y} = \sum a_i y_i + \sum (1 - a_i)\hat{R} x_i,$$

where $a_i = 1$ if y_i is a respondent and $a_i = 0$ otherwise, $\hat{R} = \sum a_i y_i / \sum a_i x_i$, and all summations are over $i \in S$. Let $\hat{U} = (\sum x_i / n)/(\sum a_i x_i / r)$, where r is the number of y-respondents. Then the correct variance estimator for $\hat{Y}$ is $v_1 - v_2$ with

$$v_1 = \frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}$$

and

$$v_2 = N\hat{U} s_d^2 + 2N\hat{U}\hat{R} s_{dx} + N\hat{R}^2 s_x^2,$$

where $s_d^2 = (r-1)^{-1} \sum a_i (y_i - \hat{R} x_i)^2$, $s_{dx} = (r-1)^{-1} \sum a_i x_i (y_i - \hat{R} x_i)$, and s_x^2 is the sample variance based on x_i's. Also,

$$\tilde{v}_1 = \left(1 - \frac{n}{N}\right)\left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}\right)$$

$$= v_1 - \frac{nN\hat{U}^2 s_d^2}{r} - 2N\hat{U}\hat{R} s_{dx} - N\hat{R}^2 s_x^2$$

and

$$\hat{v}_1 = \left(1 - \frac{r}{N}\right)\left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}\right)$$

$$= v_1 - N\hat{U}^2 s_d^2 - \frac{2rN\hat{U}\hat{R} s_{dx} + rN\hat{R}^2 s_x^2}{n}.$$

Since $v_1 - v_2$ is asymptotically unbiased, the bias of $v_{t1} = v_1$ is of the same order as v_2 and is always non-negative; the bias of $\tilde{v}_{t1} = \tilde{v}_1$ is of the same order as

$$N\hat{U} s_d^2 \left(1 - \frac{n\hat{U}}{r}\right) = -N\hat{U} s_d^2 \frac{\sum (1 - a_i) x_i}{\sum a_i x_i}$$

and is always non-positive; and the bias of $\hat{v}_{t1} = \hat{v}$ is of the same order as

$$N\hat{U}(1 - \hat{U}) s_d^2 + \left(1 - \frac{r}{n}\right)\left(2N\hat{U}\hat{R} s_{dx} + N\hat{R}^2 s_x^2\right). \quad (8)$$

The bias in (8) is non-negative if $s_{dx} \geq 0$ and $\hat{U} \approx 1$ (which is true if a_i is independent of x_i).

5. SOME SIMULATION RESULTS

To further study the biases of the variance estimators $v_{t1}, \tilde{v}_{t1}$ and $\hat{v}_{t1}$, we conducted a simulation study using a CES dataset (from 1980's) of 149,044 units as the population P. Each unit $i \in P$ has a vector $\boldsymbol{y}_i = (y^E_{t,i}, y^W_{t,i}, y^H_{t,i}, y^P_{t,i}, t = 0, 1, ..., 7)$ and a vector $\boldsymbol{r}_i$ consisting of response indicators of the components of $\boldsymbol{y}_i$, although all values of $\boldsymbol{y}_i$ are available (from administrative records). The sample S in the simulation was obtained by generating a stratified simple random sample $\{\boldsymbol{y}_i\}$ of size 23,092 from P according to the sample allocations listed in Table 1. The response indicators of $\{\boldsymbol{y}_i\}$ in the simulation were generated by drawing another (independent) stratified simple random sample $\{\boldsymbol{r}_i\}$ from P. Thus, nonrespondents in the simulation were random and distributed according to the values of the $\boldsymbol{r}_i$'s in the dataset P, but independent of the $\boldsymbol{y}_i$'s.

After the sample data and nonrespondents were generated, nonrespondents were imputed as described in section 2. Estimated monthly totals $\hat{Y}_t$ and monthly changes $\hat{Y}_t - \hat{Y}_{t-1}$ were calculated based on imputed data and their variance estimators, $\tilde{v}_{t1}, \hat{v}_{t1}, v_{t1}$, and $v_{t1} - v_{t2}$ were computed as described in sections 3 and 4. For comparison, the naive variance estimator v_{t0}, computed by treating imputed values as observed data, was also computed.

Based on 1,000 simulations, the relative biases (RB) and variances (Var) of the estimated totals $\hat{Y}_t$ and changes $\hat{Y}_t - \hat{Y}_{t-1}$, the RB and coefficient of variations (CV) of the variance estimators for $\hat{Y}_t$ and $\hat{Y}_t - \hat{Y}_{t-1}$, the coverage probability (CP) of the approximate 95% confidence intervals of the form

$$\text{the estimate} \pm 1.96 \sqrt{\text{the estimated variance}},$$

and the width (MW) of the confidence interval are given in Tables 2 through 5 respectively for 4 different variables. Estimated simulation standard errors are 2% for RB, CV, and MW, and 0.5% for CP.

Table 1
Sample Size by Stratum

SIC	SIZE	Stratum Size	Sample Size	Sampling Fraction	SIC	SIZE	Stratum Size	Sample Size	Sampling Fraction
10, 12-14	1	567	14	0.02439	50-51	1	3631	66	0.01812
	2	433	303	0.70000		2	3678	183	0.04987
	3	526	526	1.00000		3	4300	403	0.09375
	4	210	210	1.00000		4	1831	289	0.15789
	5	165	165	1.00000		5	833	320	0.38461
15-17	1	5055	129	0.02549	52-59	1	7084	149	0.02103
	2	4476	570	0.12731		2	5701	440	0.07724
	3	5281	1154	0.21854		3	8363	1037	0.12403
	4	2111	836	0.39583		4	4511	763	0.16915
	5	1005	1005	1.00000		5	4087	1002	0.24528
24-25, 32-29	1	3103	73	0.02349	60-62, 67	1	1384	17	0.01230
	2	3905	331	0.08475		2	971	38	0.03906
	3	6381	891	0.13966		3	1529	115	0.07500
	4	4273	1036	0.24242		4	981	67	0.06818
	5	4143	2127	0.51351		5	728	73	0.10000
20-23, 26-31	1	1754	40	0.02276	63-64	1	1364	15	0.01119
	2	1953	128	0.06564		2	652	20	0.03125
	3	3591	524	0.14599		3	754	87	0.11538
	4	3108	596	0.19167		4	435	48	0.11110
	5	3448	1041	0.30189		5	344	57	0.16667
40-49	1	1648	31	0.01902	7, 70-99	1	9641	230	0.02385
	2	1463	101	0.06918		2	6701	643	0.09602
	3	1988	221	0.11111		3	7833	1275	0.16275
	4	1171	211	0.18033		4	4839	1317	0.27215
	5	759	108	0.14286		5	4352	2067	0.47500

Table 2
Simulation Results for Employment

	Estimation of total			Variance estimation for estimated total																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Total*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	6.7E6	0.0	5.5E7	-37.0	47.6	85.3	7.7	-4.1	67.5	92.3	9.2	4.9	69.8	93.1	9.6	19.5	76.1	95.1	10.3	7.4	67.4	92.8	9.7
2	6.8E6	0.0	8.8E7	-34.3	28.8	86.9	9.6	-7.3	40.4	92.6	11.4	0.9	42.9	93.6	12.4	15.3	47.6	94.7	12.7	4.4	49.1	92.3	12.1
3	6.9E6	0.0	1.4E8	-26.1	30.4	88.2	12.9	-4.1	42.3	91.8	14.7	1.4	44.2	92.9	15.1	18.8	49.9	94.8	16.3	3.6	50.5	90.8	15.2
4	6.9E6	0.0	2.1E8	-22.5	32.9	89.3	16.1	-2.4	44.0	92.1	18.1	3.8	46.3	92.7	18.7	22.3	53.1	94.7	20.3	2.7	51.3	91.4	18.6
5	6.9E6	0.0	2.7E8	-21.9	35.0	88.3	18.4	-7.7	45.2	90.9	20.0	-1.1	47.9	92.0	20.7	16.2	55.6	94.4	22.4	-4.7	54.2	90.9	20.3
6	6.9E6	0.0	2.0E8	-8.8	40.5	91.7	17.1	-5.2	41.7	91.9	17.4	0.0	43.6	93.1	17.9	19.7	51.8	95.5	19.6	-3.1	52.5	90.5	17.6
7	6.9E6	0.0	1.5E8	-12.4	34.8	91.8	14.5	-8.6	36.1	92.5	14.8	-2.0	38.3	93.6	15.3	16.8	45.0	96.2	16.7	-6.6	42.4	92.7	15.0

	Estimation of change			Variance estimation for estimated change																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Change*	RB*	Var	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	8.0E4	-0.1	6.1E7	-43.0	25.4	84.9	7.5	-11.3	41.4	92.3	9.3	-4.5	43.9	93.7	9.7	9.4	48.7	95.6	10.3	8.6	51.7	93.5	10.3
3	9.7E4	-1.8	7.4E7	-35.0	31.7	85.0	8.7	-8.5	46.0	90.5	10.4	-3.2	47.7	91.0	10.7	11.7	53.1	93.4	11.5	-3.1	48.8	90.9	10.7
4	1.8E4	2.9	1.1E8	-31.8	42.3	87.4	11.0	-0.9	60.6	93.1	13.2	4.9	63.2	93.6	13.6	25.0	73.5	95.9	14.8	-2.5	47.7	89.9	13.1
5	4.4E4	3.4	1.1E8	-41.9	34.5	83.1	10.1	-10.8	57.3	91.4	12.5	-4.9	60.4	92.3	12.9	13.2	69.4	94.6	14.1	0.8	94.1	93.1	13.3
6	-1.1E4	9.3	1.1E8	-41.0	29.9	84.1	10.2	-12.6	42.0	91.1	12.4	-6.4	44.2	93.0	12.8	9.4	50.2	94.6	13.9	-4.1	53.9	93.0	13.0
7	1.6E3	3.2	1.2E8	-43.8	38.4	82.9	10.4	-15.9	57.5	89.6	12.7	-11.3	60.1	90.5	13.1	5.6	69.9	92.6	14.2	-0.2	75.5	90.0	13.8

Total: population total.
Change: population difference between the current month and the previous month.
Var: variance of the estimated total or change.
RB: relative bias = 100(bias/true value)%.
CV: coefficient of variation = 100 (standard error/true value)%.
CP: coverage probability of asymptotic confidence interval using estimated variance (in %).
MW: (mean width of asymptotic confidence interval)/10^7.
*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 3
Simulation Results for Non-supervisory Workers

	Estimation of Total			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Total*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	5.4E6	-0.1	4.6E7	-33.3	49.7	80.9	7.0	-4.4	66.1	88.1	8.4	4.6	68.6	89.9	8.8	19.8	75.5	92.3	9.4	3.4	65.3	89.9	8.7
2	5.5E6	-0.1	7.6E7	-30.6	31.4	84.0	9.2	-7.4	41.1	89.4	10.6	0.9	43.7	91.0	11.1	15.8	48.7	93.8	11.9	4.2	50.8	88.3	11.3
3	5.6E6	-0.1	1.2E8	-23.6	31.2	85.6	12.1	-4.8	41.0	89.5	13.5	0.7	42.9	90.0	13.9	18.4	48.7	93.1	15.1	3.9	50.9	90.8	14.1
4	5.6E6	-0.1	1.9E8	-19.0	34.5	88.4	15.7	-2.4	43.8	91.7	17.2	3.8	46.3	91.9	17.8	22.5	53.2	94.1	19.3	1.8	71.7	90.5	17.6
5	5.7E6	-0.1	2.4E8	-18.9	36.8	87.8	17.6	-7.1	45.3	89.7	18.9	-0.4	48.2	90.7	19.6	17.2	56.0	93.0	21.2	-4.1	54.7	90.4	19.2
6	5.7E6	0.0	1.8E8	-7.6	41.7	91.8	16.3	-4.7	42.8	92.4	16.6	0.6	44.8	92.7	17.0	20.6	53.1	95.4	18.6	-3.3	53.1	90.5	16.7
7	5.7E6	0.0	1.4E8	-10.9	36.1	91.9	14.1	-7.7	37.2	92.2	14.4	1.0	39.4	93.6	15.0	18.3	46.3	95.9	16.3	-8.5	42.5	92.6	14.3

	Estimation of Change			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Change*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	7.7E4	-0.8	5.1E7	-40.8	27.0	84.5	7.0	-12.9	41.2	91.5	8.4	-6.0	43.7	92.4	8.8	8.2	48.8	94.4	9.4	9.9	54.8	93.0	9.5
3	9.1E4	-1.4	6.2E7	-31.2	32.1	86.4	8.3	-8.7	42.8	91.2	9.5	-3.2	44.5	91.7	9.8	12.3	49.9	94.1	10.6	-3.1	47.7	91.3	9.8
4	1.6E4	19.6	9.1E7	-27.2	44.0	87.1	10.3	-1.1	59.4	92.8	12.0	4.7	62.1	94.1	12.3	24.9	73.0	95.8	13.5	-5.3	91.8	91.4	11.7
5	4.4E4	-0.4	9.5E7	-37.5	38.4	83.4	9.7	-10.0	58.6	90.8	11.7	-3.9	61.8	91.3	12.1	14.5	71.4	93.4	13.2	-2.1	79.4	92.3	12.2
6	-1.0E4	-19.3	9.0E7	-37.0	32.4	83.4	9.5	-11.1	43.1	89.6	11.3	-4.7	45.5	90.4	11.7	11.7	51.8	92.4	12.7	-3.3	54.7	90.9	11.8
7	7.9E2	48.7	1.0E8	-39.3	42.6	83.7	9.9	-14.5	59.7	89.2	11.7	-9.8	62.4	90.2	12.0	7.6	72.6	92.6	13.1	-1.3	76.8	90.6	12.6

Total: population total.
Change: population difference between the current month and the previous month.
Var: variance of the estimated total or change.
RB: relative bias = 100(bias/true value)%.
CV: coefficient of variation = 100 (standard error/true value)%.
CP: coverage probability of asymptotic confidence interval using estimated variance (in %).
MW: (mean width of asymptotic confidence interval) /10^7.
*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 4
Simulation Results for Hours

	Estimation of Total			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Total*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	1.9E8	-0.1	5.8E10	-31.5	28.0	79.0	8.0	2.3	44.4	88.3	9.7	12.3	46.5	90.5	10.2	33.4	53.4	93.6	11.1	8.0	48.7	90.9	10.0
2	2.0E8	-0.1	1.2E11	-30.2	32.8	84.7	11.6	-7.7	40.4	90.6	13.3	0.1	42.8	91.7	13.9	19.7	49.4	94.3	15.2	3.8	49.1	90.1	14.1
3	2.0E8	-0.1	1.8E11	-23.3	30.0	86.3	14.9	-6.3	36.7	90.3	16.4	-1.0	38.1	91.2	16.9	19.6	43.8	94.6	18.6	1.4	45.2	90.7	17.1
4	2.0E8	0.0	3.2E11	-20.2	35.6	90.2	20.2	-0.5	47.1	93.4	22.6	5.6	49.7	93.3	23.3	27.9	59.8	95.3	25.6	-0.4	79.7	91.2	22.6
5	2.1E8	0.0	4.4E11	-21.2	40.5	88.9	23.6	-7.9	52.3	90.7	25.5	-1.6	55.1	92.0	26.3	18.0	64.4	94.2	28.8	-5.1	64.2	90.9	25.8
6	2.1E8	0.0	3.4E11	-10.4	46.3	92.1	22.1	-5.9	48.9	92.2	22.6	-1.0	50.7	93.0	23.2	20.8	59.9	94.7	25.6	-3.3	65.7	90.3	22.9
7	2.1E8	0.0	2.3E11	-7.0	40.8	93.0	18.5	-2.2	42.8	93.2	19.0	4.2	44.7	94.1	19.6	27.2	53.2	95.8	21.6	-7.7	49.0	90.9	18.4

