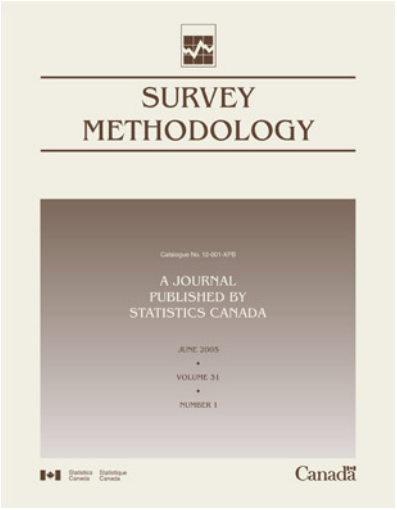
Catalogue no. 12-001-XIE

Survey Methodology

December 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Ordering and subscription information

This product, catalogue no. 12-001-XIE, is published twice a year in electronic format at a price of CAN$23.00 per issue and CAN$44.00 for a one-year subscription. To obtain a single issue or to subscribe, visit our website at www.statcan.ca and select Our Products and Services.

This product, catalogue no. 12-001-XPB, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription. The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$15.00	CAN$30.00

All prices exclude sales taxes.

The printed version of this publication can be ordered

- by phone (Canada and United States) 1 800 267-6677
- by fax (Canada and United States) 1 877 287-4369
- by e-mail infostats@statcan.ca
- by mail Statistics Canada
  Finance Division
  R.H. Coats Bldg., 6th Floor
  120 Parkdale Avenue
  Ottawa, ON K1A 0T6
- In person from authorised agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

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

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2005 • VOLUME 31 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

February 2006

Catalogue no. 12-001-XIE

Frequency: Semi-annual

ISSN 1492-0921

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 articles in English or French in electronic form, preferably in Word to the Editor, (rte@statcan.ca, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 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 statisticiens du Québec.

Survey Methodology

A journal Published by Statistics Canada

Volume 31, Number 2, December 2005

Contents


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 111-112
Statistics Canada, Catalogue No. 12-001-XPB

In This Issue

It is with great sadness that we note the recent passing of M.P. Singh, Editor of the *Survey Methodology* journal since the very first issue in 1975. This issue of the journal opens with a brief obituary in memoriam.

This issue of *Survey Methodology* also contains the fifth paper in the annual invited paper series in honour of Joseph Waksberg. A short biography of Joseph Waksberg was given in the June 2001 issue of the journal, along with the first paper in the series. I would like to thank the members of the selection committee- Michael Brick, chair, David Bellhouse, Gordon Brackstone and Paul Biemer – for having selected Jon Rao as the author of this year's Waksberg paper.

In his paper entitled "Interplay Between Sample Survey Theory and Practice: An Appraisal", Rao traces how survey methods are stimulated by new theoretical developments, and how theory is challenged by survey practice. After summarizing fifty years of contributions from 1920 to 1970, he presents more detailed discussions of more recent developments in several areas. Finally, he discusses several examples of important theory that is not yet widely applied in practice.

In their paper, Fuller and Kim develop and study an efficient hot-deck imputation method under the assumption that response probabilities are equal within imputation cells. Their proposed method is based on the idea of fractional imputation and uses regression techniques to obtain an approximation of the fully efficient version of fractional imputation. Variance estimation is developed for replication methods. Their proposed method is shown to work well in a simulation study.

The paper by Brick, Jones, Kalton and Valliant compares through a simulation study three variance estimation methods in the presence of hot-deck imputation: the model-assisted method, the adjusted jackknife method and multiple imputation. The goal of the simulation study is to study the properties of these variance estimators when their underlying assumptions do not hold. They found that the coverage rate of confidence intervals is not close to the nominal level when the point estimates are biased due failure to take into account the domains of interest at the imputation stage. They conclude by noting that the differences between the variance estimators were too small and inconsistent to support claims that any one of them is superior in general.

Little and Vartivarian study the effect of nonresponse weighting on the Mean Squared Error (MSE) of a population mean estimator. Nonresponse weighting adjustments are obtained by adjusting design weights by the inverse of response rates within cells. They come to the conclusion that a covariate must have two characteristics to reduce nonresponse bias: it needs to be related to both the probability of response and to the survey outcome. If the latter is true, nonresponse weighting can also reduce nonresponse variance. Estimates of the MSE are proposed and used to define a composite estimator. This composite estimator worked well when evaluated in a simulation study.

O'Malley and Zaslavsky present generalized variance-covariance modeling functions (GVCFs) for multivariate means of ordinal survey items, for both complete data and data with structured non-response. After developing and evaluating their methods, they give an illustration using data from the Consumer Assessments of Health Plans Study. In the concluding section they discuss some issues related to the application of GVCFs.

The paper by Singh, Shukla and Kundu develops spatial and spatial-temporal models for small area estimation, as well as estimation of the MSE of the resulting EBLUPs. The models are applied to monthly per capita consumption expenditure data, and they conclude that the models can be very effective when there are significant correlations due to neighborhood effects.

Belsby, Bjørnstad and Zhang discuss modeling to estimate the number of households of different sizes when there is nonignorable nonresponse. They model the response mechanism conditional on household size, using registered family size as supplementary data. After developing their modeling approach, they produce and evaluate estimates using data from the 1992 Norwegian Consumer Expenditure Survey.

Nandram, Cox and Choi consider an analysis for categorical data from a single two-way table with both item and unit nonresponse or, in their terminology, partial classification. They propose to use a Bayesian approach for modeling different patterns of missingness under ignorability and non-ignorability assumptions. The methods are illustrated using incompletely-observed bivariate data from the National Health and Nutrition Examination Survey where the variables subject to missingness are bone mineral density and family income.

In the first of three short notes in this issue, Beaumont discusses the use of data collection process information in nonresponse weight adjustment. He then presents an example from the Canadian Labour Force Survey using the number of attempts to contact a survey unit. An important result is that if the collection process information can be treated as random, then this approach does not introduce any bias.

Starting from basic principles, Bustos derives an explicit form for the probability function of an ordered sample. Using this function, he shows how it can be used to compute inclusion probabilities with illustrations for common sample designs. Finally, he gives the general form for the correlation matrix of sample units, which depends solely on the inclusion probabilities.

Finally, the paper by Wu briefly reviews some theory about the Pseudo Empirical Likelihood (PEL) method in survey sampling, and presents algorithms for computing maximum PEL estimators and for constructing PEL ratio confidence intervals. Functions using the statistical software R and S-PLUS are given to help implement these algorithms in real surveys or in simulation studies.

Harold Mantel

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 113
Statistics Canada, Catalogue No. 12-001-XPB

In Memoriam
M.P. Singh
(1941-2005)

Dr. Mangala P. Singh was born in India on December 26th, 1941 and received his PhD in 1969 from the Indian Statistical Institute, with a specialization in survey sampling. He joined Statistics Canada in 1970, where he rose to the position of Director of Household Survey Methods Division in 1994, a position he held at his death on August 24th, 2005.

M.P., as he was known to everyone, was a leading figure in the application of statistical methods at Statistics Canada. He was probably most closely associated with the Labour Force Survey, one of the agency's most important surveys. He directed the methodology of the LFS through redesigns in the 1970s, 1980s, 1990s and early 21st century, introducing innovations at every turn, but always ensuring that changes were well-tested and sound. In the later years of his career, he also oversaw the development of several new and innovative health surveys and directed the development of statistical programs in the areas of household expenditures, education and justice.

M.P.'s role as the Editor-in-Chief of the journal *Survey Methodology* had a transformative effect on the profession of survey methodology, both in Canada and abroad. M.P. was the founding editor of the journal, and for 30 years he guided its evolution into a flagship publication of Statistics Canada. Thanks to his ability to attract a stellar team of associate editors and contributors, *Survey Methodology* is now recognized as one of the pre-eminent journals of its kind in the world. Even in recent years, M.P. continued to introduce innovations such as the Waksberg series of papers and electronic publishing.

M.P. was a source of many other "big ideas" throughout his career at Statistics Canada. During the 1970s he was instrumental in gaining support for the idea of stable funding for methodology research, and he personally chaired the Methodology Research and Development Committee in its formative years. He encouraged numerous researchers and went out of his way to make them feel at home at Statistics Canada. Turning 60 did not stem the flow of ideas in any way. M.P. devoted considerable energy in the past four years to his proposal for a major overhaul of the way household surveys are conducted in Canada. As a result of his efforts, people throughout Statistics Canada are working on ways to implement his vision, and his influence on Canada's household surveys will be felt for many years.

M.P. had a special love for statistical research and for statistics as a profession. He personally authored over 40 papers in international journals, co-edited two books published by Wiley and Sons, and organized sessions and presented papers at numerous statistical conferences. He served on various committees and task forces of the Statistical Society of Canada, the International Statistical Institute and the American Statistical Association. He also served as Secretary of Statistics Canada's external Advisory Committee on Statistical Methods. In turn, the profession honoured him; he was elected to the International Statistical Institute in 1975, and in 1988 he became a Fellow of the American Statistical Association.

However it is his influence on an entire generation of statisticians that may be his greatest legacy. He was a mentor, a coach, a patriarch and a friend to all who knew him. He inspired others to give their best, and they did. He was always ready with a laugh, a smile and a friendly word of encouragement. He dedicated his life to the profession of statistics and it is through those whom he touched that his true contribution is measured.

He is survived by his wife Savitri, his two daughers Mala and Mamta, and his son Rahul.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 115-116
Statistics Canada, Catalogue No. 12-001-XPB

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work. The author receives a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association. Previous winners were Gad Nathan, Wayne Fuller, Tim Holt, Norman Bradburn, Jon Rao, and Alastair Scott. The first five papers in the series have already appeared in *Survey Methodology*.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)

Nominations:

The author of the 2007 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Gordon Brackstone, 78 Charing Road, Ottawa, Ontario, Canada, K2G 4C9, by email to Gordon.brackstone@sympatico.ca or by fax 1-613-951-1394. Nominations and suggestions for topics must be received by February 28, 2006.

2005 Waksberg Invited Paper

Author: J.N.K. Rao

J.N.K. Rao is Distinguished Research Professor at Carleton University, Ottawa. He has published many articles on a wide range of topics in survey sampling theory and methods and he is the author of the 2003 Wiley book "Small Area Estimation". His research interests in survey sampling include analysis of survey data, small area estimation, missing data and imputation, re-sampling methods and empirical likelihood inference. His 1981 JASA paper (with A.J. Scott) on analysis of survey data was selected as a landmark paper in survey sampling theory and methods. He has been a Member of the Advisory Committee on Statistical Methods of Statistics Canada since its inception 20 years ago. He is a Fellow of the Royal Society of Canada and received the 1994 Gold Medal of the Statistical Society of Canada.

Members of the Waskberg Paper Selection Committee (2005-2006)

Gordon Brackstone, (Chair)
Wayne Fuller, *Iowa State University*
Sharon Lohr, *Arizona State University*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 117-138
Statistics Canada, Catalogue No. 12-001-XPB

Interplay Between Sample Survey Theory and Practice: An Appraisal

J.N.K. Rao [1]

Abstract

A large part of sample survey theory has been directly motivated by practical problems encountered in the design and analysis of sample surveys. On the other hand, sample survey theory has influenced practice, often leading to significant improvements. This paper will examine this interplay over the past 60 years or so. Examples where new theory is needed or where theory exists but is not used will also be presented.

Key Words: Analysis of survey data; Early contributions; Inferential issues; Re-sampling methods; Small area estimation.

1. Introduction

In this paper I will examine the interplay between sample survey theory and practice over the past 60 years or so. I will cover a wide range of topics: early landmark contributions that have greatly influenced practice, inferential issues, calibration estimation that ensures consistency with user specified totals of auxiliary variables, unequal probability sampling without replacement, analysis of survey data, the role of resampling methods, and small area estimation. I will also present some examples where new theory is needed or where theory exists but is not used widely.

2. Some Early Landmark Contributions: 1920 – 1970

This section gives an account of some early landmark contributions to sample survey theory and methods that have greatly influenced the practice. The Norwegian statistician A.N. Kiaer (1897) is perhaps the first to promote sampling (or what was then called "the representative method") over complete enumeration, although the oldest reference to sampling can be traced back to the great Indian epic Mahabharata (Hacking 1975, page 7). In the representative method the sample should mirror the parent finite population and this may be achieved either by balanced sampling through purposive selection or by random sampling. The representative method was used in Russia as early as 1900 (Zarkovic 1956) and Wright conducted sample surveys in the United States around the same period using this method. By the 1920s, the representative method was widely used, and the International Statistical Institute played a prominent role by creating a committee in 1924 to report on the representative method. This committee's report discussed theoretical and practical aspects of the random sampling method. Bowley's (1926) contribution to this report includes his fundamental work on stratified random sampling with proportional allocation, leading to a representative sample with equal inclusion probabilities. Hubback (1927) recognized the need for random sampling in crop surveys: "The only way in which a satisfactory estimate can be found is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal limitations of the experimenter but also makes it possible to say what is the probability with which the results of a given number of samples will be within a given range from the mean. To put this into definite language, it should be possible to find out how many samples will be required to secure that the odds are at least 20:1 on the mean of the samples within one maund of the true mean". This statement contains two important observations on random sampling: (1). It avoids personal biases in sample selection. (2). Sample size can be determined to satisfy a specified margin of error apart from a chance of 1 in 20. Mahalanobis (1946b) remarked that R.A. Fisher's fundamental work at Rothamsted Experimental Station on design of experiments was influenced directly by Hubback (1927).