	Estimation of Change			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Change*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	5.0E6	0.1	8.8E10	-38.8	25.9	89.0	9.3	-9.7	35.1	92.4	11.3	-2.2	37.2	93.7	11.7	16.3	43.0	96.1	12.8	6.7	43.8	93.6	12.3
3	3.8E6	-1.0	1.1E11	-36.5	25.2	88.4	10.6	-12.6	34.5	91.9	12.4	-6.7	36.0	92.4	12.8	10.4	41.2	93.9	13.9	0.4	41.3	93.2	13.3
4	1.0E6	11.0	2.1E11	-31.2	45.6	87.3	15.2	-5.0	59.3	90.9	17.9	0.6	62.4	91.6	18.4	21.6	75.2	93.9	20.2	-3.4	98.8	91.5	18.0
5	2.1E6	-0.5	2.2E11	-41.6	39.9	85.6	14.3	-14.3	63.9	91.1	17.4	-8.4	66.6	90.1	18.0	10.5	76.0	94.9	19.7	1.5	95.1	93.2	18.9
6	-7.7E5	-7.8	1.9E11	-40.1	35.1	82.5	13.5	-12.7	47.5	89.5	16.3	-6.5	50.3	90.7	16.9	12.7	60.1	94.1	18.5	-9.5	55.1	91.1	16.6
7	2.5E5	-7.2	2.1E11	-39.0	48.4	82.9	14.3	-15.1	60.3	89.5	16.9	-10.6	62.4	90.3	17.3	8.0	72.3	94.0	19.0	-3.9	82.0	91.8	18.0

Total: population total.
Change: population difference between the current month and the previous month.
Var: variance of the estimated total or change.
RB: relative bias = 100(bias/true value)%.
CV: coefficient of variation = 100 (standard error/true value)%.
CP: coverage probability of asymptotic confidence interval using estimated variance (in %).
MW: (mean width of asymptotic confidence interval) /10^{10}.
*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 5
Simulation Results for Weekly Pay

	Estimation of Total			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Total*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	2.0E9	-0.1	9.5E12	-30.7	30.4	81.8	10.3	1.7	41.0	90.0	12.4	17.2	44.3	92.4	13.3	39.8	54.4	94.4	14.6	4.3	48.9	91.0	12.6
2	2.1E9	-0.1	1.7E13	-27.2	27.8	84.3	14.1	-3.4	38.7	89.2	16.2	7.9	41.2	91.2	17.1	31.1	48.1	93.5	18.9	3.3	51.5	91.6	16.8
3	2.1E9	-0.1	2.2E13	-14.3	34.7	85.6	17.4	1.1	42.2	88.1	18.9	8.0	43.9	89.5	19.5	34.9	51.4	93.5	21.8	2.6	50.4	91.4	19.0
4	2.2E9	-0.1	3.7E13	-12.3	40.3	90.1	22.8	6.4	50.6	92.8	25.1	13.8	53.0	94.1	26.0	41.2	63.0	96.1	28.9	-0.9	84.5	92.8	24.2
5	2.2E9	-0.1	5.0E13	-16.0	41.6	89.0	25.9	-1.5	51.8	91.4	28.1	5.9	54.8	92.0	29.1	29.3	64.6	94.3	32.2	-5.4	56.0	92.4	27.5
6	2.2E9	-0.1	4.5E13	-9.4	44.1	92.0	25.5	-3.8	46.9	92.6	26.3	1.8	48.7	92.8	27.1	27.8	57.8	95.0	30.3	-0.4	54.1	94.2	26.8
7	2.2E9	-0.1	3.5E13	-7.3	43.1	92.1	22.8	-0.7	48.3	92.8	23.6	6.8	50.0	93.9	24.5	31.9	57.0	96.4	27.2	-0.0	54.3	95.3	23.7

	Estimation of Change			Variance estimation																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Month	Change*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	6.4E7	-0.1	1.5E13	-37.6	25.7	85.4	12.2	-8.2	38.4	93.0	14.8	0.2	40.4	94.1	15.5	21.6	47.7	95.8	17.1	5.5	49.2	92.6	15.9
3	3.5E7	-1.6	1.3E13	-31.7	27.9	87.7	11.9	-5.2	42.3	92.2	14.0	2.2	43.8	92.8	14.6	22.3	48.9	94.3	15.9	3.5	43.2	93.5	14.7
4	2.1E7	6.6	2.4E13	-29.5	47.1	86.7	16.5	0.4	63.2	91.9	19.6	6.7	66.2	92.6	20.2	30.7	78.7	95.2	22.4	-4.3	96.9	90.6	19.2
5	2.1E7	-0.4	2.4E13	-40.5	34.1	83.5	15.1	-9.2	55.7	90.5	18.7	-2.4	58.9	92.0	19.4	19.9	69.2	94.9	21.5	3.6	90.0	92.5	19.9
6	1.4E7	2.0	2.3E13	-40.8	31.1	84.4	14.8	-13.5	46.0	91.4	17.8	-6.7	48.9	92.1	18.5	16.8	60.1	94.5	20.7	-4.4	53.0	91.5	18.8
7	1.1E7	-0.1	2.7E13	-40.5	42.0	83.1	16.0	-13.9	56.5	89.2	19.3	-8.7	58.7	90.6	19.9	13.0	68.8	92.8	22.1	-3.7	69.5	90.8	20.4

Total: population total.
Change: population difference between the current month and the previous month.
Var: variance of the estimated total or change.
RB: relative bias = 100(bias/true value)%.
CV: coefficient of variation = 100 (standard error/true value)%.
CP: coverage probability of asymptotic confidence interval using estimated variance (in %).
MW: (mean width of asymptotic confidence interval) /10^{12}.
*: Scientific notation (for example, 6,700,000 is 6.7E6).

From Tables 2 through 5, the relative biases of estimators of monthly totals and changes are negligible for all variables. The following is a summary for the simulation results of variance estimators in terms of RB and CV.

1. As expected, the naive variance estimator v_{t0} has a large negative relative bias.
2. The asymptotically unbiased variance estimator $v_{t1} - v_{t2}$ performs well in general. Its relative bias is always under 10% in absolute value and is frequently under 5%.
3. The variance estimator v_{t1} has a large positive relative bias in all cases. This indicates that the v_{t2} term is not negligible in the CES in which the overall sampling fraction, n/N, is about 15%.
4. The variance estimator $\tilde{v}_{t1}$, which is the same as v_{t1} but with sampling fractions n_h/N_h incorporated (section 4), has a negative relative bias in general. Its negative bias may be large, especially in the estimation of the variance for monthly changes.
5. The variance estimator $\hat{v}_{t1}$, which is the same as $\tilde{v}_{t1}$ but with sampling fractions n_h/N_h replaced by $r_{h,t}/N_h$, performs well in the simulation study, although it is not asymptotically unbiased (section 4). Its relative bias is large in a few cases, *e.g.*, in variance estimation for total of weekly pay at months 1 and 4, in variance estimation for total of hours at month 1, and in variance estimation for change of employment at month 7. In many cases, however, the performance of $\hat{v}_{t1}$ is even better than the asymptotically unbiased estimator $v_{t1} - v_{t2}$.

The following is a summary for the simulation results of confidence intervals in terms of CP and MW.

1. The CP of the confidence interval based on the naive variance estimator v_{t0} is substantially lower than the nominal level 95% in most cases.
2. The CP of the confidence interval based on the asymptotically valid variance estimator, $v_{t1} - v_{t2}$, is between 90% and 93% in most cases. This is often the case for an asymptotically valid variance estimator, *i.e.*, its relative bias is small but the CP of the related confidence interval is lower than the nominal level. One possible reason is that the convergence in distribution (asymptotic normality, which is the key for asymptotic confidence intervals) requires a larger sample size than the convergence of the second moment (in variance estimation).
3. In terms of CP, the confidence interval based on v_{t1} is the best. This might be because the overestimation in variance offsets the undercoverage in interval estimation. The mean width of the interval based on v_{t1} may

be substantially larger than those of other intervals, especially for weekly pay.

4. The CP of the confidence interval based on $\hat{v}_{t1}$, which is not asymptotically valid, is similar to that of the confidence interval based on $v_{t1} - v_{t2}$.

6. CONCLUSION AND DISCUSSION

For the survey estimators in the Current Employment Survey (CES) with imputed data, we propose an asymptotically unbiased and consistent estimator $v_{t1} - v_{t2}$ (section 3). Although v_{t1} can be easily computed using the grouped balanced half sample method, the computation of v_{t2} involves separate derivations for nonlinear estimators. Thus, several approximations, v_{t1}, $\tilde{v}_{t1}$, and $\hat{v}_{t1}$ (section 4) are considered and compared with $v_{t1} - v_{t2}$ in a simulation study in which a CES dataset is used as population. Our result shows that v_{t1} and $\tilde{v}_{t1}$ have large relative biases, due to the fact that the overall sampling fraction, 15%, is not negligible; the estimator $\hat{v}_{t1}$, which is the same as v_{t1} but incorporates an estimated sampling fraction (using the rate of response) in the balanced half sample method, performs fairly well. Thus, $\hat{v}_{t1}$ is recommended to replace $v_{t1} - v_{t2}$ if the computation of v_{t2} is too complicated. Since the use of the "observed sampling fraction" $r_{h,t}/N_h$ does not reflect the fact that information is available about the nonrespondents from previous months, $\hat{v}_{t1}$ may be improved using a more accurate estimated sampling fraction, for example, Rubin's (1987) "fraction of missing information".

Although our study is based on the CES, our results are applicable to any survey that adopts a similar sampling design and a similar imputation method. Furthermore, an extension to the case where model (2) involves $y_{t,i}, y_{t-1,i}, ..., y_{t-s,i}$ with an integer $s \geq 2$ is straightforward, although the derivation of v_{t2} (for an asymptotically valid variance estimator) is more complicated.

ACKNOWLEDGEMENTS

The authors are grateful to an Associate Editor and two referees for their helpful comments and suggestions. The research of Jun Shao was partly supported by the NSF grant DMS-9803112 and DMS-01-02223 and the NSA grant MDA 904-99-1-0032.

APPENDIX: PROOF OF (4)

It suffices to show that

$$\mathrm{Cov}_m(\hat{Y}_t, Y_t) = V_m(Y_t). \tag{9}$$

We show the case of a single imputation cell and $y_{t,i} = y_{t,i}^E$ (employment). The general case can be treated similarly.

We use mathematical induction. When $t = 1$,

$$\hat{Y}_t = \hat{\alpha}_1 Y_0.$$

By assumption (2),

$$\begin{aligned}\mathrm{Cov}_m(\hat{Y}_t, Y_t) &= \alpha_1^2 V_m(Y_0) + \sigma^2 E_m(Y_0) \\ &= N(\alpha_1^2 \nu_0 + \sigma^2 \mu_0) \\ &= V_m(Y_1).\end{aligned}$$

Suppose now that (9) is true at time $t-1$. Let E_t, V_t and Cov_t be the expectation, variance and covariance conditional on $y_{j-1,i}, R_j, j=1, ..., t$. Then

$$E_t(\hat{Y}_t) = \alpha_t \hat{Y}_{t-1}$$

and

$$\begin{aligned}\mathrm{Cov}_t(\hat{Y}_t, Y_t) &= \mathrm{Cov}_t(\hat{\alpha}_t \hat{Y}_{t-1}, Y_t) \\ &= \hat{Y}_{t-1}\, \mathrm{Cov}_t(\hat{\alpha}_t, Y_t) \\ &= \sigma^2 \hat{Y}_{t-1},\end{aligned}$$

where the last equality follows from assumption (2). By the induction assumption,

$$\mathrm{Cov}_m(\hat{Y}_t, Y_{t-1}) = V_m(Y_{t-1}).$$

Then

$$\begin{aligned}\mathrm{Cov}_m(\hat{Y}_t, Y_t) &= \mathrm{Cov}_m\left[E_t(\hat{Y}_t), E_t(Y_t)\right] + E_m\left[\mathrm{Cov}_t(\hat{Y}_t, Y_t)\right] \\ &= \alpha_t^2 \mathrm{Cov}_m(\hat{Y}_{t-1}, Y_{t-1}) + \sigma^2 E_m(\hat{Y}_{t-1}) \\ &= \sigma_t^2 V_m(Y_{t-1}) + \sigma^2 E_m(Y_{t-1}) \\ &= V_m(Y_t).\end{aligned}$$

REFERENCES

BUTANI, S., HARTER, R. and WOLTER, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 523-528.

DIPPO, C.S., FAY, R.E. and MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the Section on Survey Research Methodology*, American Statistical Association. 489-494.

LEE, H., RANCOURT, E. and SÄRNDAL C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

JUDKINS, D.R. (1990). Fay's method of variance estimation. *Journal of the Official Statistical*, 6, 223-239.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, 79, 811-822.

RAO, J.N.K., and SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

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

SHAO, J., and CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhya, B*, Special Issue on Sample Surveys, 187-201.

SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 97-101
Statistics Canada

Implementing Rao-Shao Type Variance Estimation with Replicate Weights

MICHAEL P. COHEN[1]

ABSTRACT

In estimating variances so as to account for imputation for item nonresponse, Rao and Shao (1992) originated an approach based on adjusted replication. Further developments (particularly the extension to balanced repeated replication from the jackknife replication of Rao and Shao) were made by Shao, Chen and Chen (1998). In this article we explore how these methods can be implemented using replicate weights.

KEY WORDS: Balanced Repeated Replication; Jackknife replication; Imputation; Item nonresponse; Weighted hot deck.

1. INTRODUCTION

Variance estimation by replication methods is facilitated by the use of replicate weights (Dippo, Fay and Morganstein 1984). In the past decade adjusted replication methods have been developed (Rao and Shao 1992; Shao, Chen and Chen 1998) that allow one to account for the variation due to imputation for item nonresponse in the estimation of variances. It is not, however, entirely obvious how these adjusted replication procedures can be implemented by means of replicate weights. This article explores how this can be done. The focus is on ways to prepare the dataset so that standard variance estimation software products that make use of replicate weights will work without modification. In the next to last section, however, some comments are made about whether modifying the software would help.

2. REPLICATION METHODS AND REPLICATE WEIGHTS

Wolter (1985) provides a comprehensive introduction to variance estimation for sample surveys. Chapters 3 and 4 cover the two replication methods pertinent to this article: the jackknife and balanced repeated replication. Shao and Tu (1995, chapter 6) is recommended for a more recent and advanced treatment. Variance estimation for surveys by replication continues to be an active area for research. Works that are even more recent include Brick and Morganstein (1996, 1997), Kott (2001), Rao and Shao (1996, 1999), Rust and Rao (1996), Shao (1996) and Valliant (1996).

The two replication methods work by creating subsets of the sample called *replicates*. The methods differ in the pattern by which replicates are formed. In balanced repeated replication (also called balanced half-sample replication), the replicates consist of roughly half the units in the original sample; hence they are also called *half samples*. In jackknife replication (as applied to survey data), the replicates typically consist of the original sample except that a single primary sampling unit (PSU) or a small number of PSUs in the same stratum is deleted. For both methods, the replicates can be considered samples in their own right. Therefore if $\hat{\theta}$ is an estimate of some quantity θ based on the original sample, we can form an estimate $\hat{\theta}^{(r)}$ of θ based on replicate r. If there are R replicates, we estimate the sampling variance of $\hat{\theta}$, $\text{var}(\hat{\theta})$, by

$$\widehat{\text{var}}(\hat{\theta}) = C_{M,R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2 \qquad (2.1)$$

where the constant $C_{M,R}$ depends solely on the replication method M and the number of replicates R.