Neyman's (1934) classic landmark paper laid the theoretical foundations to the probability sampling (or design-based) approach to inference from survey samples. He showed, both theoretically and with practical examples, that stratified random sampling is preferable to balanced sampling because the latter can perform poorly if the underlying model assumptions are violated. Neyman also introduced the ideas of efficiency and optimal allocation in his theory of stratified random sampling without replacement by relaxing the condition of equal inclusion probabilities. By generalizing the Markov theorem on least squares estimation, Neyman proved that the stratified mean, $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, is the best estimator of the population mean, $\bar{Y} = \sum_h W_h \bar{Y}_h$, in the linear class of unbiased estimators of the form $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, where $W_h, \bar{y}_h$ and $\bar{Y}_h$ are

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

the h^{th} stratum weight, sample mean and population mean $(h = 1, ..., L)$, and b_{hi} is a constant associated with the item value y'_{hi} observed on the i^{th} sample draw $(i = 1, ..., n_h)$ in the h^{th} stratum. Optimal allocation $(n_1, ..., n_L)$ of the total sample size, n, was obtained by minimizing the variance of $\bar{y}_{st}$ subject to $\sum_h n_h = n$; an earlier proof of Neyman allocation by Tschuprow (1923) was later discovered. Neyman also proposed inference from larger samples based on normal theory confidence intervals such that the frequency of errors in the confidence statements based on all possible stratified random samples that could be drawn does not exceed the limit prescribed in advance *"whatever the unknown properties of the population"*. Any method of sampling that satisfies the above frequency statement was called "representative". Note that Hubback (1927) earlier alluded to the frequency statement associated with the confidence interval. Neyman's final contribution to the theory of sample surveys (Neyman 1938) studied two-phase sampling for stratification and derived the optimal first phase and second phase sample sizes, n' and n, by minimizing the variance of the estimator subject to a given cost $C = n'c' + nc$, where the second phase cost per unit, c, is large relative to the first phase cost per unit, c'.

The 1930's saw a rapid growth in demand for information, and the advantages of probability sampling in terms of greater scope, reduced cost, greater speed and model-free features were soon recognized, leading to an increase in the number and type of surveys taken by probability sampling and covering large populations. Neyman's approach was almost universally accepted by practicing survey statisticians. Moreover, it inspired various important extensions, mostly motivated by practical and efficiency considerations. Cochran's (1939) landmark paper contains several important results: the use of ANOVA to estimate the gain in efficiency due to stratification, estimation of variance components in two-stage sampling for future studies on similar material, choice of sampling unit, regression estimation under two-phase sampling and effect of errors in strata sizes. This paper also introduced the super-population concept: "The finite population should itself be regarded as a random sample from some infinite population". It is interesting to note that Cochran at that time was critical of the traditional fixed population concept: "Further, it is far removed from reality to regard the population as a fixed batch of known numbers". Cochran (1940) introduced ratio estimation for sample surveys, although an early use of the ratio estimator dates back to Laplace (1820). In another landmark paper (Cochran 1942), he developed the theory of regression estimation. He derived the conditional variance of the usual regression estimator for a fixed sample and also a sample estimator of this variance, assuming a linear regression model $y = \alpha + \beta x + e$, where e has mean zero and constant variance in arrays in which x is fixed. He also noted that the regression estimator remains (model) unbiased under non-random sampling, provided the assumed linear regression model is correct. He derived the average bias under model deviations (in particular, quadratic regression) for simple random sampling as the sample size n increased. Cochran then extended his results to weighted regression and derived the now well-known optimality result for the ratio estimator, namely it is a "best unbiased linear estimate if the mean value and variance both change proportional to x". The latter model is called the ratio model in the current literature. Madow and Madow (1944) and Cochran (1946) compared the expected (or anticipated) variance under a super-population model to study the relative efficiency of systematic sampling and stratified random sampling analytically. This paper stimulated much subsequent research on the use of super-population models in the choice of probability sampling strategies, and also for model-dependent and model-assisted inferences (see section 3).

In India, Mahalanobis made pioneering contributions to sampling by formulating cost and variance functions for the design of surveys. His 1944 landmark paper (Mahalanobis 1944) provides deep theoretical results on the efficient design of sample surveys and their practical applications, in particular to crop acreage and yield surveys. The well-known optimal allocation in stratified random sampling with cost per unit varying across strata is obtained as a special case of his general theory. As early as 1937, Mahalanobis used multi-stage designs for crop yield surveys with villages, grids within villages, plots within grids and cuts of different sizes and shapes as sampling units in the four stages of sampling (Murthy 1964). He also used a two-phase sampling design for estimating the yield of cinchona bark. He was instrumental in establishing the National Sample Survey (NSS) of India, the largest multi-subject continuing survey operation with full-time staff using personal interviews for socioeconomic surveys and physical measurements for crop surveys. Several prominent survey statisticians, including D.B. Lahiri and M.N. Murthy, were associated with the NSS.

P.V. Sukhatme, who studied under Neyman, also made pioneering contributions to the design and analysis of large-scale agricultural surveys in India, using stratified multi-stage sampling. Starting in 1942–1943 he developed efficient designs for the conduct of nationwide surveys on wheat and rice crops and demonstrated high degree of precision for state estimates and reasonable margin of error for district estimates. Sukhatme's approach differed from that of Mahalanobis who used very small plots for crop cutting employing *ad hoc* staff of investigators. Sukhatme (1947) and Sukhatme and Panse (1951) demonstrated that

the use of a small plot might give biased estimates due to the tendency of placing boundary plants inside the plot when there is doubt. They also pointed out that the use of an *ad hoc* staff of investigators, moving rapidly from place to place, forces the plot measurements on only those sample fields that are ready for harvest on the date of the visit, thus violating the principle of random sampling. Sukhatme's solution was to use large plots to avoid boundary bias and to entrust crop-cutting work to the local revenue or agricultural agency in a State.

Survey statisticians at the U.S. Census Bureau, under the leadership of Morris Hansen, William Hurwitz, William Madow and Joseph Waksberg, made fundamental contributions to sample survey theory and practice during the period 1940–70, and many of those methods are still widely used in practice. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with probability proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensures self-weighting (equal overall probabilities of selection) within strata. This approach provides approximately equal interviewer work loads which is desirable in terms of field operations. It also leads to significant variance reduction by controlling the variability arising from unequal PSU sizes without actually stratifying by size and thus allowing stratification on other variables to reduce the variance. On the other hand, workloads can vary widely if the PSUs are selected by simple random sampling and then sub-sampled at the same rate within each stratum. PPS sampling of PSUs is now widely used in the design of large-scale surveys, but two or more PSUs are selected without replacement from each stratum such that the PSU inclusion probabilities are proportional to size measures (see section 5).

Many large-scale surveys are repeated over time, such as the monthly Canadian Labour Force Survey (LFS) and the U.S. Current Population Survey (CPS), with partial replacement of ultimate units (also called rotation sampling). For example, in the LFS the sample of households is divided into six rotation groups (panels) and a rotation group remains in the sample for six consecutive months and then drops out of the sample, thus giving five-sixth overlap between two consecutive months. Yates (1949) and Patterson (1950), following the initial work of Jessen (1942) for sampling on two occasions with partial replacement of units, provided the theoretical foundations for design and estimation of repeated surveys, and demonstrated the efficiency gains for level and change estimation by taking advantage of past data. Hansen, Hurwitz, Nisseslson and Steinberg (1955) developed simpler estimators, called $K-$ composite estimators, in the context of stratified multi-stage designs with PPS sampling in the first stage. Rao and Graham (1964) studied optimal replacement policies for the $K-$composite estimators. Various extensions have also been proposed. Composite estimators have been used in the CPS and other continuing large scale surveys. Only recently, the Canadian LFS adopted a type of composite estimation, called regression composite estimation, that makes use of sample information from previous months and that can be implemented with a regression weights program (see section 4).

Keyfitz (1951) proposed an ingenious method of switching to better PSU size measures in continuing surveys based on the latest census counts. His method ensures that the probability of overlap with the previous sample of one PSU per stratum is maximized, thus reducing the field costs and at the same time achieving increased efficiency by using the better size measures in PPS sampling. The Canadian LFS and other continuing surveys have used the Keyfitz method. Raj (1956) formulated the optimization problem as a "transportation problem" in linear programming. Kish and Scott (1971) extended the Keyfitz method to changing strata and size measures. Ernst (1999) has given a nice account of the developments over the past 50 years in sample coordination (maximizing or minimizing the sample overlap) using transportation algorithms and related methods; see also Mach, Reiss and Schiopu-Kratina (2005) for applications to business surveys with births and deaths of firms.

Dalenius (1957, Chapter 7) studied the problem of optimal stratification for a given number of strata, L, under the Neyman allocation. Dalenius and Hodges (1959) obtained a simple approximation to optimal stratification, called the cum $\sqrt{f}$ rule, which is extensively used in practice. For highly skewed populations with a small number of units accounting for a large share of the total Y, such as business populations, efficient stratification requires one take-all stratum $(n_1 = N_1)$ of big units and take-some strata of medium and small size units. Lavallée and Hidiroglou (1988) and Rivest (2002) developed algorithms for determining the strata boundaries using power allocation (Fellegi 1981; Bankier 1988) and Neyman allocation for the take some strata. Statistics Canada and other agencies currently use those algorithms for business surveys.

The focus of research prior to 1950 was on estimating population totals and means for the whole population and large planned sub-populations, such as states or provinces. However, users are also interested in totals and means for unplanned sub-populations (also called domains) such as age-sex groups within a province, and parameters other than totals and means such as the median and other quantiles, for example median income. Hartley (1959) developed a simple, unified theory for domain estimation applicable to any design, requiring only the standard formulae for the estimator of total and its variance estimator, denoted in the

operator notation as $\hat{Y}(y)$ and $v(y)$ respectively. He introduced two synthetic variables ${}_j y_i$ and ${}_j a_i$ which take the values y_i and 1 respectively if the unit i belongs to domain j and equal to 0 otherwise. The estimators of domain total ${}_jY = Y({}_j y)$ and domain size ${}_j N = Y({}_j a)$ are then simply obtained from the formulae for $\hat{Y}(y)$ and $v(y)$ by replacing y_i by ${}_j y_i$ and ${}_j a_i$ respectively. Similarly, estimators of domain means and domain differences and their variance estimators are obtained from the basic formulae for $\hat{Y}(y)$ and $v(y)$. Durbin (1968) also obtained similar results. Domain estimation is now routinely done using Hartley's ingenious method.

For inference on quantiles, Woodruff (1952) proposed a simple and ingenious method of getting a $(1-\alpha)-$level confidence interval under general sampling designs, using only the estimated distribution function and its standard error (see Lohr's (1999) book, pages 311–313). Note that the latter are simply obtained from the formulae for a total by changing y to an indicator variable. By equating the Woodruff interval to a normal theory interval on the quantile, a simple formula for the standard error of the p^{th} quantile estimator may also be obtained as half the length of the interval divided by the upper $\alpha/2-$point of the standard $N(0, 1)$ distribution which equals 1.96 if $\alpha = 0.05$ (Rao and Wu 1987; Francisco and Fuller 1991). A surprising property of the Woodruff interval is that it performs well even when p is small or large and sample size is moderate (Sitter and Wu 2001).

The importance of measurement errors was realized as early as the 1940s. Mahalanobis' (1946a) influential paper developed the technique of interpenetrating sub-samples (called replicated sampling by Deming 1960). This method was extensively used in large-scale sample surveys in India for assessing both sampling and measurement errors. The sample is drawn in the form of two or more independent sub-samples according to the same sampling design such that each sub-sample provides a valid estimate of the total or mean. The sub-samples are assigned to different interviewers (or teams) which leads to a valid estimate of the total variance that takes proper account of the correlated response variance component due to interviewers. Interpenetrating sub-samples increase the travel costs of interviewers, but they can be reduced through modifictions of interviewer assignments. Hansen, Hurwitz, Marks and Mauldin (1951), Sukhatme and Seth (1952) and Hansen, Hurwitz and Bershad (1961) developed basic theories under additive measurement error models, and decomposed the total variance into sampling variance, simple response variance and correlated response variance. The correlated response variance due to interviewers was shown to be of the order k^{-1} regardless of the sample size, where k is the number of interviewers. As a result, it can dominate the total variance if k is not large. The 1950 U.S. Census interviewer variance study showed that this component was indeed large for small areas. Partly for this reason, self-enumeration by mail was first introduced in the 1960 U.S. Census to reduce this component of the variance (Waksberg 1998). This is indeed a success story of theory influencing practice. Fellegi (1964) proposed a combination of interpenetration and replication to estimate the covariance between sampling and response deviations. This component is often neglected in the decomposition of total variance but it could be sizeable in practice.

Yet another early milestone in sample survey methods is the concept of design effect (DEFF) due to Leslie Kish (see Kish 1965, section 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance that would be obtained under simple random sampling of the same size. This concept is especially useful in the presentation and modeling of sampling errors, and also in the analysis of complex survey data involving clustering and unequal probabilities of selection (see section 6).

We refer the reader to Kish (1995), Kruskal and Mosteller (1980), Hansen, Dalenius and Tepping (1985) and O'Muircheartaigh and Wong (1981) for reviews of early contributions to sample survey theory and methods.

3. Inferential Issues

3.1 Unified Design-Based Framework

The development of early sampling theory progressed more or less inductively, although Neyman (1934) studied best linear unbiased estimation for stratified random sampling. Strategies (design and estimation) that appeared reasonable were entertained and relative properties were carefully studied by analytical and/or empirical methods, mainly through comparisons of mean squared errors, and sometimes also by comparing anticipated mean squared errors or variances under plausible super-population models, as noted in section 2. Unbiased estimation under a given design was not insisted upon because it "often results in much larger mean squared error than necessary" (Hansen, Hurwitz and Tepping 1983). Instead, design consistency was deemed necessary for large samples *i.e.*, the estimator approches the population value as the sample size increases. Classical text books by Cochran (1953), Deming (1950), Hansen, Hurwitz and Madow (1953), Sukhatme (1954) and Yates (1949), based on the above approach, greatly influenced survey practice. Yet, academic statisticians paid little attention to traditional sampling theory, possibly because it lacked a formal theoretical framework and was not integrated with mainstream statistical theory. Numerous prestigious statistics departments in North America did not offer graduate courses in sampling theory.

Formal theoretical frameworks and approaches to integrating sampling theory with mainstream statistical inference were initiated in the 1950s under a somewhat idealistic set-up that focussed on sampling errors assuming the absence of measurement or response errors and non-response. Horvitz and Thompson (1952) made a basic contribution to sampling with arbitrary probabilities of selection by formulating three subclasses of linear design-unbiased estimators of a total Y that include the Markov class studied by Neyman as one of the subclasses. Another subclass with design weight d_i attached to a sample unit i and depending only on i admitted the well-known estimator with weight inversely proportional to the inclusion probability π_i as the only unbiased estimator. Narain (1951) also discovered this estimator, so it should be called the Narain-Horvitz-Thompson (NHT) estimator rather than the HT estimator as it is commonly known. For simple random sampling, the sample mean is the best linear unbiased estimator (BLUE) of the population mean in the three subclasses, but this is not sufficient to claim that the sample mean is the best in the class of all possible linear unbiased estimators. Godambe (1955) proposed a general class of linear unbiased estimators of a total Y by recognizing the sample data as $\{(i, y_i), i \in s\}$ and by letting the weight depend on the sample unit i as well as on the other units in the sample s, that is, the weight is of the form $d_i(s)$. He then established that the BLUE does not exist in the general class

$$\hat{Y} = \sum_{i \in s} d_i(s) y_i, \qquad (1)$$

even under simple random sampling. This important negative theoretical result was largely overlooked for about 10 years. Godambe also established a positive result by relating y to a size measure x using a super-population regression model through origin with error variance proportional to x^2, and then showing that the NHT estimator under any fixed sample size design with π_i proportional to x_i minimizes the anticipated variance in the unbiased class (1). This result clearly shows the conditions on the design for the use of the NHT estimator. Rao (1966) recognized the limitations of the NHT estimator in the context of surveys with PPS sampling and multiple characteristics. Here the NHT estimator will be very inefficient when a characteristic y is unrelated or weakly related to the size measure x (such as poultry count y and farm size x in a farm survey). Rao proposed efficient alternative estimators for such cases that ignore the NHT weights. Ignoring the above results, some theoretical criteria were later advanced in the sampling literature to claim that the NHT estimator should be used for *any* sampling design. Using an amusing example of circus elephants, Basu (1971) illustrated the futility of such criteria. He constructed a "bad" design with π_i unrelated to y_i and then demonstrated that the NHT estimator leads to absurd estimates which prompted the famous mainstream Bayesian statistician Dennis Lindley to conclude that this counterexample destroys the design-based sample survey theory (Lindley 1996). This is rather unfortunate because NHT and Godambe clearly stated the conditions on the design for a proper use of the NHT estimator, and Rao (1966) and Hajek (1971) proposed alternative estimators to deal with multiple characteristics and bad designs, respectively. It is interesting to note that the same theoretical criteria led to a bad variance estimator of the NHT estimator as the 'optimal' choice (Rao and Singh 1973).

Attempts were also made to integrate sample survey theory with mainstream statistical inference via the likelihood function. Godambe (1966) showed that the likelihood function from the sample data $\{(i, y_i), i \in s\}$, regarding the $N-$ vector of unknown $y-$values as the parameter, provides no information on the unobserved sample values and hence on the total Y. This uninformative feature of the likelihood function is due to the label property that treats the N population units as essentially N post-strata. A way out of this difficulty is to take the Bayesian route by assuming informative (exchangeable) priors on the parameter vector (Ericson 1969). An alternative route (design-based) is to ignore some aspects of the sample data to make the sample non-unique and thus arrive at an informative likelihood function (Hartley and Rao 1968; Royall 1968). For example, under simple random sampling, suppressing the labels i and regarding the data as $\{y_i, i \in s\}$ in the absence of information relating i to y_i, leads to the sample mean as the maximum likelihood estimator of the population mean. Bayesian estimation, assuming non-informative prior distributions, leads to results similar to Ericson's (1969) but depends on the sampling design unlike Ericson's. In the case y_i is a vector that includes auxiliary variables with known totals, Hartley and Rao (1968) showed that the maximum likelihood estimator under simple random sampling is approximately equal to the traditional regression estimator of the total. This paper was the first to show how to incorporate known auxiliary population totals in a likelihood framework. For stratified random sampling, labels within strata are ignored but not strata labels because of known strata differences. The resulting maximum likelihood estimator is approximately equal to a pseudo-optimal linear regression estimator when auxiliary variables with known totals are available. The latter estimator has some good conditional design-based properties (see section 3.4). The focus of Hartley and Rao (1968) was on the estimation of a total, but the likelihood approach has much wider scope in sampling, including the estimation of distribution functions and quantiles and the construction of likelihood ratio based confidence intervals (see section 8.1). The Hartley-Rao non-parametric likelihood approach was discovered

independently twenty years later (Owen 1988) in the mainstream statistical inference under the name "empirical likelihood". It has attracted a good deal of attention, including its application to various sampling problems. So in a sense the integration efforts with mainstream statistics were partially successful. Owen's (2002) book presents a thorough account of empirical likelihood theory and its applications.

3.2 Model-Dependent Approach

The model-dependent approach to inference assumes that the population structure obeys a specified super-population model. The distribution induced by the assumed model provides inferences referring to the particular sample of units s that has been drawn. Such conditional inferences can be more relevant and appealing than repeated sampling inferences. But model-dependent strategies can perform poorly in large samples when the model is not correctly specified; even small deviations from the assumed model that are not easily detectable through model checking methods can cause serious problems. For example, consider the often-used ratio model when an auxiliary variable x with known total X is also measured in the sample:

$$y_i = \beta x_i + \varepsilon_i ;\ i = 1, \ldots, N \qquad (2)$$

where the ε_i are independent random variables with zero mean and variance proportional to x_i. Assuming the model holds for the sample, that is, no sample selection bias, the best linear model-unbiased predictor of the total Y is given by the ratio estimator $(\bar{y}/\bar{x})X$ regardless of the sample design. This estimator is not design consistent unless the design is self-weighting, for example, stratified random sampling with proportional allocation. As a result, it can perform very poorly in large samples under non-self-weighting designs even if the deviations from the model are small. Hansen *et al.* (1983) demonstrated the poor performance under a repeated sampling set-up, using a stratified random sampling design with near optimal sample allocation (commonly used to handle highly skewed populations). Rao (1996) used the same design to demonstrate poor performance under a conditional framework relevant to the model-dependent approach (Royall and Cumberland 1981). Nevertheless, model-dependent approaches can play a vital role in small area estimation where the sample size in a small area (or domain) can be very small or even zero; see section 7.

Brewer (1963) proposed the model-dependent approach in the context of the ratio model (2). Royall (1970) and his collaborators made a systematic study of this approach. Valliant, Dorfman and Royall (2000) give a comprehensive account of the theory, including estimation of the (conditional) model variance of the estimator which varies with s. For example, under the ratio model (2) the model variance depends on the sample mean $\bar{x}_s$. It is interesting to note that balanced sampling through purposive selection appears in the model-dependent approach in the context of protection against incorrect specification of the model (Royall and Herson 1973).

3.3 Model-Assisted Approach

The model-assisted approach attempts to combine the desirable features of design-based and model-dependent methods. It entertains only design-consistent estimators of the total Y that are also model unbiased under the assumed "working" model. For example, under the ratio model (2), a model-assisted estimator of Y for a specified probability sampling design is given by the ratio estimator $\hat{Y}_r = (\hat{Y}_{\mathrm{NHT}} / \hat{X}_{\mathrm{NHT}})X$ which is design consistent regardless of the assumed model. Hansen *et al.* (1983) used this estimator for their stratified design to demonstrate its superior performance over the model dependent estimator $(\bar{y}/\bar{x})X$. For variance estimation, the model-assisted approach uses estimators that are consistent for the design variance of the estimator and at the same time exactly or asymptotically model unbiased for the model variance. However, the inferences are design-based because the model is used only as a "working" model.

For the ratio estimator $\hat{Y}_r$ the variance estimator is given by

$$\mathrm{var}(\hat{Y}_r) = (X / \hat{X}_{\mathrm{NHT}})^2 v(e), \qquad (3)$$

where in the operator notation $v(e)$ is obtained from $v(y)$ by changing y_i to the residuals $e_i = y_i - (\hat{Y}_{\mathrm{NHT}} / \hat{X}_{\mathrm{NHT}})x_i$. This variance estimator is asymptotically equivalent to a customary linearization variance estimator $v(e)$, but it reflects the fact that the information in the sample varies with $\hat{X}_{\mathrm{NHT}}$: larger values lead to smaller variability and smaller values to larger variability. The resulting normal pivotal leads to valid model-dependent inferences under the assumed model (unlike the use of $v(e)$ in the pivotal) and at the same time protects against model deviations in the sense of providing asymptotically valid design-based inferences. Note that the pivotal is asymptotically equivalent to $\hat{Y}(\tilde{e})/[v(\tilde{e})]^{1/2}$ with $\tilde{e}_i = y_i - (Y/X)x_i$. If the deviations from the model are not large, then the skewness in the residuals $\tilde{e}_i$ will be small even if y_i and x_i are highly skewed, and normal confidence intervals will perform well. On the other hand, for highly skewed populations, the normal intervals based on $\hat{Y}_{\mathrm{NHT}}$ and its standard error may perform poorly under repeated sampling even for fairly large samples because the pivotal depends on the skewness of the y_i. Therefore, the population structure does matter in design-based inferences contrary to the claims of Neyman (1934), Hansen *et al.* (1983) and others. Rao, Jocelyn and Hidiroglou (2003) considered the simple linear regression

estimator under two-phase simple random sampling with only x observed in the first phase. They demonstrated that the coverage performance of the associated normal intervals can be poor even for moderately large second phase samples if the true underlying model that generated the population deviated significantly from the linear regression model (for example, a quadratic regression of y on x) and the skewness of x is large. In this case, the first phase $x-$ values are observed, and a proper model-assisted approach would use a multiple linear regression estimator with x and $z = x^2$ as the auxiliary variables. Note that for single phase sampling such a model-assisted estimator cannot be implemented if only the total X is known since the estimator depends on the population total of z.