In forming the estimate $\hat{\theta}$ of θ, use is made of the sample weights. For example, to estimate a population total for a particular item y, the estimate is the weighted sum of the values of y. Thus, if y_u and w_u are the values of y and the sample weight for sample unit u, then $\hat{\theta} = \sum_u w_u y_u$ where the sum is over all units in the sample. In addition to the sample weight w_u on the record for unit u, we can add replicate weights $w_u^{(r)}$, $r = 1$ to R, to the record on the file and calculate $\hat{\theta}^{(r)}$ in the same way as $\hat{\theta}$ except that $w_u^{(r)}$ replaces w_u for each sample unit u. Thus for the example in which $\hat{\theta}$ is the population total for y, $\hat{\theta}^{(r)} = \sum_u w_u^{(r)} y_u$. If unit u is not in replicate r, then $w_u^{(r)} = 0$. Some or all of the replicate weights for units that *are* in the replicate will be larger than their sample weights so that the units in the replicate continue to represent the entire population.

The use of replicate weights provided on the file to calculate the sampling variance estimates has advantages:

- Any statistics no matter how complicated that can be calculated for the whole sample can be calculated just as easily for each replicate. The sampling variance is then estimated by (2.1).

[1] Michael P. Cohen, Senior Mathematical Statistician, U.S. Bureau of Transportation Statistics, 400 Seventh Street SW, Washington, DC 20590 U.S.A.

- Adjustments for unit nonresponse and poststratification can (and should) be done individually for each replicate and incorporated in the replicate weights. This adjustment is usually done by an experienced sampling statistician and the adjusted replicate weights are put on the file so that the data analyst can use them without extra effort.
- Adjustments to the replicate weights put on the file can make use of auxiliary information not available to the data analyst, possibly for reasons of confidentiality. Even if not restricted, the auxiliary information may be difficult for the data analyst to obtain or use.
- General purpose software is available that employs replicate weights. Two software products that emphasize replication methods for surveys are WesVar from Westat, Inc. and VPLX from the U.S. Census Bureau. See the Web page

  //www.fas.harvard.edu/~stats/survey-soft/survey-soft.html

  for information on survey analysis software.

In this section we have ignored the complications that come from trying to capture the component of variance due to item imputation in the variance estimates. We begin to address these complications in the next section.

3. ADJUSTED REPLICATION METHODS

The works of Rao and Shao (1992) and Shao, Chen and Chen (1998) are key to this article. Shao and Chen (1999) and Shao and Steel (1999) also treat replication-based variance estimation for imputed survey data.

We begin by developing the notation, for the most part using that of Shao, Chen and Chen (1998). The population is divided into L strata with N_h clusters in the hth stratum. In the first stage of sampling in stratum $h, n_h \geq 2$ clusters are selected, the ith cluster being selected with probability $p_{hi}, i = 1, ..., N_h; h = 1, ..., L$. The clusters are selected without replacement and clusters in different strata are selected independently. The sampling fractions n_h/N_h are assumed to be small enough that no finite population correction is needed. Further stages of sampling may take place within each cluster, independently from cluster to cluster. There are N_{hi} ultimate population units in cluster i of stratum h. For population unit (h, i, j), there is a variable y_{hij} of interest. Let S be the collection of all sample units and let $\{\tilde{y}_{hij}, (h, i, j) \in S\}$ be the imputed dataset: the $\tilde{y}_{hij}$ are equal to y_{hij} when the item is observed and equal to the imputed value otherwise. The sample units are divided into *imputation classes* indexed by k and A_k is the index set of respondents for item y in imputation class k. We assume that the dataset contains identifiers ("flags") so that the nonrespondents can be identified.

In adjusted replication methods, $\tilde{y}_{hij}$ in imputation class k is adjusted to

$$\tilde{y}_{hij}^{(r)} = \begin{cases} \tilde{y}_{hij} + E_{A_k}^{(r)}(\tilde{y}_{hij}) - E_{A_k}(\tilde{y}_{hij}) \\ \quad \text{if } y_{hij} \text{ is imputed} \\ y_{hij} \\ \quad \text{if } y_{hij} \text{ is observed,} \end{cases} \tag{3.1}$$

where E_{A_k} is the expectation with respect to the original imputation procedure within imputation class k and $E_{A_k}^{(r)}$ is the expectation with respect to the imputation procedure based only on data in the rth replicate within imputation class k. This formula is given explicitly in Shao, Chen and Chen (1998, page 822) for balanced repeated replication and a variety of imputation methods. It also applies to the development in Rao and Shao (1992) for jackknife replication and weighted hot deck imputation.

We shall adopt the notation that $(h^\circ i^\circ j^\circ)$ denotes a unit that did not respond to item y and $(h'i'j')$ denotes a unit that did respond to item y. We assume that

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j') \in A_k} a_{h'i'j';\, h^\circ i^\circ j^\circ}\, y_{h'i'j'}$$

and

$$E_{A_k}^{(r)}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j') \in A_k} a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)}\, y_{h'i'j'}$$

where the $a_{h'i'j';\, h^\circ i^\circ j^\circ}$ and $a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)}$ are constants not depending on the values of the $y_{h'i'j'}$ and $a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)} = 0$ for $(h'i'j')$ not in replicate r. The $a_{h'i'j';\, h^\circ i^\circ j^\circ}$ and $a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)}$ may depend on auxiliary information available for all units in the sample. For the weighted hot deck of Rao and Shao (1992) and all of the imputation methods of Shao, Chen and Chen (1998), the expectations have this form.

3.1 Example: Ratio Imputation

This imputation method applies to situations in which there are auxiliary data $\{x_{hij}\}$ available for all sample units. Ratio imputation imputes a missing item $y_{h^\circ i^\circ j^\circ}$ by

$$x_{h^\circ i^\circ j^\circ} \sum_{(h'i'j') \in A_k} w_{h'i'j'} y_{h'i'j'} \Big/ \sum_{(h'i'j') \in A_k} w_{h'i'j'} x_{h'i'j'}.$$

So

$$a_{h'i'j';\, h^\circ i^\circ j^\circ} = x_{h^\circ i^\circ j^\circ} w_{h'i'j'} \Big/ \sum_{(h''i''j'') \in A_k} w_{h''i''j''} x_{h''i''j''}$$

and

$$a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)} = x_{h^\circ i^\circ j^\circ} w_{h'i'j'}^{(r)} \Big/ \sum_{(h''i''j'') \in A_k} w_{h''i''j''}^{(r)} x_{h''i''j''}$$

Notice that the $a_{h'i'j';\, h^\circ i^\circ j^\circ}$ and $a_{h'i'j';\, h^\circ i^\circ j^\circ}^{(r)}$ depend on the $\{x_{hij}\}$.

3.2 Example: Weighted Hot Deck Imputation

This imputation method imputes a missing item by a value randomly selected from the respondents to the same item with probability proportional to the weights of the respondents in the imputation class. See section 5 for further discussion of this method. Shao, Chen and Chen (1998, page 822) show that

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} w_{h'i'j'} y_{h'i'j'} \Big/ \sum_{(h'i'j')\in A_k} w_{h'i'j'}$$

and

$$E^{(r)}_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} w^{(r)}_{h'i'j'} y_{h'i'j'} \Big/ \sum_{(h'i'j')\in A_k} w^{(r)}_{h'i'j'}.$$

Thus

$$a_{h'i'j';h^\circ i^\circ j^\circ} = w_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w_{h''i''j''}$$

and

$$a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} = w^{(r)}_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w^{(r)}_{h''i''j''}.$$

4. THE DATA FILE FOR VARIANCE ESTIMATION

For simplicity we assume that each record contains an identifier indicating to which imputation class the unit belongs. Often the imputation class is determined by several variables on the record. A record will look something like this:

$$ID \quad IC \quad w_{hij}\, w^{(1)}_{hij} \cdots w^{(R)}_{hij} \quad \tilde{y}_{hij} \quad IF_y \quad \tilde{z}_{hij} \quad IF_z$$

where ID is the identifier for the unit, IC is the identifier for the imputation class, w_{hij} is the (full sample) weight, $w^{(1)}_{hij} \cdots w^{(R)}_{hij}$ are the replicate weights, $\tilde{y}_{hij}$ is the value (possibly imputed) of the variable y under consideration, IF_y is the imputation "flag" that indicates whether $\tilde{y}_{hij}$ is imputed, $\tilde{z}_{hij}$ is the value (possibly imputed) of another variable z and IF_z is the imputation "flag" that indicates whether $\tilde{z}_{hij}$ is imputed. There, of course, may be other variables on the files as well, for example an auxiliary variable x_{hij} available for all sample units.

We propose to add additional records, called extra records, to facilitate variance estimation. For each nonrespondent $(h^\circ i^\circ j^\circ)$ and respondent $(h'i'j')$ to item y in imputation class k, we create the record

$$ID \quad IC \quad 0 \quad \tilde{w}^{(1)}_{h^\circ i^\circ j^\circ;h'i'j'} \cdots \tilde{w}^{(R)}_{h^\circ i^\circ j^\circ;h'i'j'} \quad y_{h'i'j'} \quad IF_y \quad 0 \quad IF_z$$

where $IC = k$, ID is the identifier of the unit $(h^\circ i^\circ j^\circ)$ that did not respond to item y and

$$\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'} = (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ})\, w^{(r)}_{h^\circ i^\circ j^\circ},$$

$$r = 1, \ldots, R. \quad (4.1)$$

Note that the full sample weight is 0 on the extra records so these records do not affect the full sample estimates. The replicate estimates, though, agree with those defined by (3.1). Note also that the weights $\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'}$ may be negative.

Table 1
Numerical Illustration: Portion of Data File for Variance Estimation

ID	IC	w_{hij}	$w^{(1)}_{hij}$	...	$w^{(R)}_{hij}$	$\tilde{y}_{hij}$	IF_y	$\tilde{z}_{hij}$	IF_z
001	1	10.1	20.2000	...	0.0000	5.4	1	1.2	1
002	1	20.3	40.6000	...	0.0000	5.1	0	1.3	0
003	1	18.4	36.8000	...	0.0000	5.2	0	1.3	0
004	1	11.1	0.0000	...	22.2000	5.1	1	1.2	0
005	1	16.3	0.0000	...	32.6000	5.1	1	1.4	0
006	1	15.4	0.0000	...	30.8000	5.4	0	1.4	0
001	1	0.0	3.0162	...	0.0000	5.1	2	0.0	3
001	1	0.0	2.7339	...	0.0000	5.2	2	0.0	3
001	1	0.0	-5.7501	...	0.0000	5.4	2	0.0	3
004	1	0.0	0.0000	...	-8.3301	5.1	2	0.0	3
004	1	0.0	0.0000	...	-7.5505	5.2	2	0.0	3
004	1	0.0	0.0000	...	15.8806	5.4	2	0.0	3
005	1	0.0	0.0000	...	-12.2325	5.1	2	0.0	3
005	1	0.0	0.0000	...	-11.0876	5.2	2	0.0	3
005	1	0.0	0.0000	...	23.3201	5.4	2	0.0	3
001	1	0.0	5.5645	...	0.0000	0.0	3	1.3	2
001	1	0.0	5.0436	...	0.0000	0.0	3	1.3	2
001	1	0.0	-2.7512	...	0.0000	0.0	3	1.2	2
001	1	0.0	-4.0400	...	0.0000	0.0	3	1.4	2
001	1	0.0	-3.8169	...	0.0000	0.0	3	1.4	2

Table 1 provides a numerical illustration. In the illustration, the nine records (rows of the table) with $IF_y = 2$ are the extra records for item y. The first six records are the original records for the six sample units that constitute imputation class $IC = 1$. (The records at the end with $IF_z = 2$ are the extra records for item z and will be discussed in the next paragraph. In these records, the imputation flag for y, IF_y, has been set to 3 to indicate that these are extra records for an item other than y.) There are three respondents $(IF_y = 0)$ and three nonrespondents $(IF_y = 1)$ to item y. The method of imputation is assumed to be weighted hot deck. Only the first and last replicate weights $(w^{(1)}_{hij}$ and $w^{(R)}_{hij})$ are presented, but these are consistent with replicate weights used for the balanced repeated replication method of variance estimation. We have $\sum w_{hij}\, \tilde{y}_{hij} = 476.650$, $\sum w^{(1)}_{hij}\, \tilde{y}_{hij} = 506.048$ and $\sum w^{(R)}_{hij}\, \tilde{y}_{hij} = 455.696$ where the sums are over all the

records. The reader may verify that this agrees with $\sum w_{hij}\,\tilde{y}_{hij} = 476.650$, $\sum w^{(1)}_{hij}\,\tilde{y}^{(1)}_{hij} = 506.048$ and $\sum w^{(R)}_{hij}\,\tilde{y}^{(R)}_{hij} = 455.696$ obtained using (3.1) where the sums are over the first six records only.

Let us now consider item z. The extra records for this item have the form

$$ID\ \ IC\ \ 0\ \ \tilde{w}^{(1)}_{h^\circ i^\circ j^\circ;h'i'j'} \cdots \tilde{w}^{(R)}_{h^\circ i^\circ j^\circ;h'i'j'}\ \ 0\ \ IF_y\ \ z_{h'i'j'}\ \ IF_z$$

where $\tilde{w}^{(1)}_{h^\circ i^\circ j^\circ;h'i'j'} \cdots \tilde{w}^{(R)}_{h^\circ i^\circ j^\circ;h'i'j'}$ are computed by (4.1) but using the imputation method and response pattern for item z. The imputation method for z need not be the same as the imputation method for y but must be of the form discussed in section 3. In Table 1 the extra records for item z can be identified by having $IF_z = 2$. We have then $\sum w_{hij}\tilde{z}_{hij} = 120.130$, $\sum w^{(1)}_{hij}\,\tilde{z}_{hij} = 124.349$ and $\sum w^{(R)}_{hij}\,\tilde{z}_{hij} = 115.400$ where the sums are over all the records. This agrees with the sums obtained by (3.1).

Clearly the biggest disadvantage of this approach is the large number of extra records that have to be added to the file. This disadvantage is less severe when the imputation classes are small. (There are, however, many factors that go into determining the size of the imputation classes.) The advantages, on the other hand, include the following:

- The adjusted replicate estimates and variance estimates can be computed with any software designed to estimate variances by means of replicate weights.
- If there is another variable, say y', with the same pattern of nonresponse and the very same method of imputation as y (that is, the same a and $a^{(r)}$ values), the computation of replicate estimates for y' can be accommodated without adding more records.
- One can make estimates over subdomains, even if they cut across imputation class boundaries.
- Suppose the method of imputation is the weighted hot deck. Then one estimates the variance of a derived variable, say $\log y$ where $y > 0$, by simply adding the derived variable to each record and computing replicate estimates based on it. (We shall have more to say about the weighted hot deck in the next section.)

The data analyst may choose to delete the extra records from a copy of the data file and use the reduced file to check for outliers, formulate hypotheses, *etc.* When it comes time to estimate variances, the extra records would be merged back in.

It should be pointed out that Rao and Shao (1992) proposed and evaluated their jackknife variance estimation method only for the estimation of totals (or means). One must be cautioned against the use of the approach for more complex statistics. In the same way, Shao, Chen and Chen (1998) proposed their balanced repeated replication variance estimation method for functions of totals and for quantiles so it should not be used for other statistics.

5. THE WEIGHTED HOT DECK

The use of the weighted hot deck method of imputation (*e.g.*, Cox 1980) has a number of advantages so we devote a separate section to it. Rao and Shao (1992) concentrate on this imputation method and it is discussed also in Shao, Chen and Chen (1998). Under this method, a missing item is imputed by a value selected at random from the respondents to that item in the imputation class. The probability of selection is proportional to $w_{h'i'j'}$, the weight of the respondent. The respondents that have a positive probability of being selected are called *potential donors*; the nonrespondent being imputed is the *recipient*. If there is more than one item on the file that will be imputed by the weighted hot deck, simplifications occur if one uses complete respondents (units who responded to all items) as potential donors and uses only one donor to impute all items requiring weighted hot deck imputation for a given recipient. (The donor is selected for each sample unit having *any* item for which there is item nonresponse.)

If each unit in an imputation class has the same chance of responding to an item, the weighted hot deck yields design consistent estimates of means, totals and sample quantiles. The imputations, moreover, will be "plausible" in the sense of looking like real data.

An advantageous feature of the weighted hot deck is that it is equivariant under one-to-one transformations. To explain equivariance, consider a derived variable d where $d = g(y)$ and g is a one-to-one function. Then,using the weighted hot deck, we impute item y of unit $(h^\circ i^\circ j^\circ)$ that did not respond to the item by $\tilde{y}_{h^\circ i^\circ j^\circ}$ and use $g(\tilde{y}_{h^\circ i^\circ j^\circ})$ for d. This is equivalent to using the weighted hot deck to impute d by $\tilde{d}_{h^\circ i^\circ j^\circ}$ and using $g^{-1}(\tilde{d}_{h^\circ i^\circ j^\circ})$ for y. This feature of hot deck imputation in not shared by many other methods. For example, under *mean imputation* (in which the imputed value is the mean of the values for respondents in the imputation class), g would have to be linear for the equivariance property to hold. The pertinence of this to variance estimation by adjusted replicate methods is that when hot deck imputation is used, the data analyst can add $d = g(y)$ to the file and estimate variances for d as well as for y.

Suppose that the weighted hot deck is employed for several variables on a file and suppose that only complete respondents are used as potential donors. In this case, even if the patterns of nonresponse are different for the variables being imputed, the implementation of the adjusted replication by replicate weights described in the previous section can be carried out with the same set of extra replicate weights

$$\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'} = (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ})\, w^{(r)}_{h^\circ i^\circ j^\circ}$$

for each variable.

6. ALTERNATIVES

In this section we consider alternative methods including one that requires modifying the software.

6.1 First Alternative

One way to reduce the number of records is to have extra records of the form

$$ID' \quad IC \quad 0 \; \ddot{w}^{(1)}_{h'i'j'} \cdots \ddot{w}^{(R)}_{h'i'j'} \; y_{h'i'j'} \; IF_y \quad 0 \quad IF_z$$

where ID' is the identifier of the *potential donor* unit $(h'i'j')$ that responded to item y, B_k is the index set of units not responding to item y in imputation class k and

$$\ddot{w}^{(r)}_{h'i'j'} = \sum_{(h^\circ i^\circ j^\circ) \in B_k} (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ}) \, w^{(r)}_{h^\circ i^\circ j^\circ},$$

$$r = 1, \ldots, R.$$

Under this setup, for a given item there is only one extra record per potential donor. The chief disadvantage is that, because of the summation, estimates for subdomains that cut across imputation classes cannot be computed.

6.2 Second Alternative

Perhaps the most obvious implementation would be to add the $\tilde{y}^{(r)}_{hij}$ to the (hij) record and modify software to use the $\tilde{y}^{(r)}_{hij}$ rather than $\tilde{y}_{hij}$ when computing replicate estimates. The chief drawbacks are (1) sophisticated reprogramming of software would be needed, (2) if multiple variables may require imputation, the number of fields needed expands greatly and (3) it is unclear how a data analyst would estimate the variance of a derived variable, say d, unless the $\tilde{d}^{(r)}_{hij}$ were put on the file in advance. The favorable features of this implementation are (1) no extra records are needed and (2) variance estimates for subdomains do not require additional work.

7. CONCLUDING REMARKS

The adjusted replication methods of Rao and Shao (1992) and Shao, Chen and Chen (1998) provide a way of computing variance estimates that account for imputation for item nonresponse. An important next step is the development of ways to facilitate the computation. This article explored implementations based on the use of replicate weights.

ACKNOWLEDGEMENTS

The idea for this article came from a question posed by Robert E. Fay at a Washington Statistical Society presentation. The author is also grateful to both referees, the associate editor and the editor for helpful comments. The author worked for the National Center for Education Statistics when the initial version of the article was written. The views in this paper are those of the author and no official support by the U.S. Department of Education or the U.S. Department of Transportation is intended or should be inferred.

REFERENCES

BRICK, J.M., and MORGANSTEIN, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference,* U.S. Bureau of the Census, 861-866.

BRICK, J.M., and MORGANSTEIN, D. (1997). Computing sampling errors from clustered unequally weighted data using replication: WesVarPC. *Bulletin of the International Statistical Institute, Proceedings,* 1, 479-482.

COX, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 721-726.

DIPPO, C.S., FAY, R.E. and MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 489-494.

KOTT, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., and SHAO, J. (1996). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Society*, 91, 343-348.

RAO, J.N.K., and SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RUST, K., and RAO, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381-397.

SHAO, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, 27, 203-254.

SHAO, J., and CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhyā*, B, Special Issue on Sample Surveys, 187-201.

SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Society*, 93, 819-831.

SHAO, J., and STEEL, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Society*, 94, 254-265.

SHAO, J., and TU, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

VALLIANT, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.

WOLTER, K.M. (1985). *Introduction to Variance Estimation* New York: Springer-Verlag.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 103-114
Statistics Canada

Variance Estimation for the General Regression Estimator

RICHARD VALLIANT[1]

ABSTRACT

A variety of estimators of the variance of the general regression (GREG) estimator of a mean have been proposed in the sampling literature, mainly with the goal of estimating the design-based variance. Estimators can be easily constructed that, under certain conditions, are approximately unbiased for both the design-variance and the model-variance. Several dual-purpose estimators are studied here in single-stage sampling. These choices are robust estimators of a model-variance even if the model that motivates the GREG has an incorrect variance parameter.

A key feature of the robust estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. We also show that the delete-one jackknife implicitly includes the leverage adjustments and is a good choice from either the design-based or model-based perspective. In a set of simulations, these variance estimators have small bias and produce confidence intervals with near-nominal coverage rates for several sampling methods, sample sizes, and populations in single-stage sampling.

We also present simulation results for a skewed population where all variance estimators perform poorly. Samples that do not adequately represent the units with large values lead to estimated means that are too small, variance estimates that are too small, and confidence intervals that cover at far less than the nominal rate. These defects need to be avoided at the design stage by selecting samples that cover the extreme units well. However, in populations with inadequate design information this will not be feasible.

KEY WORDS: Confidence interval coverage; Hat matrix; Jackknife; Leverage; Model unbiased; Skewness.

1. INTRODUCTION

Robust variance estimation is a key consideration in the prediction approach to finite population sampling. Valliant, Dorfman, and Royall (2000) synthesize much of the model-based literature. In that approach, a working model is formulated that is used to construct a point estimator of a mean or total. Variance estimators are created that are robust in the sense of being approximately model-unbiased and consistent for the model-variance even when the variance specification in the working model is incorrect. In this paper, that approach is extended to the general regression estimator (GREG) to construct variance estimators that are approximately model-unbiased but are also approximately design-unbiased in single-stage sampling. A number of alternatives are compared including the jackknife and some variants of the jackknife. We will use a particular class of linear models along with Bernoulli or Poisson sampling as motivation for the variance estimators. However, some of these estimators can often be successfully applied in practice to single-stage designs where selections are not independent.

Associated with each unit in the population is a target variable Y_i and a p-vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ where $i = 1, ..., N$. The population vector of totals of the auxiliaries is $\mathbf{T}_x = (T_{x1}, ..., T_{xp})'$ where $T_{xk} = \sum_{i=1}^{N} x_{ki}$, $k = 1, ..., p$. The general regression estimator, defined below, is motivated by a linear model in which the Y's are independent random variables with

$$E_M(Y_i) = \mathbf{x}_i'\beta$$
$$\text{var}_M(Y_i) = v_i. \qquad (1.1)$$

In most situations (1.1) is a "working" model that is likely to be incorrect to some degree.

Assume that a probability sample s is selected and that the selection probability of sample unit i is $P(\delta_i = 1) = \pi_i$ where δ_i is a 0-1 indicator for whether a unit is in the sample or not. We assume that the sample selection mechanism is ignorable. Roughly speaking, ignorability means that the joint distribution of the Y's and the sample indicators, given the x's, can be factored into the product of the distribution for Y given x and the distribution for the indicators given x (see Sugden and Smith 1984 for a formal definition). In that case, model-based inference can proceed using the model and ignoring the selection mechanism.

The n-vector of targets for the sample units is $\mathbf{Y}_s = (Y_1, ..., Y_n)'$, and the $n \times p$ matrix of auxiliaries for the sample units is $\mathbf{X}_s = (\mathbf{x}_1, ..., \mathbf{x}_n)'$. Define the diagonal matrix of selection probabilities as $\boldsymbol{\Pi}_s = \text{diag}(\pi_i)$, $i \in s$, and the diagonal matrix of model-variances as $\mathbf{V}_s = \text{diag}(v_i)$. The GREG estimator of the total, $T = \sum_{i=1}^{N} Y_i$, is then defined as the Horvitz-Thompson estimator or π-estimator, $\hat{T}_\pi = \sum_s Y_i/\pi_i$, plus an adjustment:

$$\hat{T}_G = \hat{T}_\pi + \hat{\mathbf{B}}'(\mathbf{T}_x - \hat{\mathbf{T}}_x) \qquad (1.2)$$

where $\hat{\mathbf{B}} = \mathbf{A}_{\pi s}^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{Y}_s$ with $\mathbf{A}_{\pi s} = \mathbf{X}_s'\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s$, and $\hat{\mathbf{T}}_x = \sum_s \mathbf{x}_i/\pi_i$. The GREG estimator can also be written as

[1] Richard Valliant, Westat, 1650 Research Boulevard, Rockville, MD 20850.

$$\hat{T}_G = \mathbf{g}_s' \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s \tag{1.3}$$

with $\mathbf{g}_s = \mathbf{V}_s^{-1}\mathbf{X}_s\mathbf{A}_{\pi s}^{-1}(\mathbf{T}_x - \hat{\mathbf{T}}_x) + \mathbf{1}_s$ and $\mathbf{1}_s$ being an n-vector of 1's. Expression (1.3) will be useful for subsequent calculations.

A variant of the GREG, referred to as a "cosmetic" estimator, was introduced by Särndal and Wright (1984) and amplified by Brewer (1995, 1999). A cosmetic estimator also has design-based and model-based interpretations. The variance estimators in this paper could also be adapted to cover cosmetic estimation.

Assuming that N is known, the GREG estimator of the mean is simply $\hat{\bar{Y}}_G = \hat{T}_G / N$. We will concentrate on the analysis of $\hat{\bar{Y}}_G$. (In some situations, particularly ones where multi-stage sampling is used, the population size is unknown and an estimate, $\hat{N}$, must be used in the denominator of $\hat{\bar{Y}}_G$. The following analysis for the mean does not apply in that case.) Either quantitative or qualitative auxiliaries (or both) can be used in the GREG. If a qualitative variable like gender (male or female) is used, then two or more columns in $\mathbf{X}_s$ will be linearly dependent, in which case a generalized inverse, denoted by $\mathbf{A}_{\pi s}^{-}$, will be used in (1.2) and (1.3). Note that, although $\mathbf{A}_{\pi s}^{-}$ is not unique, the GREG estimator $\hat{\bar{Y}}_G$ is invariant to the choice of generalized inverse. The proof is similar to Theorem 7.4.1 in Valliant *et al.* (2000).

The GREG estimator is model-unbiased under (1.1) and is approximately design-unbiased in large probability samples. Note that the model-unbiasedness requires only that $E_M(Y_i) = \mathbf{x}_i'\beta$; if the variance parameters in (1.1) are misspecified, the GREG will still be model-unbiased. On the other hand, if $E_M(Y_i)$ is incorrectly specified, the GREG is model-biased and the model mean squared error may contain an important bias-squared term. The estimation error of the GREG $\hat{\bar{Y}}_G$ is defined as

$$\hat{\bar{Y}}_G - \bar{Y} = N^{-1}(\mathbf{a}_s' \mathbf{Y}_s - \mathbf{1}_r' \mathbf{Y}_r)$$

where $\bar{Y} = T/N$, $\mathbf{a}_s = \boldsymbol{\Pi}_s^{-1}\mathbf{g}_s - \mathbf{1}_s$, $\mathbf{Y}_r$ is the $(N-n)$-vector of target variables for the nonsample units, and $\mathbf{1}_r$ is a vector of $N-n$ 1's. Next, suppose that the true model for Y_i is

$$E_M(Y_i) = \mathbf{x}_i'\beta$$

$$\mathrm{var}_M(Y_i) = \psi_i, \tag{1.4}$$

i.e., the variance specification is different from (1.1) but $E_M(Y_i)$ is the same. Using the estimation error, the error-variance of $\hat{\bar{Y}}_G$ is then

$$\mathrm{var}_M(\hat{\bar{Y}}_G - \bar{Y}) = N^{-2}(\mathbf{a}_s' \boldsymbol{\Psi}_s \mathbf{a}_s + \mathbf{1}_r' \boldsymbol{\Psi}_r \mathbf{1}_r)$$

where the $n \times n$ covariance matrix for $\mathbf{Y}_s$ is $\boldsymbol{\Psi}_s = \mathrm{diag}(\psi_i)$ and $\boldsymbol{\Psi}_r$ is the $(N-n)\times(N-n)$ covariance matrix for $\mathbf{Y}_r$. When the sample and population sizes are both large and the sampling fraction, $f = n/N$, is negligible, the error-variance is approximately

$$\mathrm{var}_M(\hat{\bar{Y}} - \bar{Y}) \cong N^{-2}\sum_{i\in s} a_i^2 \psi_i. \tag{1.5}$$

Note that this variance depends on the true variance parameters, ψ_i, and on the working model variance parameters, v_i, because v_i is part of a_i. Since $\mathbf{a}_s$ is approximately the same as $\boldsymbol{\Pi}_s^{-1}\mathbf{g}_s$ when selection probabilities are small, the error variance in that case is also approximately

$$\mathrm{var}_M(\hat{\bar{Y}}_G - \bar{Y}) \cong N^{-2}\sum_{i\in s} \frac{g_i^2}{\pi_i^2}\psi_i. \tag{1.6}$$

For model-based variance estimation, we will take either of the asymptotic forms in (1.5) or (1.6) as the target. However, when the sampling fraction is substantial, the term $\mathbf{1}_r'\boldsymbol{\Psi}_r\mathbf{1}_r/N^2$ can be an important part of the error-variance and (1.5) or (1.6) may be poor approximations.