Särndal, Swenson and Wretman (1992) provide a comprehensive account of the model-assisted apporach to estimating the total Y of a variable y under the working linear regression model

$$y_i = x_i'\beta + \varepsilon_i\,;\; i = 1,\, ...,\, N \qquad (4)$$

with mean zero, uncorrelated errors ε_i and model variance $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma_i^2$ where the q_i are known constants and the $x-$vectors have known totals X (the population values x_1, ..., x_N may not be known). Under this set-up, the model-assisted approach leads to the generalized regression (GREG) estimator with a closed-form expression

$$\hat{Y}_{gr} = \hat{Y}_{\mathrm{NHT}} + \hat{B}'(X - \hat{X}_{\mathrm{NHT}}) =: \sum_{i\in s} w_i(s)\, y_i, \qquad (5)$$

where

$$\hat{B} = \hat{T}^{-1}\left(\sum_s \pi_i^{-1} x_i\, y_i / q_i\right) \qquad (6)$$

with $\hat{T} = \sum_s \pi_i^{-1} x_i x_i' / q_i$ is a weighted regression coefficient, and $w_i(s) = g_i(s)\pi_i^{-1}$ with $g_i(s) = 1 + (X - \hat{X}_{\mathrm{NHT}})'\hat{T}^{-1} x_i / q_i$, known as "$g-$weights". Note that the GREG estimator (5) can also be written as $\sum_{i\in U}\hat{y}_i + \hat{E}_{\mathrm{NHT}}$, where $\hat{y}_i = x_i'\hat{B}$ is the predictor of y_i under the working model and $\hat{E}_{\mathrm{NHT}}$ is the NHT estimator of the total prediction error $E = \sum_{i\in U} e_i$ with $e_i = y_i - \hat{y}_i$. This representation shows the role of the working model in the model-assisted approach. The GREG estimator (5) is design-consistent as well as model-unbiased under the working model (4). Moreover, it is nearly "optimal" in the sense of minimizing the asymptotic anticipated MSE (model expectation of the design MSE) under the working model, provided the inclusion probability, π_i, is proportional to the model standard deviation σ_i. However, in surveys with multiple variables of interest, the model variance may vary across variables. Because one must use a general-purpose design such as the design with inclusion probabilities proportional to sizes, the optimality result no longer holds, even if the same vector x_i is used for all the variables y_i in the working model.

The GREG estimator simplifies to the 'projection' estimator $X'\hat{B} = \sum_s w_i(s)\, y_i$ with $g_i(s) = X'\hat{T}^{-1} x_i / q_i$ if the model variance σ_i^2 is proportional to $\lambda' x_i$ for some λ. The ratio estimator is obtained as a special case of the projection estimator by letting $q_i = x_i$, leading to $g_i(s) = X/\hat{X}_{\mathrm{HT}}$. Note that the GREG estimator (5) requires only the population totals X and not necessarily the individual population values x_i. This is very useful because the auxiliary population totals are often ascertained from external sources such as demographic projections of age and sex counts. Also, it ensures consistency with the known totals X in the sense of $\sum_s w_i(s) x_i = X$. Because of this property, GREG is also a calibration estimator.

Suppose there are p variables of interest, say $y^{(1)}$, ..., $y^{(p)}$, and we want to use the model-assisted approach to estimate the corresponding population totals $Y^{(1)}$, ..., $Y^{(p)}$. Also, suppose that the working model for $y^{(j)}$ is of the form (4) but requires possibly different $x-$vector $x^{(j)}$ with known total $X^{(j)}$ for each $j = 1,\, ...,\, p$:

$$y_i^{(j)} = x_i^{(j)'}\beta^{(j)} + \varepsilon_i^{(j)},\; i = 1,\, ...,\, N. \qquad (7)$$

In this case, the $g-$weights depend on j and in turn the final weights $w_i(s)$ also depend on j. In practice, it is often desirable to use a single set of final weights for all the p variables to ensure internal consistency of figures when aggregated over different variables. This property can be achieved only by enlarging the $x-$vector in the model (7) to accommodate all the variables $y^{(j)}$, say $\tilde{x}$ with known total $\tilde{X}$ and then using the working model

$$y_i^{(j)} = \tilde{x}_i'\beta^{(j)} + \varepsilon_i^{(j)},\; i = 1,\, ...,\, N. \qquad (8)$$

However, the resulting weighted regression coefficients could become unstable due to possible multicolinearity in the enlarged set of auxiliary variables. As a result, the GREG estimator of $Y^{(j)}$ under model (8) is less efficient compared to the GREG estimator under model (7). Moreover, some of the resulting final weights, say $\tilde{w}_i(s)$, may not satisfy range restrictions by taking either values smaller than 1 (including negative values) or very large positive values. A possible solution to handle this problem is to use a generalized ridge regression estimator of $Y^{(j)}$ that is model-assisted under the enlarged model (Chambers 1996; Rao and Singh 1997).

For variance estimation, the model-assisted approach attempts to used design-consistent variance estimators that are also model-unbiased (at least for large samples) for the conditional model variance of the GREG estimator. Denoting the variance estimator of the NHT estimator of Y by $v(y)$ in an operator notation, a simple Taylor linearization variance estimator satisfying the above property is given by $v(ge)$, where $v(ge)$ is obtained by changing y_i to $g_i(s)e_i$ in the formula for $v(y)$; see Hidiroglou, Fuller

and Hickman (1976) and Särndal, Swenson and Wretman (1989).

In the above discussion, we have assumed a working linear regression model for all the variables $y^{(j)}$. But in practice a linear regression model may not provide a good fit for some of the $y-$variables of interest, for example, a binary variable. In the latter case, logistic regression provides a suitable working model. A general working model that covers logistic regression is of the form $E_m(y_i) = h(x_i'\beta) = \mu_i$, where $h(.)$ could be non-linear; model (5) is a special case with $h(a) = a$. A model-assisted estimator of the total under the general working model is the difference estimator $\hat{Y}_{\text{NHT}} + \sum_U \hat{\mu}_i - \sum_s \pi_i^{-1}\hat{\mu}_i$, where $\hat{\mu}_i = h(x_i'\hat{\beta})$ and $\hat{\beta}$ is an estimator of the model parameter β. It reduces to the GREG estimator (5) if $h(a) = a$. This difference estimator is nearly optimal if the inclusion probability π_i is proportional to σ_i, where σ_i^2 denotes the model variance, $V_m(y_i)$.

GREG estimators have become popular among users because many of the commonly used estimators may be obtained as special cases of (5) by suitable specifications of x_i and q_i. A Generalized Estimation System (GES) based on GREG has been developed at Statistics Canada.

Kott (2005) has proposed an alternative paradigm inference, called the randomization-assisted model-based approach, which attempts to focus on model-based inference assisted by randomization (or repeated sampling). The definition of anticipated variance is reversed to the randomization-expected model variance of an estimator, but it is identical to the customary anticipated variance when the working model holds for the sample, as assumed in the paper. As a result, the choices of estimator and variance estimator are often similar to those under the model-assisted approach. However, Kott argues that the motivation is clearer and "the approach proposed here for variance estimation leads to logically coherent treatment of finite population and small-sample adjustments when needed".

3.4 Conditional Design-Based Approach

A conditional design-based approach has also been proposed. This approach attempts to combine the conditional features of the model-dependent approach with the model-free features of the design-based approach. It allows us to restrict the reference set of samples to a "relevant" subset of all possible samples specified by the design. Conditionally valid inferences are obtained in the sense that the conditional bias ratio (*i.e.*, the ratio of conditional bias to conditional standard error) goes to zero as the sample size increases. Approximately $100(1-\alpha)\%$ of the realized confidence intervals in repeated sampling from the conditional set will contain the unknown total Y.

Holt and Smith (1979) provide compelling arguments in favour of conditional design based inference, even though the discussion was confined to one-way post-stratification of a simple random sample in which case it is natural to make inferences conditional on the realized strata sample sizes. Rao (1992, 1994) and Casady and Valliant (1993) studied conditional inference when only the auxiliary total X is known from external sources. In the latter case, conditioning on the NHT estimator $\hat{X}_{\text{NHT}}$ may be reasonable because it is "approximately" an ancillary statistic when X is known and the difference $\hat{X}_{\text{NHT}} - X$ provides a measure of imbalance in the realized sample. Conditioning on $\hat{X}_{\text{NHT}}$ leads to the "optimal" linear regression estimator which has the same form as the GREG estimator (5) with $\hat{B}$ given by (6) replaced by the estimated optimal value $\hat{B}_{\text{opt}}$ of the regression coefficient which involves the estimated covariance of $\hat{Y}_{\text{NHT}}$ and $\hat{X}_{\text{NHT}}$ and the estimated variance of $\hat{X}_{\text{NHT}}$. This optimal estimator leads to conditionally valid design-based inferences and model-unbiased under the working model (4). It is also a calibration estimator depending only on the total X and it can be expressed as $\sum_{i\in s}\tilde{w}_i(s)y_i$ with weights $\tilde{w}_i(s) = d_i\tilde{g}_i(s)$ and the calibration factor $\tilde{g}_i(s)$ depending only on the total X and the sample $x-$values. It works well for stratified random sampling (commonly used in establishment surveys). However, $\hat{B}_{\text{opt}}$ can become unstable in the case of stratified multistage sampling unless the number of sample clusters minus the number of strata is fairly large. The GREG estimator does not require the latter condition but it can perform poorly in terms of conditional bias ratio and conditional coverage rates, as shown by Rao (1996). The unbiased NHT estimator can be very bad conditionally unless the design ensures that the measure of imbalance as defined above is small. For example, in the Hansen *et al.* (1983) design based on efficient $x-$stratification, the imbalance is small and the NHT estimator indeed performed well conditionally.

Tillé (1998) proposed an NHT estimator of the total Y based on approximate conditional inclusion probabilities given $\hat{X}_{\text{NHT}}$. His method also leads to conditionally valid inferences, but the estimator is not calibrated to X unlike the "optimal" linear regression estimator. Park and Fuller (2005) proposed a calibrated GREG version based on Tillé's estimator which leads to non-negative weights more often than GREG.

I believe practitioners should pay more attention to conditional aspects of design-based inference and seriously consider the new methods that have been proposed.

Kalton (2002) has given compelling arguments for favoring design-based approaches (possibly model-assisted and/or conditional) for inference on finite population descriptive parameters. Smith (1994) named design-based inference as "procedural inference" and argued that

procedural inference is the correct approach for surveys in the public domain. We refer the reader to Smith (1976) and Rao and Bellhouse (1990) for reviews of inferential issues in sample survey theory.

4. Calibration Estimators

Calibration weights $w_i(s)$ that ensure consistency with user-specified auxiliary totals X are obtained by adjusting the design weights $d_i = \pi_i^{-1}$ to satisfy the benchmark constraints $\sum_{i \in s} w_i(s) x_i = X$. Estimators that use calibration weights are called calibration estimators and they use a single set of weights $\{w_i(s)\}$ for all the variables of interest. We have noted in section 3.4 that the model-assisted GREG estimator is a calibration estimator, but a calibration estimator may not be model-assisted in the sense that it could be model-biased under a working model (4) unless the $x-$variables in the model exactly match the variables corresponding to the user-specified totals. For example, suppose the working model suggested by the data is a quadratic in a scalar variable x while the user-specified total is only its total X. The resulting calibration estimator can perform poorly even in fairly large samples, as noted in section 3.3, unlike the model-assisted GREG estimator based on the working quadratic model that requires the population total of the quadratic variables x_i^2 in addition to X.

Post-stratification has been extensively used in practice to ensure consistency with known cell counts corresponding to a post-stratification variable, for example counts in different age groups ascertained from external sources such as demographic projections. The resulting post-stratified estimator is a calibration estimator. Calibration estimators that ensure consistency with known marginal counts of two or more post-stratification variables have also been employed in practice; in particular raking ratio estimators that are obtained by benchmarking to the marginal counts in turn until convergence is approximately achieved, typically in four or less iterations. Raking ratio weights $w_i(s)$ are always positive. In the past, Statistics Canada used raking ratio estimators in the Canadian Census to ensure consistency of 2B–item estimators with known 2A–item counts. In the context of the Canadian Census, Brackstone and Rao (1979) studied the efficiency of raking ratio estimators and also derived Taylor linearization variance estimators when the number of iterations is four or less. Raking ratio estimators have also been employed in the U.S. Current Population Survey (CPS). It may be noted that the method of adjusting cell counts to given marginal counts in a two-way table was originally proposed in the landmark paper by Deming and Stephan (1940).

Unified approaches to calibration, based on minimizing a suitable distance measure between calibration weights and design weights subject to benchmark constraints, have attracted the attention of users due to their ability to accommodate arbitrary number of user-specified benchmark constraints, for example, calibration to the marginal counts of several post-stratification variables. Calibration software is also readily available, including GES (Statistics Canada), LIN WEIGHT (Statistics Netherlands), CALMAR (INSEE, France) and CLAN97 (Statistics Sweden).

A chi-squared distance, $\sum_{i \in s} q_i (d_i - w_i)^2 / d_i$, leads to the GREG estimator (5), where the x–vector corresponds to the user-specified benchmark constraints (BC) and $w_i(s)$ is denoted as w_i for simplicity (Huang and Fuller 1978; Deville and Särndal 1992). However, the resulting calibration weights may not satisfy desirable range restrictions (RR), for example some weights may be negative or too large especially when the number of constraints is large and the variability of the design weights is large. Huang and Fuller (1978) proposed a scaled modified chi-squared distance measure and obtained the calibration weights through an iterative solution that satisfies BC at each iteration. However, a solution that satisfies BC and RR may not exist. Another method, called shrinkage minimization (Singh and Mohl 1996) has the same difficulty. Quadratic programming methods that minimize the chi-squared distance subject to both BC and RR have also been proposed (Hussain 1969) but the feasible set of solutions satisfying both BC and RR can be empty. Alternative methods propose to change the distance function (Deville and Särndal 1992) or drop some of the BC (Bankier, Rathwell and Majkowski 1992). For example, an information distance of the form $\sum_{i \in s} q_i \{w_i \log(w_i / d_i) - w_i + d_i\}$ gives raking ratio estimators with non-negative weights w_i, but some of the weights can be excessively large. "Ridge" weights obtained by minimizing a penalized chi-squared distance have also been proposed (Chambers 1996), but no guarantee that either BC or RR are satisfied, although the weights are more stable than the GREG weights. Rao and Singh (1997) proposed a "ridge shrinkage" iterative method that ensures convergence for a specified number of iterations by using a built-in tolerance specification to relax some BC while satisfying RR. Chen, Sitter and Wu (2002) proposed a similar method.