We will consider the design variance under two single-stage plans-Bernoulli and Poisson. In Poisson sampling, the indicators δ_i for whether a unit is in the sample or not are independent with $P(\delta_i = 1) = 1 - P(\delta_i = 0) = \pi_i$ (see Särndal, Swensson, and Wretman 1992, section 3.5, for a more detailed description). Bernoulli sampling is a special case of Poisson sampling in which each unit has the same inclusion probability. Under these two plans, the approximate design-variance of $\hat{\bar{Y}}_G$ is

$$\mathrm{var}_\pi(\hat{\bar{Y}}_G) \cong N^{-2}\sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} E_i^2 \tag{1.7}$$

where $E_i = Y_i - \mathbf{x}_i'\mathbf{B}$ and $\mathbf{B} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ is the regression parameter estimator evaluated for the full finite population. Särndal (1996) recommends using the GREG in conjunction with sampling plans for which (1.7) is valid on the grounds that the variance (1.7) is simple and that the use of regression estimation can often more than compensate for the random sample sizes that are a consequence of such designs.

The Bernoulli and Poisson designs and the linear models (1.1) and (1.4) serve mainly as motivation for the variance estimators presented in sections 2 and 3. As noted by Yung and Rao (1996, page 24), it is common practice to use variance estimators that are appropriate to a design with independent selections or to a with-replacement design even when a sample has been selected without replacement. Likewise, variance estimators motivated by a linear model are often applied in cases where departures from the model are anticipated. This practical approach underlies the thinking in this paper and is illustrated in the simulation study reported in section 4.

2. VARIANCE ESTIMATORS

Our general goal in variance estimation will be to find estimators that are consistent and approximately unbiased under both a model and a design. Kott (1990) also

considered this problem. Note that the goal here is not the estimation of a combined (or anticipated) model-design variance,

$$E_M E_\pi \left\{ \left[\left(\hat{\bar{Y}}_G - \bar{Y} \right) - E_M E_\pi \left(\hat{\bar{Y}}_G - \bar{Y} \right) \right]^2 \right\}.$$

Rather we seek estimators that are useful for both $\mathrm{var}_M(\hat{\bar{Y}} - \bar{Y})$ and $\mathrm{var}_\pi(\hat{\bar{Y}})$. The arguments given here are largely heuristic ones used to motivate the forms of the variance estimators. Additional, formal conditions such as those found in Royall and Cumberland (1978) or Yung and Rao (2000) are needed for model-based and design-based consistency and approximate unbiasedness.

First, consider estimation of the approximate model-variance given in (1.5). In the following development, we assume that, as N and n become large,

(i) $N \max_i(\pi_i) = O(n)$ and

(ii) $\mathbf{A}_{\pi s}/N$ converges to a matrix of constants, $\mathbf{A}_o$.

A residual associated with sample unit i is $r_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \mathbf{x}_i' \hat{\mathbf{B}}$. The vector of predicted values for the sample units can be written as

$$\hat{\mathbf{Y}}_s = \mathbf{H}_s \mathbf{Y}_s \tag{2.1}$$

where $\mathbf{H}_s = \mathbf{X}_s \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1}$. The predicted value for an individual unit is $\hat{Y}_i = \sum_{j \in s} h_{ij} Y_j$ where $h_{ij} = \mathbf{x}_i' \mathbf{A}_{\pi s}^{-1} \mathbf{x}_j / (v_j \pi_j)$ is the $(ij)^{\text{th}}$ element of $\mathbf{H}_s$. The matrix $\mathbf{H}_s$ is the analog to the usual hat matrix (Belsley, Kuh and Welsch 1980) from standard regression analysis. The diagonal elements of the hat matrix are known as leverages and are a measure of the effect that a unit has on its own predicted value. Notice that the inverses of the selection probabilities are involved in (2.1), although these would have no role in purely model-based analysis.

The following lemma, which is a variation of some results in Lemma 5.3.1 of (Valliant *et al.* 2000), gives some properties of the leverages and the hat matrix.

Lemma 1. Assume that (i) and (ii) hold. For $\mathbf{H}_s = \mathbf{X}_s \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1}$ the following properties hold for all $i \in s$:

(a) $h_{ij} = O(n^{-1})$

(b) $\mathbf{H}_s$ is idempotent.

(c) $0 \le h_{ii} \le 1$.

Proof: Since $h_{ij} = \mathbf{x}_i' \mathbf{A}_{\pi s}^{-1} \mathbf{x}_j / (v_j \pi_j)$, conditions (i) and (ii) imply that $h_{ij} = O(n^{-1})$. Part (b) follows from direct multiplication, using the definition of $\mathbf{H}_s$. To prove (c) note that $h_{ii} \ge 0$ since it is a quadratic form. Part (b) implies that $h_{ii} = h_{ii}^2 + \sum_{j \ne i} h_{ij} h_{ji}$ which can hold only if $h_{ii} \le 1$.

Next, we write the residual as $r_i = Y_i(1 - h_{ii}) - \sum_{j \in s(i)} h_{ij} Y_j$ where $s(i)$ is the sample excluding unit i. Since $E_M(r_i) = 0$, we have $E_M(r_i^2) = \mathrm{var}_M(r_i)$ and

$$E_M(r_i^2) = \Psi_i(1 - h_{ii})^2 + \sum_{j \in s(i)} h_{ij}^2 \Psi_j \tag{2.2}$$

under model (1.4). Using Lemma 1(a), we have $h_{ii} = o(1)$, $h_{ij}^2 = o(1)$, and consequently, $E_M(r_i^2) \cong \Psi_i$. Thus, in large samples, r_i^2 is an approximately unbiased estimator of the correct model-variance even though the variance specification in model (1.1) was incorrect. As a result, r_i^2 is a robust estimator of the model-variance for unit i regardless of the form of Ψ_i. A simple, robust estimator of the approximate model-variance (1.5) is then

$$v_{R1}(\hat{\bar{Y}}_G) = N^{-2} \sum_s a_i^2 r_i^2 \tag{2.3}$$

which is a type of "sandwich" estimator (see, *e.g.*, White 1982). (Note that a formal argument that v_{R1} is robust would require conditions such that $n^{-1} E_M(v_{R1})$ and $n^{-1} N^{-2} \sum_s a_i^2 \Psi_i$ converge to the same quantity.) Another variance estimator, similar to v_{R1} if $\mathbf{a}_s \cong \boldsymbol{\Pi}_s^{-1} \mathbf{g}_s$, is

$$v_{R2}(\hat{\bar{Y}}_G) = N^{-2} \sum_s \frac{g_i^2}{\pi_i^2} r_i^2. \tag{2.4}$$

An estimator of the approximate design-variance in (1.7) is

$$v_\pi(\hat{\bar{Y}}_G) = N^{-2} \sum_s \frac{1 - \pi_i}{\pi_i^2} r_i^2. \tag{2.5}$$

An alternative suggested by Särndal *et al.* (1989) as having better conditional properties is

$$v_{\mathrm{SSW}}(\hat{\bar{Y}}_G) = N^{-2} \sum_s \frac{1 - \pi_i}{\pi_i^2} g_i^2 r_i^2. \tag{2.6}$$

Another, similar estimator, used in the SUPERCARP software (Hidiroglou, Fuller and Hickman 1980) and derived using Taylor series methods, is

$$v_T(\hat{\bar{Y}}_G) = N^{-2} \frac{n}{n-1} \sum_s \left(\frac{g_i r_i}{\pi_i} - \frac{1}{n} \sum_s \frac{g_i r_i}{\pi_i} \right)^2. \tag{2.7}$$

As shown in the Appendix, the second term in parentheses in (2.7) converges in probability to zero under model (1.1). Thus, $v_T \approx v_{R2}$ in large samples.

When the selection probability of each unit is small, v_{SSW} will be similar to v_{R1}, v_{R2}, and v_T. All three will be approximately model-unbiased under (1.4) and approximately design-unbiased under Bernoulli and Poisson sampling. On the other hand, v_π is approximately design-unbiased but ignores the g_i coefficients and is biased under either model (1.1) or (1.4).

As a simple example, consider Bernoulli sampling with $\pi_i = n/N$ and the working model $E_M(Y_i) = x_i \beta$, $\mathrm{var}_M(Y_i) = \sigma^2 x_i$. Then the GREG is the ratio estimator

$\hat{\bar{Y}}_G = \bar{Y}_s \bar{x}/\bar{x}_s$ where $\bar{x}$ is a finite population mean. The approximate model variance under the more general specification, $\text{var}_M(Y_i) = \psi_i$, is $(\bar{\Psi}_s/n)(\bar{x}/\bar{x}_s)^2$ where $\bar{\Psi}_s = \sum_s \Psi_i/n$. The approximate design-variance is $(1-f)/(nN) \sum_{i=1}^{N}(Y_i - x_i\bar{Y}/\bar{x})^2$ where $\bar{Y}$ is a finite population mean. The estimator $v_{R2} = n^{-2}(\bar{x}/\bar{x}_s)^2 \sum_s (Y_i - x_i\bar{Y}_s/\bar{x}_s)^2$ is approximately unbiased for the model- variance and, because $\bar{x}/\bar{x}_s \xrightarrow{P} 1$ in large Bernoulli samples, v_{R2} is also approximately unbiased for the design-variance as long as f is small. In contrast, $v_\pi = n^{-2}(1-f) \sum_s (Y_i - x_i \bar{Y}_s/\bar{x}_s)^2$ is approximately design- unbiased but is model-unbiased only in balanced samples where $\bar{x} = \bar{x}_s$. Royall and Cumberland (1981) noted similar results for the ratio estimator in simple random sampling without replacement.

3. ALTERNATIVE VARIANCE ESTIMATORS USING ADJUSTED SQUARED RESIDUALS

The first alternative variance estimator we consider is the jackknife. The particular version to be studied is defined as

$$v_J = \frac{n-1}{n} \sum_{i=1}^{n} \left[\hat{\bar{Y}}_{G(i)} - \hat{\bar{Y}}_{G(\cdot)} \right]^2 \qquad (3.1)$$

where $\hat{\bar{Y}}_{G(i)}$ has the same form as the full sample estimator after omitting sample unit i. If the selection probability has the form $\pi_i = np_i$, then (3.1) can be rewritten. Using the convention that the subscript (i) means that sample unit i has been omitted, we have

$$\hat{\bar{Y}}_{G(i)} = \hat{T}_{G(i)}/N, \hat{\bar{Y}}_{G(\cdot)} = \sum_{i \in s} \hat{\bar{Y}}_{G(i)}/n, \hat{T}_{G(i)}$$

$$= \hat{T}_{\pi(i)} + \hat{\mathbf{B}}_{(i)}'(\mathbf{T}_x - \hat{\mathbf{T}}_{x(i)}),$$

$$\hat{T}_{\pi(i)} = n \sum_{i \in s(i)} Y_j/[\pi_j(n-1)], \hat{T}_{x(i)}$$

$$= n \sum_{j \in s(i)} x_j/[\pi_j(n-1)], \text{ and}$$

$$\hat{B}_{(i)} = A_{\pi s(i)}^{-1} X_{s(i)}' V_{s(i)}^{-1} \Pi_{s(i)}^{-1} Y_{s(i)} \text{ with}$$

$$A_{\pi s(i)} = X_{s(i)}' V_{s(i)}^{-1} \Pi_{s(i)}^{-1} X_{s(i)}$$

Another more conservative, but asymptotically equivalent, version of the jackknife replaces $\hat{\bar{Y}}_{G(\cdot)}$ with the full sample estimator $\hat{\bar{Y}}_G$. Design-based properties of the jackknife in (3.1) are usually studied in samples selected with replacement (see, *e.g.*, Krewski and Rao 1981, Rao and Wu 1985, Yung and Rao 1996), but applied in practice to without-replacement designs. Note that for the linear estimator $\hat{\bar{Y}}_\pi = N^{-1} \sum_{i \in s} Y_i/\pi_i$ in probability proportional to size without-replacement sampling, neither the jackknife, v_J, nor the approximations to v_J given later in this section, reduce to the usual Horvitz-Thompson or Yates-Grundy variance estimators.

With some effort we can write the jackknife in a form that involves the residuals and the leverages. The rewritten form will make clear the relationship of the jackknife to the variance estimators in section 2. First, note the following equalities that are easily verified:

$$\hat{T}_{\pi(i)} = \frac{n}{n-1}\left(\hat{T}_{\pi(i)} - \frac{Y_i}{\pi_i}\right), \hat{T}_{x(i)} = \frac{n}{n-1}\left(\hat{\mathbf{T}}_{x(i)} - \frac{\mathbf{x}_i}{\pi_i}\right) \qquad (3.2)$$

$$\mathbf{X}_{s(i)}' \mathbf{V}_{s(i)}^{-1} \boldsymbol{\Pi}_{s(i)}^{-1} \mathbf{Y}_{s(i)} = \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s - \mathbf{x}_i Y_i/v_i \pi_i,$$

$$\mathbf{A}_{\pi s(i)} = \mathbf{A}_{\pi s} - \mathbf{x}_i \mathbf{x}_i'/v_i \pi_i \qquad (3.3)$$

Using a standard formula for the inverse of the sum of two matrices, the slope estimator, omitting sample unit i, equals

$$\mathbf{B}_{(i)} = \hat{\mathbf{B}} + n^{-1} \sum_s \frac{\mathbf{A}_{\pi s}^{-1} \mathbf{x}_i}{1 - h_{ii}} \frac{r_i}{v_i \pi_i}.$$

Details of this and the succeeding computations are sketched in the Appendix. After a considerable amount of algebra, we have

$$\hat{T}_{G(i)} - \hat{T}_{G(\cdot)} = -\frac{n}{n-1}(D_i - \bar{D}_s) + \frac{n}{n-1} F_i$$

where

$$D_i = \frac{g_i r_i}{\pi_i (1 - h_{ii})}$$

and F_i is defined in the Appendix. The jackknife in (3.1) is then equal to

$$v_J\left(\hat{\bar{Y}}_G\right) = N^{-2} \frac{n}{n-1} \times$$

$$\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2 \sum_s F_i (D_i - \bar{D}_s) \right]. \qquad (3.4)$$

Expression (3.4) is an exact equality and could be used as a computational formula for the jackknife. This would sidestep the need to mechanically delete a unit, compute $\hat{\bar{Y}}_{G(i)}$, and so on, through the entire sample.

In large samples the first term in brackets in (3.4) is dominant while the second and third are near zero under some reasonable conditions. Thus, in large samples the jackknife is approximated by $v_J(\hat{\bar{Y}}_G) \cong N^{-2} \sum_s (D_i - \bar{D}_s)^2$, or, equivalently,

$$v_J\left(\hat{\bar{Y}}_G\right) \cong \frac{1}{N^2} \times$$

$$\sum_s \left[\frac{g_i r_i}{\pi_i (1 - h_{ii})}\right]^2 - \frac{1}{N^2 n}\left[\sum_s \frac{g_i r_i}{\pi_i (1 - h_{ii})}\right]^2. \qquad (3.5)$$

As shown in the Appendix, the second term in (3.5) converges in probability to zero under model (1.1). Consequently, a further approximation to the jackknife is

$$v_J(\hat{\bar{Y}}_G) \approx \frac{1}{N^2}\sum_s \left[\frac{g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}\right]^2. \qquad (3.6)$$

As (3.5) and (3.6) show, the jackknife implicitly incorporates the g_i^2 coefficients needed for estimating the model-variance. The right-hand side of (3.6) is itself an alternative estimator that we will denote by $v_J^*(\hat{\bar{Y}}_G)$.