GREG calibration weights have been used in the Canadian Labour Force Survey and more recently it has been extended to accommodate composite estimators that make use of sample information in previous months, as noted in section 2 (Fuller and Rao 2001; Gambino, Kennedy and Singh 2001; Singh, Kennedy and Wu 2001). GREG-type calibration estimators have also been used for the integration of two or more independent surveys from the

same population. Such estimators ensure consistency between the surveys, in the sense that the estimates from the two surveys for common variables are identical, as well as benchmarking to known population totals (Renssen and Nieuwenbroek 1997; Singh and Wu 1996; Merkouris 2004). For the 2001 Canadian Census, Bankier (2003) studied calibration weights corresponding to the "optimal" linear regression estimator (section 3.3) under stratified random sampling. He showed that the "optimal" calibration method performed better than the GREG calibration used in the previous census, in the sense of allowing more BC to be retained while at the same time allowing the calibration weights to be at least one. The "optimal" calibration weights can be obtained from GES software by including the known strata sizes in the BC and defining the tuning constant q_i suitably. Note that the "optimal" calibration estimator also has desirable conditional design properties (section 3.4). Weighting for the 2001 Canadian census switched from projection GREG (used in the 1996 census) to "optimal" linear regression.

Demnati and Rao (2004) derived Taylor linearization variance estimators for a general class of calibration estimators with weights $w_i = d_i F(x_i'\hat{\lambda})$, where the LaGrange multiplier $\hat{\lambda}$ is determined by solving the calibration constraints. The choice $F(a) = 1 + a$ gives GREG weights and $F(a) = e^a$ leads to raking ratio weights. In the special case of GREG weights, the variance estimator reduces to $v(ge)$ given in section 3.3.

We refer the reader to the Waksberg award paper of Fuller (Fuller 2002) for an excellent overview and appraisal of regression estimation in survey sampling, including calibration estimation.

5. Unequal Probability Sampling Without Replacement

We have noted in section 2 that PPS sampling of PSUs within strata in large-scale surveys was practically motivated by the desire to achieve approximately equal workloads. PPS sampling also achieves significant variance reduction by controlling on the variability arising from unequal PSU sizes without actually stratifying by size. PSUs are typically sampled without replacement such that the PSU inclusion probability, π_i, is proportional to PSU size measure x_i. For example, systematic PPS sampling, with or without initial randomization of the PSU labels, is an inclusion probability proportional to size (IPPS) design (also called πPS design) that has been used in many complex surveys, including the Canadian LFS. The estimator of a total associated with an IPPS design is the NHT estimator.

Development of suitable (IPPS, NHT) strategies raises theoretically challenging problems, including the evaluation of exact joint inclusion probabilities, π_{ij}, or accurate approximations to π_{ij} requiring only the individual π_i s, that are needed in getting unbiased or nearly unbiased variance estimator. My own 1961 Ph.D. thesis at Iowa State University addressed the latter problem. Several solutions, requiring sophisticated theoretical tools, have been published since then by talented mathematical statisticians. However, this theoretical work is often classified as "theory without application" because it is customary practice to treat the PSUs as if sampled with replacement since that leads to great simplification. The variance estimator is simply obtained from the estimated PSU totals and, in fact, this assumption is the basis for re-sampling methods (section 6). This variance estimator can lead to substantial over-estimation unless the overall PSU sampling fraction is small. The latter may be true in many large-scale surveys. In the following paragraphs, I will try to demonstrate that the theoretical work on (IPPS, NHT) strategies as well as some non-IPPS designs have wide practical applicability.

First, I will focus on (IPPS, NHT) strategies. In Sweden and some other countries in Europe, stratified single-stage sampling is often used because of the availability of list frames and IPPS designs are attractive options, but sampling fractions are often large. For example, Rosén (1991) notes that Statistics Sweden's Labour Force Barometer samples some 100 different populations using systematic PPS sampling and that the sampling rates can exceed 50%. Aires and Rosén (2005) studied Pareto πPS sampling for Swedish surveys. This method has attractive properties, including fixed sample size, simple sample selection, good estimation precision and consistent variance estimation regardless of sampling rates. It also allows sample coordination through permanent random numbers (PRN) as in Poisson sampling, but the latter method leads to variable sample size. Because of these merits, Pareto πPS has been implemented in a number of Statistics Sweden surveys, notably in price index surveys. Ohlsson (1995) described PRN techniques that are commonly used in practice.

The method of Rao-Sampford (see Brewer and Hanif 1983, page 28) leads to exact IPPS designs and non-negative unbiased variance estimators for arbitrary fixed sample sizes. It has been implemented in the new version of SAS. Stehman and Overton (1994) note that variable probability structure arises naturally in environmental surveys rather than being selected just for enhanced efficiency, and that the π_i s are only known for the units i in the sample s. By treating the sample design as randomized systematic PPS, Stehman and Overton obtained approximations to the π_{ij} s that depend only π_i, $i \in s$, unlike the original approximations of Hartley and Rao (1962) that require the sum of squares of all the π_i s in the population. In the Stehman and Overton applications, the sampling rates are

substantial enough to warrant the evaluation of the joint inclusion probabilities.

I will now turn to non-IPPS designs using estimators different from the NHT estimator that ensure zero variance when y is exactly proportional to x. The random group method of Rao, Hartley and Cochran (1962) permits a simple non-negative variance estimator for any fixed sample size and yet compares favorably to (IPPS, NHT) strategies in terms of efficiency and is always more efficient than the PPS with replacement strategy. Schabenberger and Gregoire (1994) noted that (IPPS, NHT) strategies have not enjoyed much application in forestry because of difficulty in implementation and recommended the Rao-Hartley-Cochran strategy in view of its remarkable simplicity and good efficiency properties. It is interesting to note that this strategy has been used in the Canadian LFS on the basis of its suitability for switching to new size measures, using the Keyfitz method within each random group. On the other hand, (IPPS, NHT) strategies are not readily suitable for this purpose (Fellegi 1966). I understand that the Rao-Hartley-Cochran strategy is often used in audit sampling and other accounting applications.

Murthy (1957) used a non-IPPS design based on drawing successive units with probabilities $p_i, p_j/(1-p_i)$, $p_k/(1-p_i-p_j)$ and so on, and the following estimator:

$$\hat{Y}_M = \sum_{i \in s} y_i \frac{p(s|i)}{p(s)}, \qquad (9)$$

where $p(s|i)$ is the conditional probability of obtaining the sample s given that unit i was selected first. He also provided a non-negative variance estimator requiring the conditional probabilities, $p(s|i, j)$, of obtaining s given i and j are selected in the first two draws. This method did not receive practical attention for several years due to computational complexity, but more recently it has been applied in unexpected areas, including oil discovery (Andreatta and Kaufmann 1986) and sequential sampling including inverse sampling and some adaptive sampling schemes (Salehi and Seber 1997). It may be noted that adaptive sampling has received a lot of attention in recent years because of its potential as an efficient sampling method for estimating totals or means of rare populations (Thompson and Seber 1996). In the oil discovery application, the successive sampling scheme is a characterization of discovery and the order in which fields are discovered is governed by sampling proportional to field size and without replacement, following the industry folklore "on the average, the big fields are found first". Here $p_i = y_i/Y$ and the total oil reserve Y is assumed to be known from geological considerations. In this application, geologists are interested in the size distribution of all fields in the basin and when a basin is partially explored the sample is composed of magnitudes y_i of discovered deposits. The size distribution function $F(a)$ can be estimated by using Murthy's estimator (9) with y_i replaced by the indicator variable $I(y_i \le a)$. The computation of $p(s|i)$ and $p(s)$, however, is formidable even for moderate sample sizes. To overcome this computational difficulty, Andreatta and Kaufman (1986) used integral representations of these quantities to develop asymptotic expansions of Murthy's estimator, the first few terms of which are easily computable. Similarly, they obtain computable approximations to Murthy's variance estimator. Note that the NHT estimator of $F(a)$ is not feasible here because the inclusion probabilities are functions of all the $y-$values in the population.

The above discussion is intended to demonstrate that a particular theory can have applications in diverse practical areas even if it is not needed in a particular situation, such as large-scale surveys with negligible first stage sampling fractions. Also it shows that unequal probability sampling designs play a vital role in survey sampling, despite Särndal's (1996) contention that simpler designs, such as stratified SRS and stratified Bernoulli sampling, together with GREG estimators should replace strategies based on unequal probability sampling without replacement.

6. Analysis of Survey Data and Resampling Methods

Standard methods of data analysis are generally based on the assumption of simple random sampling, although some software packages do take account of survey weights and provide correct point estimates. However, application of standard methods to survey data, ignoring the design effect due to clustering and unequal probabilities of selection, can lead to erroneous inferences even for large samples. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated, type I error rates of tests of hypotheses can be much bigger than the nominal levels, and standard model diagnostics, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of those problems and emphasized the need for new methods that take proper account of the complexity of data derived from large-scale surveys. Fuller (1975) developed asymptotically valid methods for linear regression analysis, based on Taylor linearization variance estimators. Rapid progress has been made over the past 20 years or so in developing suitable methods. Resampling methods play a vital role in developing methods that take account of survey design in the analysis of data. All one needs is a data file containing the observed data, the final survey weights and the corresponding final weights for each pseudo-replicate generated by the re-sampling method. Software packages that take account of survey weights in

the point estimation of parameters of interest can then be used to calculate the correct estimators and standard errors, as demonstrated below. As a result, re-sampling methods of inference have attracted the attention of users as they can perform the analyses themselves very easily using standard software packages. However, releasing public-use data files with replicate weights can lead to confidentiality issues, such as the identification of clusters from replicate weights. In fact, at present a challenge to theory is to develop suitable methods that can preserve confidentiality of the data. Lu, Brick and Sitter (2004) proposed grouping strata and then forming pseudo-replicates using the combined strata for variance estimation, thus limiting the risk of cluster identification from the resulting public-use data file. Grouping strata and/or PSUs within strata simplifies variance estimation by reducing the number of pseudo-replicates used in variance estimation compared to the commonly used delete-cluster jackknife discussed below. A method of inverse sampling to undo the complex survey data structure and yet provide protection against revealing cluster labels (Hinkins, Oh and Scheuren 1997; Rao, Scott and Benhin 2003) appears promising, but much work on inverse sampling methods remains to be done before it becomes attractive to the user.

Rao and Scott (1981, 1984) made a systematic study of the impact of survey design effect on standard chi-squared and likelihood ratio tests associated with a multi-way table of estimated counts or proportions. They showed that the test statistic is asymptotically distributed as a weighted sum of independent χ_1^2 variables, where the weights are the eigenvalues of a "generalized design effects" matrix. This general result shows that the survey design can have a substantial impact on the type I error rate. Rao and Scott proposed simple first-order corrections to the standard chi-squared statistics that can be computed from published tables that include estimates of design effects for cell estimates and their marginal totals, thus facilitating secondary analyses from published tables. They also derived second order corrections that are more accurate, but require the knowledge of a full estimated covariance matrix of the cell estimates, as in the case of familiar Wald tests. However, Wald tests can become highly unstable as the number of cells in a mult-way table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal levels, unlike the Rao-Scott second order corrections (Thomas and Rao 1987). The first and second order corrections are now known as Rao-Scott corrections and are given as default options in the new version of SAS. Roberts, Rao and Kumar (1987) developed Rao-Scott type corrections to tests for logistic regression analysis of estimated cell proportions associated with a binary response variable. They applied the methods to a two-way table of employment rates from the Canadian LFS 1977 obtained by cross-classifying age and education groups. Bellhouse and Rao (2002) extended the work of Roberts *et al.* to the analysis of domain means using generalized linear models. They applied the methods to domain means from a Fiji Fertility Survey cross-classified by education and years since the woman's first marriage, where a domain mean is the mean number of children ever born for women of Indian race belonging to the domain.

Re-sampling methods in the context of large-scale surveys using stratified multi-stage designs have been studied extensively. For inference purposes, the sample PSUs are treated as if drawn with replacement within strata. This leads to over-estimation of variances but it is small if the overall PSU sampling fraction is negligible. Let $\hat{\theta}$ be the survey-weighted estimator of a "census" parameter of interest computed from the final weights w_i, and let the corresponding weights for each pseudo-replicate r generated by the re-sampling method be denoted by $w_i^{(r)}$. The estimator based on the pseudo-replicate weights $w_i^{(r)}$ is denoted as $\hat{\theta}^{(r)}$ for each $r = 1, ..., R$. Then a re-sampling variance estimator of $\hat{\theta}$ is of the form

$$v(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \qquad (10)$$

for specified coefficients c_r in (10) determined by the re-sampling method.

Commonly used re-sampling methods include (a) delete-cluster (delete-PSU) jackknife, (b) balanced repeated replication (BRR) particularly for $n_h = 2$ PSUs in each stratum h and (c) the Rao and Wu (1988) bootstrap. Jackknife pseudo-replicates are obtained by deleting each sample cluster $r = (hj)$ in turn, leading to jackknife design weights $d_i^{(r)}$ taking the value 0 if the sample unit i is in the deleted cluster, $n_h d_i /(n_h - 1)$ if i is not in the deleted cluster but in the same stratum, and unchanged if i is in a different stratum. The jackknife design weights are then adjusted for unit non-response and post-stratification, leading to the final jackknife weights $w_i^{(r)}$. The jackknife variance estimator is given by (10) with $c_r = (n_h - 1)/n_h$ for $r = (hj)$. The delete-cluster jackknife method has two possible disadvantages: (1) When the total number of sampled PSUs, $n = \sum n_h$, is very large, R is also very large because $R = n$. (2) It is not known if the delete-jackknife variance estimator is design-consistent in the case of non-smooth estimators $\hat{\theta}$, for example the survey-weighted estimator of the median. For simple random sampling, the jackknife is known to be inconsistent for the median or other quantiles. It would be theoretically challenging and practically relevant to find conditions for the consistency of the delete-cluster jackknife variance estimator of a non-smooth estimator $\hat{\theta}$.

BRR can handle non-smooth $\hat{\theta}$, but it is readily applicable only for the important special case of $n_h = 2$ PSUs per stratum. A minimal set of balanced half-samples can be constructed from an $R \times R$ Hadamard matrix by selecting H columns, excluding the column of +1's, where $H + 1 \leq R \leq H + 4$ (McCarthy 1969). The BRR design weights $d_i^{(r)}$ equal $2d_i$ or 0 according as whether or not i is in the half-sample. A modified BRR, due to Bob Fay, uses all the sampled units in each replicate unlike the BRR by defining the replicate design weights as $d_i^{(r)}(\varepsilon) = (1+\varepsilon)d_i$ or $(1-\varepsilon)d_i$ according as whether or not i is in the half-sample, where $0 < \varepsilon < 1$; a good choice of ε is 1/2. The modified BRR weights are then adjusted for non-response and post-stratification to get the final weights $w_i^{(r)}(\varepsilon)$ and the estimator $\hat{\theta}^{(r)}(\varepsilon)$. The modified BRR variance estimator is given by (10) divided by ε^2 and $\hat{\theta}^{(r)}$ replaced by $\hat{\theta}^{(r)}(\varepsilon)$; see Rao and Shao (1999). The modified BRR is particularly useful under independent re-imputation for missing item responses in each replicate because it can use the donors in the full sample to impute unlike the BRR that uses the donors only in the half-sample.

The Rao-Wu bootstrap is valid for arbitrary $n_h (\geq 2)$ unlike the BRR, and it can also handle non-smooth $\hat{\theta}$. Each bootstrap replicate is constructed by drawing a simple random sample of PSUs of size $n_h - 1$ from the n_h sample clusters, independently across the strata. The bootstrap design weights $d_i^{(r)}$ are given by $[n_h/(n_h-1)]m_{hi}^{(r)}d_i$ if i is in stratum h and replicate r, where $m_{hi}^{(r)}$ is the number of times sampled PSU (hi) is selected, $\sum_i m_{hi}^{(r)} = n_h - 1$. The weights $d_i^{(r)}$ are then adjusted for unit non-response and post-stratification to get the final bootstrap weights and the estimator $\hat{\theta}^{(r)}$. Typically, $R = 500$ bootstrap replicates are used in the bootstrap variance estimator (10). Several recent surveys at Statistics Canada have adopted the bootstrap method for variance estimation because of the flexibility in the choice of R and wider applicability. Users of Statistics Canada survey micro data files seem to be very happy with the bootstrap method for analysis of data.

Early work on the jackknife and the BRR was largely empirical (*e.g.*, Kish and Frankel 1974). Krewski and Rao (1981) formulated a formal asymptotic framework appropriate for stratified multi-stage sampling and established design consistency of the jackknife and BRR variance estimators when $\hat{\theta}$ can be expressed as a smooth function of estimated means. Several extensions of this basic work have been reported in the recent literature; see the book by Shao and Tu (1995, Chapter 6). Theoretical support for re-sampling methods is essential for their use in practice.

In the above discussion, I let $\hat{\theta}$ denote the estimator of a "census" parameter. The census parameter θ_C is often motivated by an underlying super-population model and the census is regarded as a sample generated by the model, leading to census estimating equations whose solution is θ_C. The census estimating functions $U_C(\theta)$ are simply population totals of functions $u_i(\theta)$ with zero expectation under the assumed model, and the census estimating equations are given by $U_C(\theta) = 0$ (Godambe and Thompson 1986). Kish and Frankel (1974) argued that the census parameter makes sense even if the model is not correctly specified. For example, in the case of linear regression, the census regression coefficient could explain how much of the relationship between the response variable and the independent variables is accounted by a linear regression model. Noting that the census estimating functions are simply population totals, survey weighted estimators $\hat{U}(\theta)$ from the full sample and $\hat{U}^{(r)}(\theta)$ from each pseudo-replicate are obtained. The solutions of corresponding estimating equations $\hat{U}(\theta) = 0$ and $\hat{U}^{(r)}(\theta) = 0$ give $\hat{\theta}$ and $\hat{\theta}^{(r)}$ respectively. Note that the re-sampling variance estimators are designed to estimate the variance of $\hat{\theta}$ as an estimator of the census parameters but not the model parameters. Under certain conditions, the difference can be ignored but in general we have a two-phase sampling situation, where the census is the first phase sample from the super-population and the sample is a probability sample from the census population. Recently, some useful work has been done on two-phase variance estimation when the model parameters are the target parameters (Graubard and Korn 2002; Rubin-Bleuer and Schiopu-Kratina 2005), but more work is needed to address the difficulty in specifying the covariance structure of the model errors.

A difficulty with the bootstrap is that the solution $\hat{\theta}^{(r)}$ may not exist for some bootstrap replicates r (Binder, Kovacevic and Roberts 2004). Rao and Tausi (2004) used an estimating function (EF) bootstrap method that avoids the difficulty. In this method, we solve $\hat{U}(\theta) = \hat{U}^{(r)}(\hat{\theta})$ for θ using only one step of the Newton-Raphson iteration with $\hat{\theta}$ as the starting value. The resulting estimator $\tilde{\theta}^{(r)}$ is then used in (10) to get the EF bootstrap variance estimator of $\hat{\theta}$ which can be readily implemented from the data file providing replicate weights, using slight modifications of any software package that accounts for survey weights. It is interesting to note that the EF bootstrap variance estimator is equivalent to a Taylor linearization sandwich variance estimator that uses the bootstrap variance estimator of $\hat{U}(\theta)$ and the inverse of the observed information matrix (derivative of $-\hat{U}(\theta)$), both evaluated at $\theta = \hat{\theta}$ (Binder *et al.* 2004).

Taylor linearization methods provide asymptotically valid variance estimators for general sampling designs, unlike re-sampling methods, but they require a separate formula for each estimator $\hat{\theta}$. Binder (1983), Rao, Yung and Hidiroglou (2002) and Demnati and Rao (2004) have provided unified linearization variance formulae for estimators defined as solutions to estimating equations.

Pfeffermann (1993) discussed the role of design weights in the analysis of survey data. If the population model holds for the sample (*i.e.*, if there is no sample selection bias), then model-based unweighted estimators will be more efficient than the weighted estimators and lead to valid inferences, especially for data with smaller sample sizes and larger variation in the weights. However, for typical data from large-scale surveys, the survey design is informative and the population model may not hold for the sample. As a result, the model-based estimators can be seriously biased and inferences can be erroneous. Pfeffermann and his colleagues initiated a new approach to inference under informative sampling; see Pfeffermann and Sverchkov (2003) for recent developments. This approach seems to provide more efficient inferences compared to the survey weighted approach, and it certainly deserves the attention of users of survey data. However, much work remains to be done, especially in handling data based on multi-stage sampling.

Excellent accounts of methods for analysis of complex survey data are given in Skinner, Holt and Smith (1989), Chambers and Skinner (2003) and Lehtonen and Pahkinen (2004).

7. Small Area Estimaton

Previous sections of this paper have focussed on traditional methods that use direct domain estimators based on domain-specific sample observations along with auxiliary population information. Such methods, however, may not provide reliable inferences when the domain sample sizes are very small or even zero for some domains. Domains or sub-populations with small or zero sample sizes are called small areas in the literature. Demand for reliable small area statistics has greatly increased in recent years because of the growing use of small area statistics in formulating policies and programs, allocation of funds and regional planning. Clearly, it is seldom possible to have a large enough overall sample size to support reliable direct estimates for all domains of interest. Also, in practice, it is not possible to anticipate all uses of survey data and "the client will always require more than is specified at the design stage" (Fuller 1999, page 344). In making estimates for small areas with adequate level of precision, it is often necessary to use "indirect" estimators that borrow information from related domains through auxiliary information, such as census and current administrative data, to increase the "effective" sample size within the small areas.

It is now generally recognized that explicit models linking the small areas through auxiliary information and accounting for residual between – area variation through random small area effects are needed in developing indirect estimators. Success of such model-based methods heavily depends on the availability of good auxiliary information and thorough validation of models through internal and external evaluations. Many of the random effects methods used in mainstream statistical theory are relevant to small area estimation, including empirical best (or Bayes), empirical best linear unbiased prediction and hierarchical Bayes based on prior distributions on the model parameters. A comprehensive account of such methods is given in Rao (2003). Practical relevance and theoretical interest of small area estimation have attracted the attention of many researchers, leading to important advances in point and mean squared error estimation. The "new" methods have been applied successfully worldwide to a variety of small area problems. Model-based methods have been recently used to produce county and school district estimates of poor school-age children in the U.S.A. The U.S. Department of Education allocates annually over $7 billion of funds to counties on the basis of model-based county estimates. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. We refer to Rao (2003, example 7.1.2) for details of this application. In the United Kingdom, the Office for National Statistics established a Small Area Estimation Project to develop model-based estimates at the level of political wards (roughly 2,000 households). The practice and estimation methods of U.S. federal statistical programs that use indirect estimators to produce published estimates are documented in Schaible (1996). Singh, Gambino and Mantel (1994) and Brackstone (2002) discuss some practical issues and strategies for small area statistics.

Small area estimation is a striking example of the interplay between theory and practice. The theoretical advances are impressive, but many practical issues need further attention of theory. Such issues include: (a) Benchmarking model-based estimators to agree with reliable direct estimators at large area levels. (b) Developing and validating suitable linking models and addressing issues such as errors in variables, incorrect specification of the linking model and omitted variables. (c) Development of methods that satisfy multiple goals: good area-specific estimates, good rank properties and good histogram for small areas.

8. Some Theory Deserving Attention of Practice and Vice Versa

In this section, I will briefly mention some examples of important theory that exists but not widely used in practice.

8.1 Empirical Likelihood Inference

Traditional sampling theory largely focused on point estimation and associated standard errors, appealing to normal approximations for confidence intervals on parameters

of interest. In mainstream statistics, the empirical likelihood (EL) approach (Owen 1988) has attracted a lot of attention due to several desirable properties. It provides a non-parametric likelihood, leading to EL ratio confidence intervals similar to the parametric likelihood ratio intervals. The shape and orientation of EL intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting, and are particularly useful in providing balanced tail error rates, unlike the symmetric normal theory intervals. As noted in section 3.1, the EL approach was in fact first introduced in the sample survey context by Hartley and Rao (1968), but their focus was on inferential issues related to point estimation. Chen, Chen and Rao (2003) obtained EL intervals on the population mean under simple random and stratified random sampling for populations containing many zeros. Such populations are encountered in audit sampling, where y denotes the amount of money owed to the government and the mean $\overline{Y}$ is the average amount of excessive claims. Previous work on audit sampling used parametric likelihood ratio intervals based on parametric mixture distributions for the variable y. Such intervals perform better than the standard normal theory intervals, but EL intervals perform better under deviations from the assumed mixture model, by providing non-coverage rate below the lower bound closer to the nominal error rate and also larger lower bound. For general designs, Wu and Rao (2004) used a pseudo-empirical likelihood (Chen and Sitter 1999) to obtain adjusted pseudo-EL intervals on the mean and the distribution function that account for the design features, and showed that the intervals provide more balanced tail error rates than the normal theory intervals. The EL method also provides a systematic approach to calibration estimation and integration of surveys. We refer the reader to the review papers by Rao (2004) and Wu and Rao (2005).

Further refinements and extensions remain to be done, particularly on the pseudo-empirical likelihood, but the EL theory in the survey context deserves the attention of practice.