Yung and Rao (1996) also derived an approximation to the jackknife for the GREG in multistage sampling. For single-stage sampling, their approximation is equal to v_T, defined in (2.7), which is the same as (3.5) if the leverages are zero. Duchesne (2000) also presented a formula for the jackknife, which he denoted as $\hat{V}_{JK2}$, that involved sample leverages. The advantage of (3.4) is that it makes clear which parts of the jackknife are negligible in large samples. Duchesne also presented an estimator, denoted by $\hat{V}_{JK2}^*$, that is essentially the same as v_{R2} and is an approximation to the jackknife.

Expressions (3.5) and (3.6) explicitly show how the leverages affect the size of the jackknife. Weighted leverages, h_{ii}, that are not near zero will inflate v_J. Depending on the configuration of the x's, this could be a substantial effect on some samples.

Since h_{ii} approaches zero with increasing sample size, v_J, v_{R2}, v_{SSW}, and v_T have the same asymptotic properties. In particular, the jackknife is approximately unbiased with respect to either the model or the design and is robust to misspecification of the variances in model (1.1). However, the factor $(1-h_{ii})$ in (3.6) is less than or equal to 1 and will make the jackknife larger than the other variance estimators. This will typically result in confidence intervals based on the jackknife covering at a higher rate than ones using v_{R2}, v_{SSW}, or v_T.

Note, also, that if a without-replacement sample is used, and some first-order or second-order selection probabilities are not small, the choices, v_{R2}, v_D, v_J, and v_J^* will be over-estimates of either the design-variance or the model-variance. To account for non-negligible selection probabilities, we can make some simple adjustments. An adjusted version of $v_J^*(\hat{\bar{Y}}_G)$, patterned after v_{SSW}, is

$$v_{JP}^*(\hat{\bar{Y}}_G) = \frac{1}{N^2}\sum_s \frac{(1-\pi_i)g_i^2 r_i^2}{\pi_i^2(1-h_{ii})^2}.$$

This expression is similar to $\hat{V}_{JK3}^*$ of Duchesne (2000), although $\hat{V}_{JK3}^*$ omits the leverages. Expression (3.6) also suggests another alternative that is closely related to an estimator of the error variance of the best linear unbiased predictor of the mean under model (1.1) (see, Valliant *et al.* 2000, chapter 5). This estimator is somewhat less conservative than (3.6), but still adjusts using the leverages:

$$v_D(\hat{\bar{Y}}_G) = \frac{1}{N^2}\sum_s \frac{g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}.$$

Because $h_{ii} = o(1)$, v_D is also approximately model and design-unbiased. A variant of this that may perform better when some selection probabilities are large is

$$v_{DP}(\hat{\bar{Y}}_G) = \frac{1}{N^2}\sum_s \frac{(1-\pi_i)g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}.$$

4. SIMULATION RESULTS

To check the performance of the variance estimators, we conducted several simulation studies using three different populations. The first is the Hospitals population listed in Valliant *et al.* (2000, Appendix B). The second population is the Labor Force population described in Valliant (1993). The third is a modification of the Labor Force population. In all three populations, sampling is done without replacement, as described below. These sampling plans will test the notion that variance estimators motivated, in part, by with-replacement designs can still be useful when applied to without-replacement designs.

The Hospitals population has $N = 393$ and a single auxiliary value x, which is the number of inpatient beds in each hospital. The Y variable is the number patients discharged during a particular time period. The GREG estimator for this population is based on the model $E_M(Y) = \beta_1 x^{1/2} + \beta_2 x$, $\text{var}_M(Y) = \sigma^2 x$. Samples of size 50 and 100 were selected using simple random sampling without replacement (srswor) and probability proportional to size (pps) without replacement with the size being the square root of x. For each combination of selection method and sample size, 3,000 samples were selected. The estimators $\hat{\bar{Y}}_G$, v_π, v_{R1}, v_{R2}, v_{SSW}, v_D, v_{DP}, v_J^*, v_{JP}^*, and v_J were calculated for each sample. For comparison we also included the π-estimator, $\hat{\bar{Y}}_\pi = \hat{T}_\pi / N$. The variance estimator v_T was included but is not reported here since results were little different from v_{R2}.

The Labor Force population contains 10,841 persons. The auxiliary variables used were age, sex, and number of hours worked per week. The Y variable was total weekly wages. Age was grouped into four categories: 19 years and under, 20-24, 25-34, and 35 or more. The model for the GREG included an intercept, main effects for age and sex, and the quantitative variable, hours worked. A constant model-variance was used. Samples of size 50, 100, and 250 were selected. The two selection methods used were srswor and sampling without replacement with probability proportional to hours worked. (This population has some clustering but this was ignored in these simulations.)

The third population was a version of Labor Force designed to inject some outliers or skewness into the weekly wages variable. We denote this new version as

"LF(mod)" for reference. In the original Labor Force population, weekly wages were top-coded at \$999. For each such top-coded wage, a new wage was generated equal to \$1,000 plus a lognormal random variable whose distribution had scale and shape parameters of 6.9 and 1. Recoded wages were generated for 4.4% of the population. Prior to recoding, the annualized mean wage was \$19,359, and the maximum was \$51,948; after recoding, the mean was \$23,103 and the maximum was \$608,116. Thus, LF(mod) exhibits more of the skewness in income that would be found in a real population.

The resulting LF(mod) distribution is shown in Figure 1 where weekly wages is plotted against hours worked for subgroups defined by age. In each panel the black points are for males while the open circles are for females. A horizontal reference line is drawn in each panel at \$999. Although there is a considerable amount of over-plotting, the general features are clear. Wage levels and spread go up as age increases, hours worked per week is related, though somewhat weakly, to wages, and wages are most skewed for age groups 25-34 and 35+. Less evident is the fact that wages for males are generally higher than ones for females.

Table 1 shows the empirical percentage relative biases, defined as the average over the samples of $(\hat{T} - T)/T$ for the π-estimator and general regression estimator for the various populations and sample sizes. Root mean square errors (rmse's), defined as the square root of the average over the samples of $(\hat{T} - T)^2$, are also shown. In the Hospitals population, both estimators have negligible bias at either sample size. The GREG is considerably more efficient in Hospitals than the π-estimator because of a strong relationship of Y to x. In the two Labor Force populations, both the π-estimator and the GREG are nearly unbiased while the GREG is somewhat more efficient as measured by the rmse for all sample sizes and selection methods.

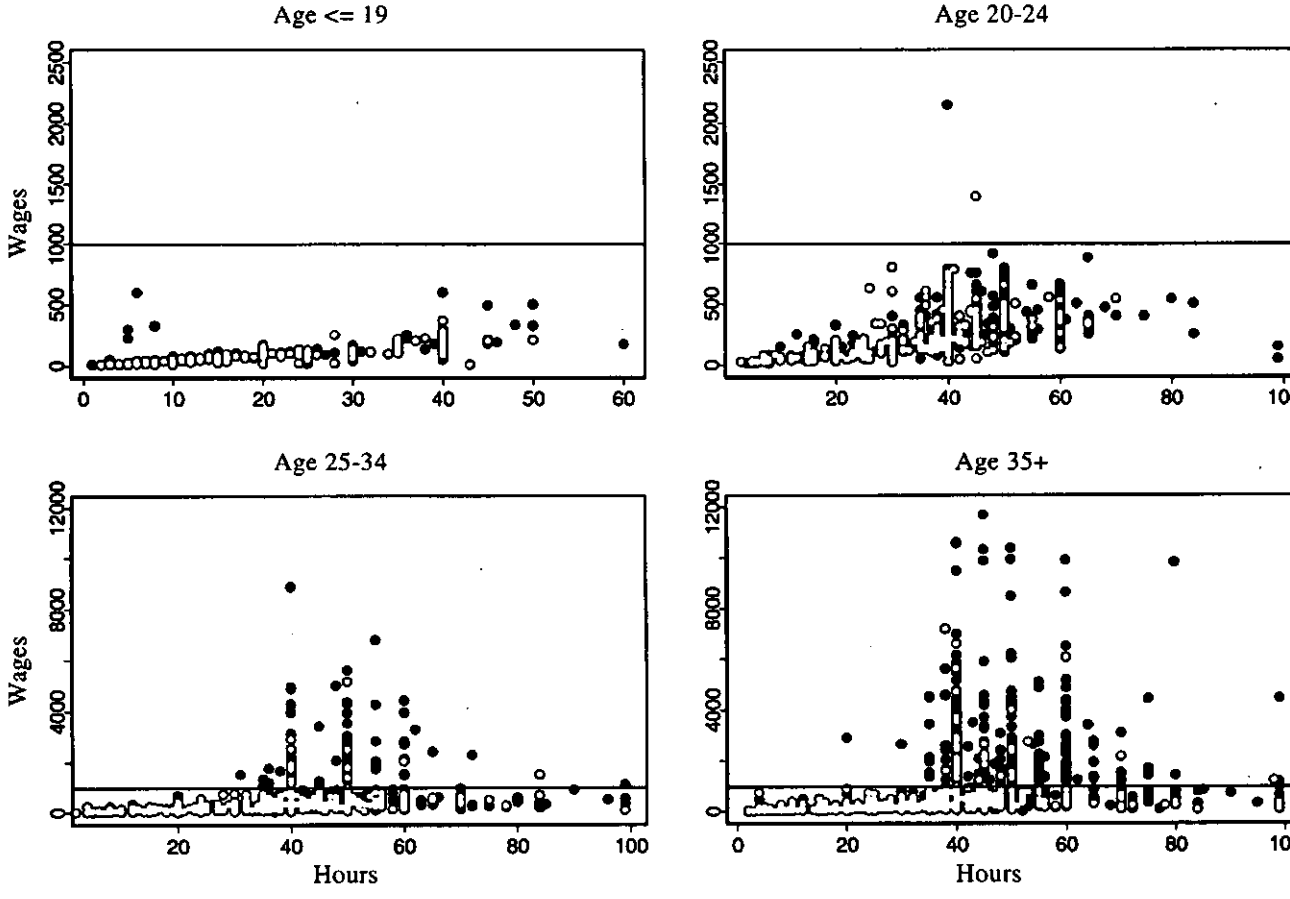


Figure 1. Scatterplots of Weekly Wages versus Hours Worked per Week in Four Age Groups for the LF(mod) population. Open circles are for females. Black circles are for males. A horizontal line is drawn at \$999 per week, the maximum value in the original Labor Force population.

Table 1
Relative biases and root means square errors (rmse's) of the π-estimator and the general regression in different simulation studies of 3,000 samples each.

	Hospitals		Labor Force			LF(mod)		
	n=50	n=100	n=50	n=100	n=250	n=50	n=100	n=250
Simple random samples								
$\hat{Y}_\pi$								
Relbias (%)	0.2	-0.1	-0.6	0	0	-0.1	0	-0.3
rmse	76.6	50.7	34.2	24.1	15.5	88.6	61.2	38.8
$\hat{Y}_G$								
Relbias (%)	0.2	0.2	0.1	0.1	0.2	0.4	0.2	-0.1
rmse	32.6	21.1	28.3	19.9	12.4	86.0	57.4	36.0
Probability proportional to size samples								
$\hat{Y}_\pi$								
Relbias (%)	-0.1	0.1	-0.5	0	0	0	-0.1	-0.1
rmse	37.6	24.4	28.2	20.3	12.6	80.6	54.6	34.1
$\hat{Y}_G$								
Relbias (%)	0.1	0.1	-0.10	0.10	0	-0.6	-0.7	-0.4
rmse	27.2	16.9	28.2	19.3	12.0	81.8	55.1	33.5

Table 2 lists the empirical relative biases (relbiases) of the nine variance estimators, defined as $100(\bar{v}$ - mse)/mse, where $\bar{v}$ is the average of a variance estimator over the 3,000 samples and mse is the empirical mean square error of the GREG. The rows of the table are sorted by the size of the relbias in LF(mod) for srswor's of size 50, although the ordering would be similar for the other populations, sample sizes, and selection methods. In the Hospitals population, the sampling fraction is substantial, especially when n=100. As might be expected, this results in the estimators that omit any type of finite population correction (fpc)–v_{R2}, v_D, v_J^*, and v_J–being severe over-estimates in either srswor or pps samples. Because v_{R1} lacks a term to reflect the model-variance of the nonsample sum, it under-estimates the mse badly when the sampling fraction is large.

In the Labor Force and LF(mod) populations, increasing sample size leads to decreasing bias. The estimators v_π, v_{R1}, v_{R2}, and v_{SSW} have negative biases that tend to be less severe as the sample size increases. The jackknife v_J and its variants, v_J^*, v_{JP}^*, are over-estimates, especially at n=50. The estimators, v_D and v_{DP}, are more nearly unbiased at each of the sample sizes than most of the other estimators.

The empirical coverages of 95% confidence intervals across the 3,000 samples in each set are shown in Table 3 for the Hospitals population. The three choices of variance estimator that use the leverage adjustments but not fpc's–v_D, v_J^*, and v_J^*–are larger and, thus, have higher coverage rates than v_π, v_{R2}, and v_{SSW}. The tendency of the jackknife to be larger than other variance estimates for the GREG has also been noted by Stukel, Hidiroglou, and Särndal (1996). This is an advantage for the smaller sample size, n=50. When n=100 and the sampling fraction is large, the estimators with the fpc's–v_π, v_{SSW}, v_{DP}, and v_J^*,–have closer to the nominal 95% coverage rates while v_{R2}, v_D, v_J^*, and v_J cover in about 97 or 98% of the samples. The estimator v_{JP}^*, that approximates the jackknife but includes an fpc, is a good choice at either sample size or sampling plan.

Table 2
Relative biases of nine variance estimators for the general regression estimator in different simulation studies of 3,000 samples each.

	Hospitals		Labor Force			LF(mod)		
	n=50	n=100	n=50	n=100	n=250	n=50	n=100	n=250
Simple random samples								
v_π	-8.6	-4.2	-18.1	-12.3	-7.5	-16.3	-2.8	-2.6
v_{R1}	-18.9	-27.0	-11.3	-9.9	-8.0	-9.6	-0.7	-3.3
v_{SSW}	-7.6	-3.0	-10.9	-9.1	-5.9	-9.3	0.1	-1.1
v_{R2}	5.9	30.1	-10.5	-8.2	-3.7	-8.8	1.0	1.3
v_{DP}	-1.4	0.2	0.1	-3.8	-3.8	0.6	5.1	0.8
v_D	13.0	34.3	0.6	-2.9	-1.6	1.0	6.1	3.2
v_J	18.4	37.4	13.9	2.2	0.3	11.2	10.5	4.8
v_{JP}^*	5.4	3.5	14.0	2.1	-1.7	12.4	10.5	2.7
v_J^*	20.8	38.8	14.5	3.1	0.7	12.9	11.5	5.2
Probability proportional to size samples								
v_π	-5.9	-0.9	-22.1	-12.1	-6.8	-16.5	-10.6	-0.3
v_{R1}	-19.7	-32.4	-11.9	-7.7	-7.1	-9.1	-8.2	-2.7
v_{SSW}	-4.0	0.0	-11.6	-7.0	-4.9	-8.7	-7.3	-0.1
v_{R2}	16.0	52.6	-11.2	-6.0	-2.5	-8.3	-6.3	2.6
v_{DP}	0.1	2.0	0.8	-0.3	-1.6	0.9	-2.5	2.1
v_D	20.8	55.6	1.3	0.7	0.8	1.4	-1.5	4.8
v_J	23.6	57.2	22.6	11.8	5.3	14.6	4.7	7.3
v_{JP}^*	4.4	4.0	19.7	9.3	3.1	14.8	3.9	4.9
v_J^*	26.1	58.8	20.3	10.3	5.5	15.4	5.0	7.7

Table 3
95% confidence interval coverage rates for simulations using the Hospitals population and nine variance estimators. 3,000 simple random samples and probability proportional to size were selected without replacement for samples of size 50 and 100. L is percent of samples with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1.96$; M is percent with $|\hat{\bar{Y}}_G - \bar{Y}|/v^{1/2} \le 1.96$; U is percent with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < 1.96$.

	n=50			n=100		
	L	M	U	L	M	U
Simple random samples						
v_π	3.1	92.1	4.8	2.6	93.6	3.9
v_{R1}	4.2	91.0	4.7	4.8	89.8	5.5
v_{SSW}	3.3	92.5	4.2	2.8	94.0	3.1
v_{R2}	2.8	93.9	3.3	1.4	97.0	1.6
v_{DP}	3.1	93.0	3.9	2.7	94.3	2.9
v_D	2.4	94.6	3.0	1.2	97.3	1.5
v_J	2.2	95.0	2.8	1.2	97.3	1.5
v_{JP}^*	2.9	93.6	3.5	2.6	94.6	2.9
v_J^*	2.2	95.1	2.8	1.2	97.4	1.4
Probability proportional to size samples						
v_π	2.9	93.9	3.2	2.6	94.6	2.8
v_{R1}	4.1	92.0	3.9	5.0	89.3	5.7
v_{SSW}	2.9	94.2	2.9	2.6	94.8	2.6
v_{R2}	2.1	95.8	2.1	0.9	98.3	0.8
v_{DP}	2.7	94.5	2.8	2.5	95.0	2.5
v_D	1.9	96.2	1.9	0.9	98.3	0.8
v_J	1.8	96.3	1.9	0.9	98.4	0.7
v_{JP}^*	2.6	94.8	2.6	2.4	95.4	2.2
v_J^*	1.7	96.5	1.8	0.8	98.4	0.7

Tables 4 and 5 show the coverage rates for the Labor Force and LF(mod) populations. For the former, v_{DP}, v_D, v_J, v_{JP}^*, and v_J^* are clearly better in Labor Force at n=50 for both srswor and pps samples. But, for n=250, coverages rates are similar for all estimators. The purely design- based estimator, v_π, is unsatisfactory at the smaller sample sizes for either sampling plan. As in Hospitals, v_{JP}^* gives near nominal coverage at each sample size in the Labor Force population.

The most striking results in Tables 4 and 5 are for LF(mod) where all variance estimators give poor coverage. Coverages range from 78.0% for the combination (v_π, n=50, srswor) to 90.7% for (v_J and v_J^*, n = 250, pps). Virtually all cases of non-coverage are because $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1.96$, where v is any of the variance estimators. The poor coverage rates occur even though the π-estimator and GREG are unbiased over all samples (see Table 1) and, in the cases of v_J, v_{JP}^*, and v_J^*, the variance estimators are overestimates (see Table 2).

Table 4

95% confidence interval coverage rates for simulations using the Labor Force and LF(mod) populations and nine variance estimators. 3,000 simple random samples were selected without replacement for samples of size 50 and 100. L is percent of samples with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1.96$; M is percent with $|\hat{\bar{Y}}_G - \bar{Y}|/v^{1/2} \le 1.96$; U is percent with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} > 1.96$.

	n=50			n=100			n=250		
	L	M	U	L	M	U	L	M	U
Labour Force									
v_π	5.3	91.4	3.2	4.3	92.8	2.9	2.8	94.1	3.1
v_{R1}	4.9	92.4	2.7	4.3	93.0	2.7	2.8	93.9	3.3
v_{SSW}	4.9	92.5	2.6	4.3	93.1	2.7	2.8	94.1	3.1
v_{R2}	4.9	92.5	2.6	4.2	93.2	2.6	2.5	94.6	2.9
v_{DP}	4.2	93.6	2.2	3.9	93.7	2.4	2.6	94.5	2.9
v_D	4.2	93.6	2.2	3.9	93.9	2.2	2.4	94.9	2.7
v_J	3.0	95.1	1.9	3.4	94.7	1.9	2.4	95.0	2.7
v_{JP}^*	3.0	95.1	1.9	3.3	94.7	1.9	2.5	94.8	2.7
v_J^*	3.0	95.1	1.9	3.3	94.8	1.9	2.4	95.0	2.7
LF(mod)									
v_π	21.0	78.0	0.9	14.1	85.5	0.4	9.9	89.7	0.4
v_{R1}	20.9	78.7	0.3	14.1	85.7	0.2	10.2	89.5	0.3
v_{SSW}	20.9	78.8	0.3	14.0	85.8	0.2	9.9	89.9	0.3
v_{R2}	20.8	78.8	0.3	13.8	86.0	0.2	9.7	90.1	0.3
v_{DP}	19.7	80.0	0.2	13.4	86.5	0.1	9.7	90.1	0.3
v_D	19.7	80.0	0.2	13.2	86.7	0.1	9.6	90.1	0.3
v_J	18.4	81.4	0.2	12.7	87.2	0.1	9.4	90.3	0.3
v_{JP}^*	18.4	81.4	0.2	12.7	87.2	0.1	9.5	90.2	0.3
v_J^*	18.3	81.5	0.2	12.6	87.3	0.1	9.3	90.4	0.3

Negative estimation errors, $\hat{\bar{Y}}_G - \bar{Y}$, occur in samples that include relatively few persons with large weekly wages. Figure 2 is a plot of t-statistics based on $\sqrt{v_{JP}^*}$, *i.e.*, $(\hat{\bar{Y}}_G - \bar{Y})/\sqrt{v_{JP}^*}$, versus the number of sample persons with weekly wages of $1,000 or more in sets of 1,000 samples for (srswor; n=50, 100, 250). The negative estimation errors in samples with few persons with high incomes lead to negative t-statistics, and confidence intervals that miss the population mean on the low side. The problem decreases with increasing sample size, but the convergence to the nominal coverage rates is slow and occurs "from the bottom up." Regardless of the variance estimator used, coverage will be less than 95% unless the sample is quite large.

Table 5

95% confidence interval coverage rates for simulations using the Labor Force and LF(mod) populations and nine variance estimators. 3,000 probability proportional to size samples were selected without replacement for samples of size 50, 100 and 250. L is percent of samples with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1.96$; M is percent with $|\hat{\bar{Y}}_G - \bar{Y}|/v^{1/2} \le 1.96$; U is percent with $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} > 1.96$.

	n=50			n=100			n=250		
	L	M	U	L	M	U	L	M	U
Labour Force									
v_π	5.7	90.2	4.1	3.7	92.9	3.4	3.1	94.3	2.6
v_{R1}	5.3	92.1	2.6	3.3	93.8	2.9	3.5	94.0	2.5
v_{SSW}	5.2	92.2	2.6	3.2	94.0	2.9	3.3	94.4	2.2
v_{R2}	5.2	92.3	2.6	3.1	94.1	2.8	3.0	94.8	2.2
v_{DP}	4.3	93.6	2.0	2.9	94.7	2.4	3.0	94.9	2.1
v_D	4.3	93.7	2.0	2.9	94.7	2.4	2.8	95.1	2.1
v_J	3.3	95.5	1.2	2.4	95.8	1.7	2.6	95.5	1.9
v_{JP}^*	3.3	95.4	1.3	2.6	95.5	1.9	2.7	95.3	1.9
v_J^*	3.3	95.4	1.3	2.6	95.6	1.8	2.6	95.6	1.8
LF(mod)									
v_π	19.6	79.7	0.7	15.0	84.4	0.7	9.9	89.8	0.4
v_{R1}	20.2	79.6	0.2	15.9	83.8	0.3	10.3	89.4	0.3
v_{SSW}	20.1	79.7	0.2	15.8	84.0	0.3	10.0	89.8	0.2
v_{R2}	20.1	79.7	0.2	15.6	84.1	0.2	9.8	90.0	0.2
v_{DP}	18.7	81.1	0.2	14.8	85.0	0.1	9.7	90.0	0.2
v_D	18.7	81.1	0.2	14.7	85.2	0.1	9.4	90.4	0.2
v_J	16.6	83.2	0.1	13.6	86.4	0.0	9.1	90.7	0.2
v_{JP}^*	16.6	83.3	0.1	13.9	86.1	0.0	9.4	90.4	0.2
v_J^*	16.5	83.4	0.1	13.8	86.2	0.0	9.1	90.7	0.2

We also examined how well the variance estimators perform, conditional on sample characteristics. We present only results related to bias of the variance estimators to conserve space. For the Hospitals population, we sorted the samples based on $D_x = \mathbf{1}'(\hat{\mathbf{T}}_x - \mathbf{T}_x)$, which is the sum of the differences of the π-estimates of the totals of $x^{1/2}$ and x from their population totals. Twenty groups of 150 samples each were then formed. In each group, we computed the bias of $\hat{\bar{Y}}_G$ along with the rmse, and the square root of the average of each variance estimator. The results are plotted in Figure 3 for srswor with n=50 and 100 and for pps with n=50 and 100. A subset of the variance estimators is plotted. The horizontal axis in each panel gives values of D_x. Since v_J, v_J^*, v_D, and v_{R2} are similar through most of the range of D_x, only the jackknife v_J is plotted. Also, v_{DP} and v_{JP}^* are close, and only the latter is plotted. The GREG does have a conditional bias that affects the rmse in off-balance samples. The poor conditional properties of v_π are most evident in the simple random samples where the bias of v_π as an estimate of the mse runs from negative to positive over the range of D_x. Among the other variance estimates, conditional biases are similar to the unconditional biases in Table 2. Both v_{JP}^* and v_{SSW} are in theory approximately design and model-unbiased, and both track the rmse well.

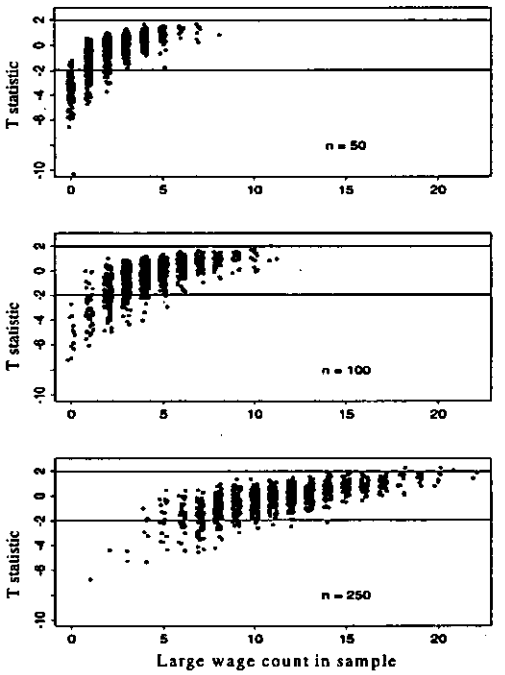


Figure 2. Plot of t-statistics versus the number of sample persons with weekly wages greater than \$1,000 in the sets of 1,000 simple random samples of size $n = 50, 100, 250$ from the LF(mod) population. Horizontal reference lines are drawn at ± 1.96. Points are jittered to minimize overplotting.

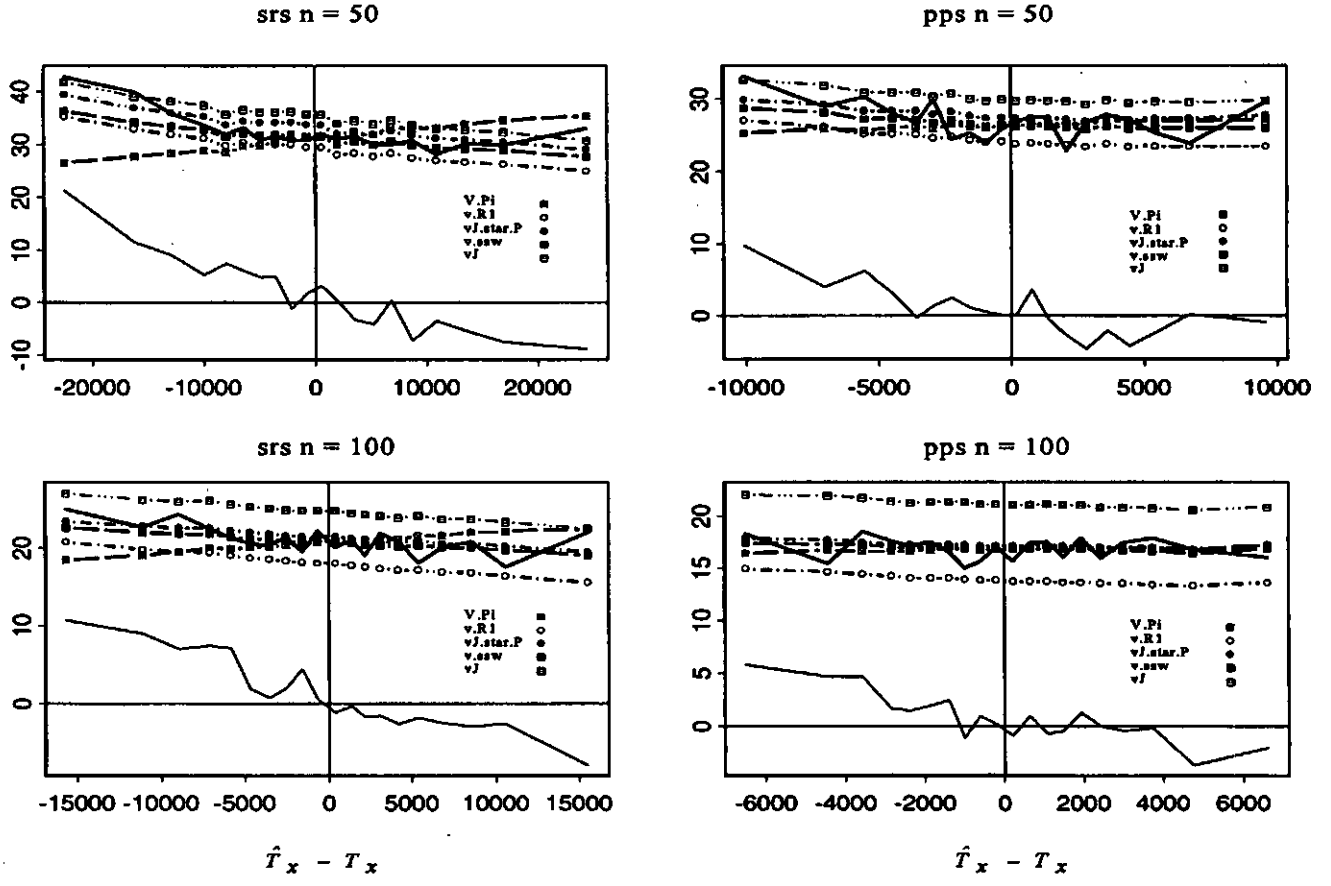


Figure 3. Plot of conditional biases, rmse's, and means of standard error estimates of the GREG for the samples from the Hospitals population. Horizontal and vertical reference lines are drawn at 0. The lowest curve each panel is the bias of the GREG. The thick solid line is the conditional root mean square error.