8.2 Exploratory Analyses of Survey Data

In section 6 we discussed methods for confirmatory analysis of survey data taking the design into account, such as point estimation of model (or census) parameters and associated standard errors and formal tests of hypotheses. Graphical displays and exploratory data analyses of survey data are also very useful. Such methods have been extensively developed in the mainstream literature. Only recently, some extensions of these modern methods are reported in the survey literature and deserve the attention of practice. I will briefly mention some of those developments. First, non-parametric kernel density estimates are commonly used to display the shape of a data set without relying on parametric models. They can also be used to compare different sub-populations.

Bellhouse and Stafford (1999) provided kernel density estimators that take account of the survey design and studied their properties and applied the methods to data from the Ontario Health Survey. Buskirk and Lohr (2005) studied asymptotic and finite sample properties of kernel density estimators and obtained confidence bands. They applied the methods to data from the US National Crime Victimization Survey and the US National Health and Nutrition Examination Survey.

Secondly, Bellhouse and Stafford (2001) developed local polynomial regression methods, taking design into account, that can be used to study the relationship between a response variable and predictor variables, without making strong parametric model assumptions. The resulting graphical displays are useful in understanding the relationships and also for comparing different sub-populations. Bellhouse and Stafford (2001) illustrated local polynomial regression on the Ontario Health Survey data; for example, the relationship between body mass index of females and age. Bellhouse, Chipman and Stafford (2004) studied additive models for survey data via penalized least squares method to handle more than one predictor variable, and illustrated the methods on the Ontario Health Survey data. This approach has many advantages in terms of graphical display, estimation, testing and selection of "smoothing" parameters for fitting the models.

8.3 Measurement Errors

Typically, measurement errors are assumed to be additive with zero means. As a result, usual estimators of totals and means remain unbiased or consistent. However, this nice feature may not hold for more complex parameters such as distribution functions, quantiles and regression coefficients. In the latter case, the usual estimators will be biased, even for large samples, and hence can lead to erroneous inferences (Fuller 1995). It is possible to obtain bias-adjusted estimators if estimates of measurement error variances are available. The latter may be obtained by allocating resources at the design stage to make repeated observations on a sub-sample. Fuller (1975, 1995) has been a champion of proper methods in the presence of measurement errors and the bias-adjusted methods deserve the attention of practice.

Hartley and Rao (1978) and Hartley and Biemer (1978) provided interviewer and coder assignment conditions that permit the estimation of sampling and response variances for the mean or total from current surveys. Unfortunately, current surveys are often not designed to satisfy those conditions and even if they do the required information on

interviewer and coder assignments is seldom available at the estimation stage.

Linear components of variance models are often used to estimate interviewer variability. Such models are appropriate for continuous responses, but not for binary responses. The linear model approach for binary responses can result in underestimating the intra-interviewer correlations. Scott and Davis (2001) proposed multi-level models for binary responses to estimate interviewer variability. Given that responses are often binary in many surveys, practice should pay attention to such models for proper analyses of survey data with binary responses.

8.4 Imputation for Missing Survey Data

Imputation is commonly used in practice to fill in missing item values. It ensures that the results obtained from different analyses of the completed data set are consistent with one another by using the same survey weight for all items. Marginal imputation methods, such as ratio, nearest neighbor and random donor within imputation classes are used by many statistical agencies. Unfortunately, the imputed values are often treated as if they were true values and then used to compute estimates and variance estimates. The imputed point estimates of marginal parameters are generally valid under an assumed response mechanism or imputation model. But the "naïve" variance estimators can lead to erroneous inferences even for large samples; in particular, serious underestimation of the variance of the imputed estimator because the additional variability due to estimating the missing values is not taken into account. Advocates of Rubin's (1987) multiple imputation claim that the multiple imputation variance estimator can fix this problem because a between imputed estimators sum of squares is added to the average of naïve variance estimators resulting from the multiple imputations. Unfortunately, there are some difficulties associated with multiple imputation variance estimators, as discussed by Kott (1995), Fay (1996), Binder and Sun (1996), Wang and Robins (1998), Kim, Brick, Fuller and Kalton (2004) and others. Moreover, single imputation is often preferred due to operational and cost considerations. Some impressive advances have been made in recent years on making efficient and asymptotically valid inferences from singly imputed data sets. We refer the reader to review papers by Shao (2002) and Rao (2000, 2005) for methods of variance estimation under single imputation. Kim and Fuller (2004) studied fractional imputation using more than one randomly imputed value and showed that it also leads to asymptotically valid inferences; see also Kalton and Kish (1984) and Fay (1996). An advantage of fractional imputation is that it reduces the imputation variance relative to single imputation using one randomly imputed value. The above methods of variance estimation deserve the attention of practice.

8.5 Multiple Frame Surveys

Multiple frame surveys employ two or more overlapping frames that can cover the target poulation. Hartley (1962) studied the special case of a complete frame B and an incomplete frame A and simple random sampling independently from both frames. He showed that an "optimal" dual frame estimator can lead to large gains in efficiency for the same cost over the single complete frame estimator, provided the cost per unit for frame A is significantly smaller than the cost per unit for frame B. Multiple frame surveys are particularly suited for sampling rare or hard-to-reach populations, such as homeless populations and persons with AIDS, when incomplete list frames contain high proportions of individuals from the target population. Hartley's (1974) landmark paper derived "optimal" dual frame estimators for general sampling designs and possibly different observational units in the two frames. Fuller and Burmeister (1972) proposed improved "optimal" estimators. However, the optimal estimators use different sets of weights for each item y, which is not desirable in practice. Skinner and Rao (1996) derived pseudo-ML (PML) estimators for dual frame surveys that use the same set of weights for all items y, similar to "single frame" estimators (Kalton and Anderson 1986), and maintain efficiency. Lohr and Rao (2005) developed a unified theory for the multiple frames setting with two or more frames, by extending the optimal, pseudo-ML and single frame estimators. Lohr and Rao (2000, 2005) obtained asymptotically valid jackknife variance estimators. Those general results deserve the attention of practice when dealing with two or more frames. Dual frame telephone surveys based on cell phone and landline phone frames need the attention of theory because it is unclear how to weight in the cell phone survey: some families share a cell phone and others have a cell phone for each person.

8.6 Indirect Sampling

The method of indirect sampling can be used when the frame for a target population U^B is not available but the frame for another population U^A, *linked* to U^B, is employed to draw a probability sample. The links between the two populations are used to develop suitable weights that can provide unbiased estimators and variance estimators. Lavallée (2002) developed a unified method, called Generalized Weight Sharing, (GWS), that covers several known methods: the weight sharing method of Ernst (1989) for cross sectional estimation from longitudinal household surveys, network sampling and multiplicity estimation (Sirken 1970) and adaptive cluster sampling (Thompson and Seber 1996). Rao's (1968) theory for sampling from a frame containing an unknown amount of duplication may be regarded as a special case of GWS. Multiple frames can also be handled by GWS and the resulting estimators are simple

but not necessarily efficient compared to the optimal estimators of Hartley (1974) or the PML estimators. The GWS method has wide applicability and deserves the attention of practice.

9. Concluding Remarks

Joe Waksberg's contributions to sample survey theory and methods truly reflect the interplay between theory and practice. Working at the US Census Bureau and later at Westat, he faced real practical problems and produced sound theoretical solutions. For example, his landmark paper (Waksberg 1978) studied an ingenious method (proposed by Warren Mitofsky) for random digit dialing (RDD) that significantly reduces the survey costs compared to dialing numbers completely at random. He presented sound theory to demonstrate its efficiency. The widespread use of RDD surveys is largely due to the theoretical development in Waksberg (1978) and subsequent refinements. Joe Waksberg is one of my heroes in survey sampling and I feel greatly honored to have received the 2005 Waksberg award for survey methodology.

Acknowledgements

I would like to thank David Bellhouse, Wayne Fuller, Jack Gambino, Graham Kalton, Fritz Scheuren and Sharon Lohr for useful comments and suggestions.

References

Aires, N., and Rosén, B. (2005). On inclusion probabilities and relative estimator bias for Pareto πps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.

Andreatta, G., and Kaufmann, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bankier, M.D. (2003). 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Technical Report no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

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

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Bellhouse, D.R., and Rao, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.

Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Bellhouse, D.R., and Stafford, J.E. (2001). Local polynomial regression in complex surveys. *Survey Methodology*, 27, 197-203.

Bellhouse, D.R., Chipman, H.A. and Stafford, J.E. (2004). Additive models for survey data via penalized least squares. Technical Report.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A., and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.

Binder, D.A., Kovacevic, M. and Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6-62.

Brackstone, G. (2002). Strategies and approaches for small area statistics. *Survey Methodology*, 28, 117-123.

Brackstone, G., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 42, 97-114.

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., and Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag.

Buskirk, T.D., and Lohr, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Casady, R.J., and Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

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

Chambers, R.L., and Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Chichester: Wiley.

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.

Chen, J., Chen, S.Y. and Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

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

Cochran, W.G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.

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

Cochran, W.G. (1953). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wicksell.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Deming, W.E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.

Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.

Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.

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

Durbin, J. (1968). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.

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

Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh), New York: John Wiley & Sons, Inc., 135-169.

Ernst, L.R. (1999). The maximization and minimization of sample overlap problem: A half century of results. *Bulletin of the International Statistical Institute*, Vol. LVII, Book 2, 293-296.

Fay, R.E. (1996). Alternative peradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Fellegi, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS sampling without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington DC, 434-442.

Fellegi, I.P. (1981). Should the census counts be adjusted for allocation purposes? – Equity considerations. In *Current Topics in Survey Sampling* (Eds. D. Krewski, R. Platek and J.N.K. Rao). New York: Academic Press, 47-76.

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

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

Fuller, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-147.

Fuller, W.A. (1999). Environmental surveys over time, *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

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.

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. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 28, 310-328.

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

Graubard, B.I., and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.

Hacking, I. (1975). *The Emergence of Probability.* Cambridge: Cambridge University Press.

Hajék, J. (1971). Comments on a paper by Basu, D. In *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston,

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M.H., Dalenius, T. and Tepping, B.J. (1985). The development of sample surveys of finite populations. Chapter 13 in *A Celebration of Statistics*. The ISI Centenary Volume, Berlin: Springer-Verlag.

Hansen, M.H., Hurwitz, W.N. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I and II. New York: John Wiley & Sons, Inc.

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, 776-793.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. and Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190

Hansen, M.H., Hurwitz, W.N., Nisseslson, H. and Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

Hartley, H.O. (1959). Analytical studies of survey data. In Volume in Honour of Corrado Gini, Instituto di Statistica, Rome, 1-32.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Mutiple frame methdology and selected applications. *Sankhyā, Series C*, 36, 99-118.

Hartley, H.O., and Biemer, P. (1978). The estimation of nonsampling variances in current surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 257-262.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hartley, H.O., and Rao, J.N.K. (1978). The estimation of nonsampling variance components in sample surveys. In *Survey Measurement* (Ed. N.K. Namboodiri), New York: Academic Press, 35-43.

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 11-21.

Holt, D., and Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

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.

Hubback, J.A. (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (represented in *Sankhyā*, 1946, vol. 7, 281-294).

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

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, 129-154.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Series A, 149, 65-82.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, A13, 1919-1939.

Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.

Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.

Kim, J., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2004). On the bias of the multiple imputation variance estimator in survey sampling. Technical Report.

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

Kish, L. (1995). The hundred year's wars of survey sampling. *Statistics in Transition*, 2, 813-830.

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

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

Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389.

Kott, P.S. (2005). Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.

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

Kruskal, W.H., and Mosteller, F. (1980). Representative sampling IV: The history of the concept in Statistics, 1895-1939. *International Statistical Review*, 48, 169-195.

Laplace, P.S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951.

Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du partage des poids.* Éditions de l'Université de Burxelles, Belgique, Éditions Ellipse, France.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

Lindley, D.V. (1996). Letter to the editor. *American Statistician*, 50, 197.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Pacific Grove: Duxbury.

Lohr, S.L., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 2710280.

Lohr, S.L., and Rao, J.N.K. (2005). Multiple frame surveys: point estimation and inference. *Journal of the American Statistical Association* (under revision).

Lu, W.W., Brick, M. and Sitter, R.R. (2004). Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Technical Report, Westat, Rockville, Maryland.

Mach, L, Reiss, P.T. and Schiopu-Kratina, I. (2005). The use of the transportation problem in co-ordinating the selection of samples for business surveys. Technical Report HSMD-2005-006E, Statistics Canada, Ottawa.

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

Mahalanobis, P.C. (1944). On large scale sample surveys. *Philosophical Transactions of the Royal Society*, London, Series B, 231, 329-451.

Mahalanobis, P.C. (1946a). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mahalanobis, P.C. (1946b). Sample surveys of crop yields in India. *Sankhyā*, 7, 269-280.

McCarthy, P.J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37, 239-264.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Murthy, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods. In *Contributions to Statistics*: Presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday, Calcutta, Statistical Publishing Society: 283-316.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Aricultural Statistics*, 3, 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Ohlsson, E. (1995). Coordination of samples using permanent random members. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley & Sons, Inc., 153-169.