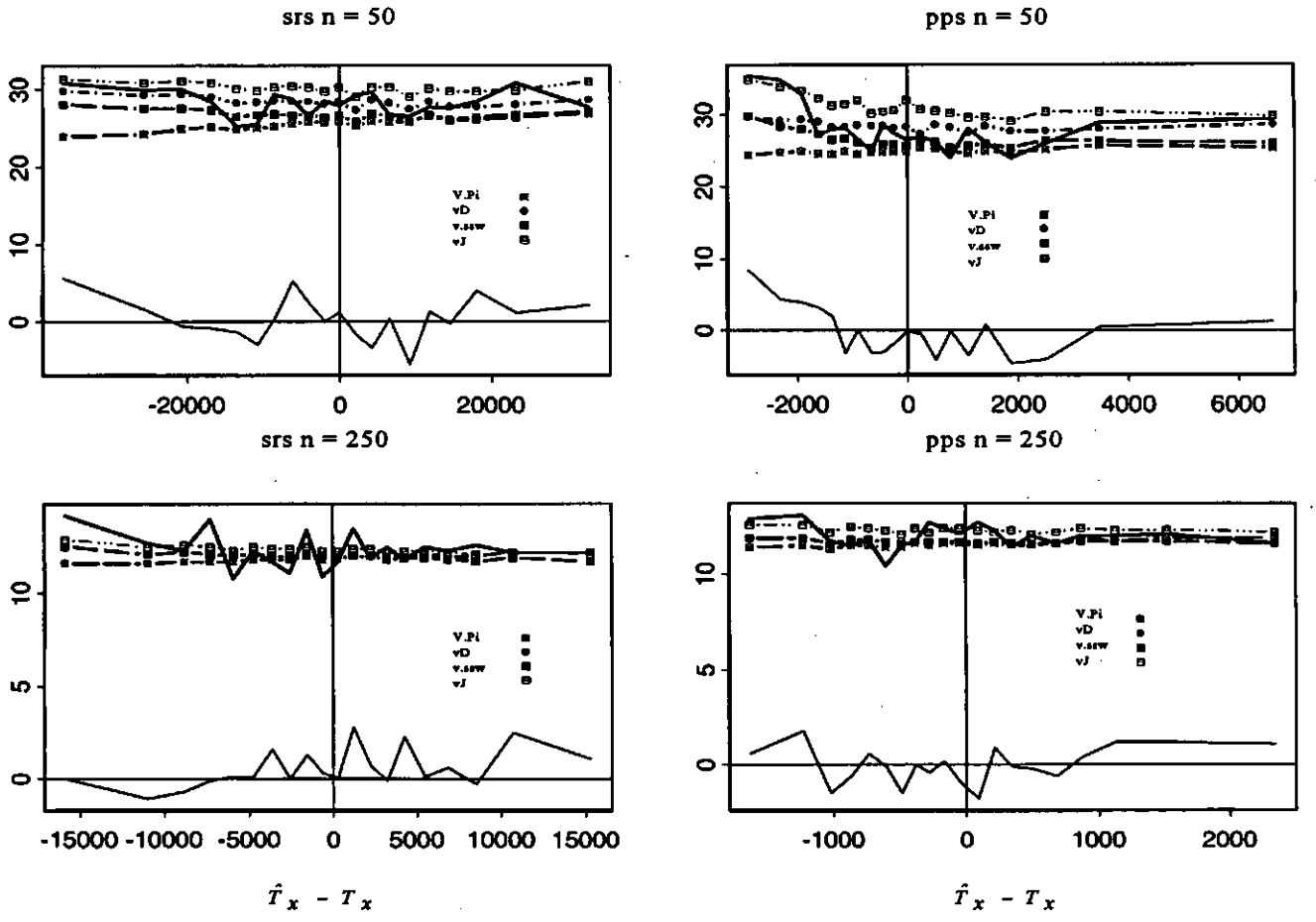


Figure 4. Plot of conditional biases, rmse's, and means of standard error estimates of the GREG for the samples from the Labor Force population. Horizontal and vertical reference lines are drawn at 0. The lowest curve in each panel is the bias of the GREG. The thick solid line is the conditional root mean square error.

Figure 4 is a similar plot for the samples from the Labor Force population. The following sets of estimates are very similar and only the first in each set is included in the plots: (v_D, v_{DP}), and (v_J, v_J^*, v_{JP}). Only the srswor and pps samples of size $n = 50$ and 250 are included. The horizontal axis is again D_x, which is the sum of differences between the π-estimates and the population values of the totals for age and sex groups and the number of hours worked per week. The conditional bias of v_π is evident in samples with the smallest values of D_x but the problem diminishes for the larger sample size in both srswor and pps samples. The jackknife v_J is, on average, the largest of the variance estimators throughout the range of D_x. The differences among the variance estimates and their biases are less for the larger sample size. The estimators v_D, v_{SSW}, and v_J all track the rmse reasonably well except when D_x is most negative, where all are somewhat low.

5. CONCLUSION

A variety of estimators of the variance of the general regression estimator have been proposed in the sampling literature, mainly with the goal of estimating the design-based variance. Estimators can be easily constructed that are approximately unbiased for both the design-variance and, under certain models, the model-variance. Moreover, the dual-purpose estimators studied here are robust estimators of a model-variance even if the model that motivates the GREG has an incorrect variance parameter.

A key feature of the best of these estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. The desirability of using leverage corrections to regression variance estimators in order to combat heteroscedasticity is well-known in econometrics, having been proposed by MacKinnon and White (1985) and recently revisited by Long and Ervin (2000). One of the best choices is an approximation to the jackknife, denoted here by v_{JP}^*, that includes a type of finite population correction.

The robust estimators studied here are quite useful for variables whose distributions are reasonably "well behaved." They adjust variance estimators in small and moderate size samples in a way that often results in better confidence interval coverage. However, they are no defense when variables are extremely skewed, and large observations are not well represented in a sample. Whether one refers to this problem as one of skewness or of outliers, the effect is clear. A sample that does not include a sufficient

number of units with large values will produce an estimated mean that is too small. A variance estimator that is small often accompanies the small estimated mean. As the simulations in section 4 illustrate, in such samples even the best of the proposed variance estimators will not yield confidence intervals that cover at the nominal rate. The transformation methods of Chen and Chen (1996) might hold some promise, but that approach would have to be tested for the more complex GREG estimators studied here.

The most effective solution to the skewness problem does not appear to be to make better use of the sample data. Rather, the sample itself needs to be designed to include good representation of the large units. In many cases, however, like a survey of households to measure income or capital assets, this may be difficult or impossible if auxiliary information closely related to the target variable is not available. Better use of the sample data employing models for skewed variables may then be useful (see, *e.g.*, Karlberg 2000).

ACKNOWLEDGEMENT

The author is indebted to Alan Dorfman whose ideas were the impetus for this work and to the Associate Editor and two referees for their careful reviews.

APPENDIX: Details of Jackknife Calculations

Using (3.2), (3.3), and the standard matrix result in Lemma 5.4.1 of Valliant *et al.* (2000), we have

$$\mathbf{A}_{\pi s(i)}^{-1} = \left[\mathbf{A}_{\pi s}^{-1} + \frac{\mathbf{A}_{\pi s}^{-1}\mathbf{x}_i\mathbf{x}_i'\mathbf{A}_{\pi s}^{-1}/v_i\pi_i}{1-h_{ii}}\right].$$

From this and the definition of $\hat{\mathbf{B}}_{(i)}$, the slope estimator, omitting unit i, is $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} + n^{-1}\sum_s \mathbf{Q}_i$ where

$$\mathbf{Q}_i = \frac{\mathbf{A}_{\pi s}\mathbf{x}_i}{1-h_{ii}}\frac{r_i}{v_i\pi_i}.$$

The GREG estimator, after deleting unit i, is

$$\hat{T}_{G(i)} = \frac{n}{n-1}\left(\hat{T}_\pi - \frac{Y_i}{\pi_i}\right) + \left(\hat{\mathbf{B}} - \mathbf{Q}_i\right)'\left[\mathbf{T}_x - \frac{n}{n-1}\left(\hat{\mathbf{T}}_x - \frac{\mathbf{x}_i}{\pi_i}\right)\right].$$

After some rearrangement, this can be rewritten as

$$\hat{T}_{G(i)} = \frac{n}{n-1}\hat{T}_G - \frac{n}{n-1}\left[\frac{g_i r_i}{\pi_i(1-h_{ii})}\right] + \frac{n}{n-1}G_i + \frac{1}{n-1}K_i$$

where

$$G_i = \frac{h_{ii}Y_i - \hat{Y}_i}{\pi_i(1-h_{ii})}$$

and

$$K_i = (\hat{\mathbf{B}} - \mathbf{Q}_i)'\left(\frac{n\mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x\right).$$

It follows that $\hat{T}_{G(i)} - \hat{T}_{G(\cdot)} = -n(n-1)^{-1}(D_i - \bar{D}_s) + n(n-1)^{-1}F_i$ where $F_i = (G_i - \bar{G}_s) + n^{-1}(K_i - \bar{K}_s)$ with $\bar{G}_s$ and $\bar{K}_s$ being sample means with the obvious definitions. Substituting in the jackknife formula (3.1) gives

$$v_J\left(\hat{\bar{Y}}_G\right) = N^{-2}\frac{n}{n-1}\times$$

$$\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2\sum_s F_i(D_i - \bar{D}_s)\right]. \quad \text{(A.1)}$$

Formula (A.1) is exact, but with some further approximations we can get the relative sizes of the terms. Using the values of G_i and K_i above and the fact that h_{ii} and the elements of $\mathbf{Q}_i$ are $o(1)$, we have

$$G_i + n^{-1}K_i = \frac{h_{ii}Y_i - \hat{Y}_i}{\pi_i(1-h_{ii})} + \frac{1}{n}(\hat{\mathbf{B}} - \mathbf{Q}_i)'\left(\frac{n\mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x\right)$$

$$\approx -\frac{\hat{Y}_i}{\pi_i} + \hat{\mathbf{B}}'\frac{\mathbf{x}_i}{\pi_i} - \frac{1}{n}\hat{\mathbf{B}}'\hat{\mathbf{T}}_x$$

$$= -\frac{1}{n}\hat{\mathbf{B}}'\hat{\mathbf{T}}_x$$

where $\approx$ denotes "asymptotically equivalent to." It follows that $F_i \approx 0$ and that $v_J(\hat{\bar{Y}}_G) \approx \sum_s (D_i - \bar{D}_s)^2$, *i.e.*, (3.5) holds.

Next, we can show that the second term in (3.5) converges in probability to zero. The vector of residuals can be expressed as $\mathbf{r}_s = (\mathbf{I} - \mathbf{H}_s)\mathbf{Y}_s$, and the second term in (3.5) is equal to $N^{-2}n^{-1}\mathbf{g}_s'\boldsymbol{\Pi}_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s\mathbf{r}_s'\mathbf{U}^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{g}_s$ where $\mathbf{U} = \text{diag}(1-h_{ii})$, $i \in s$. Thus, the second term in (3.5) is the square of $B = N^{-1}n^{-1/2}\mathbf{g}_s'\boldsymbol{\Pi}_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s$ which has expectation zero under any model with $E_M(r_i) = 0$. The model-variance of B is

$$N^{-2}n^{-1}\,\text{var}_M(\mathbf{g}_s'\boldsymbol{\Pi}_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s) =$$

$$N^{-2}n^{-1}\mathbf{g}_s'\boldsymbol{\Pi}_s^{-1}\mathbf{U}^{-1}(\mathbf{I}-\mathbf{H}_s)\times \quad \text{(A.2)}$$

$$\mathbf{V}_s(\mathbf{I}-\mathbf{H}_s)'\mathbf{U}^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{g}_s$$

which has order of magnitude n^{-2} under the assumptions we have made. Consequently, the second term in (3.5) is the square of a term with mean zero and a model-variance that approaches zero as the sample size increases. The second term in (3.5) then converges to zero by Chebyshev's inequality. This justifies (3.6).

REFERENCES

BELSLEY, D.A., KUH, E. and WELSCH, R.E. (1980). *Regression Diagnostics*. New York: John Wiley & Sons, Inc.

BREWER, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 in *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College, and P.S. Kott). New York: John Wiley & Sons, Inc., 589-606.

BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 205-212.

CHEN, G., and CHEN, J. (1996) A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.

DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.

HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1980). SUPERCARP. Department of Statistics. Ames, Iowa: Iowa State University.

KARLBERG, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-243.

KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.

KREWSKI 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.

LONG, J.S., and ERVIN, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54, 217-224.

MACKINNON, J.G., and WHITE, H. (1985). Some heteroskedastic consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 53-57.

RAO, J.N.K., and WU, C.J.F. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÄRNDAL, C.-E., and WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandanavian Journal of Statistics*, 11, 146-156.

STUKEL, D., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.

SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

VALLIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

YUNG, W., and RAO, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 18, No. 1, 2002

Evaluating Socio-economic (SES) Bias in Survey Nonresponse
John Goyder, Keith Warriner, and Susan Miller 1

Are Nonrespondents to Health Surveys Less Healthy than Respondents
G. Cohen and J.C. Duffy 13

Accounting for Biases in Election Surveys; The Case of the 1998 Quebec Election
Claire Durand, Andre Blais, and Sebastien Vachon 25

Small Area Estimation via Generalized Linear Models
Alastair Noble, Stephen Haslett, and Greg Arnold 45

Generalized Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI)
Jan de Haan 61

Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation
Risto Lehtonen, Erkki Pahkinen, and Carl-Erik Särndal 87

Book and Software Reviews 111

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 30, No. 1, March/mars 2002, 1-176

Richard A. LOCKHART
Editor's Report/Rapport du rédacteur en chef 1

Yong YOU and J.N.K. RAO
Small area estimation using unmatched sampling and linking models 3

Debbie J. DUPUIS and Stephan MORGENTHALER
Robust weighted likelihood estimators with an application to bivariate extreme value problems 17

Patrick BÉLISLE, Lawrence JOSEPH, David B. WOLFSON and Xiaojie ZHOU
Bayesian estimation of cognitive decline in patients with Alzheimer's disease 37

Joseph G. IBRAHIM, Ming-Hui CHEN and Stuart R. LIPSITZ
Bayesian methods for generalized linear models with covariates missing at random 55

Mario TROTTINI and Fulvio SPEZZAFERRI
A generalized predictive criterion for model selection 79

Pamela OHMAN-STRICKLAND and George CASELLA
Approximate and estimated saddlepoint approximations 97

Yong B. LIM, Jerome SACKS, W.J. STUDDEN and William J. WELCH
Design and analysis of computer experiments when the output is highly correlated over the input space 109

Boxin TANG, Fengshi MA, Debra INGRAM and Hong WANG
Bounds on the maximum number of clear two-factor interactions for 2^{m-p} designs of resolution III and IV 127

Case study in data analysis: The genetic analysis of inflammatory bowel disease 137

Lucia MIREA, Shelley B. BULL, Mark S. SILVERBERG and Katherine A. SIMINOVITCH
Introduction and Analysis 1: The genetic analysis of a complex disease 138

Jiahua CHEN, John D. KALBFLEISCH and Sandra ROMERO-HIDALGO
Analysis 2: Genetic data analysis of affected sib pairs 145

Gerarda A. DARLINGTON and Andrew D. PATERSON
Analysis 3: Genetic analysis of chromosome 6 in inflammatory bowel disease 152

Nicole M. ROSLIN, J.C. LOREDO-OSTI, Celia M.T. GREENWOOD and Kenneth MORGAN
Analysis 4: Genetic analysis of the role of the HLA region in inflammatory bowel disease 158

Christopher A. FIELD and Bruce SMITH
Discussion of the evaluation of a candidate genetic locus in a genome scan of complex disease 167

Acknowledgement of referees' service/Remerciements aux membres des jurys 175

Forthcoming Papers/Articles à paraître 176

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ x 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.