O'Muircheartaigh, C.A., and Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: A review. *Bulletin of the International Statistical Institute*, Invited Papers, 49, No. 1, 465-493.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2002). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Park, M., and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, 31, 85-93.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

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

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data* (Eds. R.L. Chambers and C.J. Skinner), Chichester: Wiley, 175-195.

Raj, D. (1956). On the method of overlapping maps in sample surveys. *Sankhyā*, 17, 89-98.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.

Rao, J.N.K. (1968). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

Rao, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Proceedings of the workshop on uses of auxiliary information in surveys*, Statistics Sweden.

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. (1996). Developments in sample survey theory: An appraisal. *The Canadian Journal of Statistics*, 25, 1-21.

Rao, J.N.K. (2000). Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 599-608.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken: Wiley.

Rao, J.N.K. (2004). Empirical likelihood methods for sample survey data: An overview. *Proceedings of the Survey Methods Section*, SSC Annual Meeting, in press.

Rao, J.N.K. (2005). Re-sampling variance estimation with imputed survey data: overview. *Bulletin of the International Statistical Institute.*

Rao, J.N.K., and Bellhouse, D.R. (1990). History and development of the theoretical foundations of survey based estimation and analysis. *Survey Methodology*, 16, 3-29.

Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

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.

Rao, J.N.K., and Singh, M.P. (1973). On the choice of estimators in survey sampling. *Australian Journal of Statistics*, 15, 95-104.

Rao, J.N.K., and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics – Theory and Methods*, 33, 2087-2095.

Rao, J.N.K., and Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: Some recent work. *Bulletin of the International Statistical Institute.*

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

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., Jocelyn, W. and Hidiroglou, M.A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19.

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling. *Survey Methodology*, 29, 107-128.

Rao, J.N.K., Yung, W. and Hidiroglou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, Series A, 64, 364-378.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.

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

Rosén, B. (1991). Variance estimation for systematic pps-sampling. Technical Report, Statistics Sweden.

Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

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

Royall, R.M., and Cumberland, W.G. (1981). An empirical study of the ratio estimate and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

Royall, R.M., and Herson, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 and 890-893.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc.

Rubin-Bleuer, S., and Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, (to appear).

Salehi, M., and Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacements, *Biometrika*, 84, 209-219.

Särndal, C.-E. (1996). Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical 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.

Schabenberger, O., and Gregoire, T.G. (1994). Competitors to genuine πps sampling designs. *Survey Methodology*, 20, 185-192.

Schaible, W.L. (Ed.) (1996). Indirect Estimation in U.S. Federal Programs. New York: Springer

Scott, A., and Davis, P. (2001). Estimating interviewer effects for survey responses. *Proceedings of Statistics Canada Symposium 2001*.

Shao, J. (2002). Resampling methods for variance estimation in complex surveys with a complex design. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 303-314.

Shao, J., and Tu, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer Verlag.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 27, 33-44.

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

Singh, A.C., and Wu, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, Series A, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation? *International Statistical Review*, 62, 5-34.

Stehman, S.V., and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

Sukhatme, P.V. (1947). The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys, with Applications.* Ames: Iowa State College Press.

Sukhatme, P.V., and Panse, V.G. (1951). Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.

Sukhatme, P.V., and Seth, G.R. (1952). Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.

Thomas, D.R., and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

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

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the U.S. Census Bureau. *Journal of Official Statistics*, 14, 119-135.

Wang, N., and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., and Rao, J.N.K. (2004). Empirical likelihood ratio confidence intervals for complex surveys. Submitted for publication.

Wu, C., and Rao, J.N.K. (2005). Empirical likelihood approach to calibration using survey data. Paper presented at the 2005 International Statistical Institute meetings, Sydney, Australia.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Griffin.

Zarkovic, S.S. (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society*, Series A, 119, 336-338.

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 139-149
Statistics Canada, Catalogue No. 12-001-XPB

Hot Deck Imputation for the Response Model

Wayne A. Fuller and Jae Kwang Kim [1]

Abstract

Hot deck imputation is a procedure in which missing items are replaced with values from respondents. A model supporting such procedures is the model in which response probabilities are assumed equal within imputation cells. An efficient version of hot deck imputation is described for the cell response model and a computationally efficient variance estimator is given. An approximation to the fully efficient procedure in which a small number of values are imputed for each nonrespondent is described. Variance estimation procedures are illustrated in a Monte Carlo study.

Key Words: Nonresponse; Fractional imputation; Response probability; Replication variance estimation.

1. Introduction

Imputation is used in sample surveys as a method of handling item nonresponse. In hot deck imputation, the imputed values are functions of the respondents in the current sample. Sande (1983) and Ford (1983) contain descriptions of hot deck imputation. Kalton and Kasprzyk (1986) and Little and Rubin (2002) review various imputation procedures.

In one version of hot deck imputation, the imputed value is the value of a respondent in the same imputation cell, where the imputation cells form an exhaustive and mutually exclusive subdivision of the population. In random hot deck imputation, respondents are assigned values at random from respondents in the same imputation cell. The record providing the value is called the *donor* and the record with the missing value is called the *recipient*.

The variance of the imputed estimator is generally larger than the complete sample variance because nonresponse reduces sample size and because the imputed estimator may contain a component due to random imputation. Rao and Shao (1992) proposed an adjusted jackknife method for hot-deck imputation where the first phase units are selected with-replacement. Rao and Sitter (1995) discussed the adjusted jackknife variance estimation method for ratio imputation. Rao (1996) and Sitter (1997) applied the adjusted jackknife method to regression imputation. Shao, Chen and Chen (1998) apply the idea of Rao and Shao (1992) to the balanced repeated replication method. Shao and Steel (1999) propose variance estimation for survey data with composite imputation, where more than one imputation method is used, and the sampling fractions are included in the variance expressions. Yung and Rao (2000) applied the adjusted jackknife method to imputed estimators constructed with a poststratified sample. Rubin (1987) and Rubin and Schenker (1986) suggested multiple imputation procedures. Tollefson and Fuller (1992), and Särndal (1992) proposed imputation methods and corresponding variance estimators. Kim and Fuller (2004) studied the use of fractional imputation for the model in which observations in an imputation cell are independently and identically distributed.

In this paper, we consider hot deck imputation for a population divided into imputation cells. The response model is described in section 2. In section 3, we introduce fully efficient fractional imputation and present a variance estimation method for the imputation estimator, under the assumptions that the probability of nonresponse is constant within a cell. In section 4 we suggest a modification of the fully efficient method that uses a smaller number of donors. In section 5, an example is introduced to illustrate the actual implementation of the proposed method. In section 6, results of a simulation study are reported. Summary is presented in the last section.

2. Basic Setup

Consider a population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$. Associated with each unit i in the population there is a study variable y_i and a vector $\mathbf{x}_i$ of auxiliary information. The set of vectors, $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$, is denoted by F.

Let A denote the indices of the elements in a sample selected by a set of probability rules called the *sampling mechanism*. Let the population quantity of interest be θ_N, let $\hat{\theta}$ be a full sample, linear-in-y, estimator of θ_N, and write

$$\hat{\theta} = \sum_{i \in A} w_i \, y_i. \qquad (1)$$

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011 U.S.A.; Jae Kwang Kim, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea.

If w_i is the inverse of the selection probability, then $\hat{\theta}$ is unbiased for the population total.

Let A_R and A_M denote the set of indices of the sample respondents and sample nonrespondents, respectively. Define the response indicator function

$$R_i = \begin{cases} 1 & \text{if } i \in A_R \\ 0 & \text{if } i \in A_M \end{cases} \tag{2}$$

and let $\mathbf{R} = \{(i, R_i); i \in A\}$. The distribution of $\mathbf{R}$ is called the *response mechanism*.

Assume that the finite population U is made up of G imputation cells, where the set of elements in cell g is U_g. Let n_g be the number of sample elements in imputation cell g and let $r_g, r_g > 0,$ be the number of respondents in imputation cell g. Assume the within-cell uniform response model in which the r_g responses in a cell are equivalent to a Poisson sample selected with equal probabilities from the n_g elements.

Fractional imputation is a procedure in which more than one donor is used per recipient. Kalton and Kish (1984) suggested fractional imputation as an efficient imputation r>*r*, pf 3(t)]TJ4T1_2 1 Tf0 Tc 0r>r

where $e_{iv} = y_{iv} - \bar{Y}_{gv}$, A_{gv} is the set of sample indices in the g^{th} cell for the v^{th} sample, $\bar{Y}_{gv}$ is the population mean of the y–variable in cell gv of population F_v, π_{gv} is the probability that an element in cell gv responds, and F_v denotes the v^{th} population. Also

$$V(\tilde{\theta}_{\text{FE}v} \mid F_v) = V(\hat{\theta}_v \mid F_v) + E\left\{ \sum_{g_v=1}^{Gv} \pi_{gv}^{-1}(1-\pi_{gv}) \sum_{i \in Agv} w_{iv}^2 \, e_{iv}^2 \mid F_v \right\}, \quad (10)$$

where

$$\tilde{\theta}_{\text{FE}v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1) e_{iv}.$$

The estimator (7) can be implemented by using fractional imputation in which every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell. Then, the estimator (7) can be written as the fractionally imputed estimator

$$\hat{\theta}_{\text{FEFI}} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i, \quad (11)$$

where w_j w_{ij}^* is the weight of donor i for recipient j, w_{ij}^* is the imputation fraction of donor i for recipient j defined in (3), and

$$w_{ij}^* = \begin{cases} \left(\sum_{s \in A_R \cap U_g} w_s \right)^{-1} w_i R_i & \text{if } R_j = 0 \\ 1 & \text{if } R_j = 1 \text{ and } i = j. \end{cases} \quad (12)$$

The estimator (11) with w_{ij}^* of (12), algebraically equivalent to (7), is called the *fully efficient fractionally imputed* (FEFI) estimator. The fractionally imputed estimator has the advantage that functions of y such as the fraction less than a given number can be directly estimated from the fractionally imputed data set.

To consider replication variance estimation, let a replication variance estimator for the complete sample be

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

where $\hat{\theta}^{(k)}$ is the k^{th} estimate of θ_N based on the observations included in the k^{th} replicate, L is the number of replicates, and c_k is a factor associated with replicate k determined by the replication method. For a discussion of replication for survey samples see Krewski and Rao (1981) and Rao, Wu and Yue (1992). When the original estimator $\hat{\theta}$ is a linear estimator of the form (1), the k^{th} replicate estimate of $\hat{\theta}$ can be written

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \quad (14)$$

where $w_i^{(k)}$ denotes the replicate weight for the i^{th} unit of the k^{th} replication.

A proposed replicate for the estimator $\hat{\theta}_{\text{FEFI}}$ is

$$\hat{\theta}_{\text{FEFI}}^{(k)} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} =: \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i. \quad (15)$$

Using the replicates (15), the replicate variance estimator can be written as

$$\hat{V}_{\text{FEFI}} = \sum_{k=1}^{L} c_k (\hat{\theta}_{\text{FEFI}}^{(k)} - \hat{\theta}_{\text{FEFI}})^2. \quad (16)$$

The replicates in (15) can be computed in two steps. First, create the usual replicate by defining the weights $w_i^{(k)}$ for every element. Second, for a nonrespondent, the replicate imputation fraction for donor i to recipient j is

$$w_{ij}^{*(k)} = \frac{w_i^{(k)}}{\sum_{s \in A_R \cap U_g} w_s^{(k)}}.$$

Note that the sum of the fractional replication weights of the donor records for each recipient is the same as the replication weight for that unit in a complete sample.

The suggested procedure is closely related to the Rao and Shao (1992) variance estimator. See also Yung and Rao (2000). However, the use of fractional imputation greatly simplifies variance estimation. In the creation of replicates, only the weights on the imputed values are changed. No recomputing of imputed values is required, and once computed, the replicate weights can be used for any smooth function of the vector y. Also, the fractional replicates make the estimator (16) appropriate for a vector of y–variables.

Theorem 3.1 of Kim, Navarro and Fuller (2005) can be used to show that, given a consistent full sample replication procedure,

$$\hat{V}_{\text{FEFI}} = V(\tilde{\theta}_{\text{FE}v} \mid F_v) - N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) e_{iv}^2 + o_p(n_v^{-1}), \quad (17)$$

where $\tilde{\theta}_{\text{FE}v}$ is defined in (10), and the distribution is with respect to the sampling and response mechanisms.

If the finite population correction can be ignored, the estimator (16) is consistent for $V\{\hat{\theta}_{\text{FE}}\}$. If the sample size is large relative to N, then an estimator of

$$N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) \, e_{iv}^2$$

should be added to (16).

The imputation and variance estimation procedure outlined for the response model also produces consistent estimators for the cell mean model. Under the cell mean model, the elements within a cell of the finite population are a realization of independently and identically distributed random variables. The imputation procedure based on the response model is not necessarily fully efficient for the population mean under the cell mean model, but it can be shown that the estimator of the mean and the estimator of the variance of the estimated mean are consistent.

4. Approximations to the Fully Efficient Procedure

In the previous sections, the estimator $\hat{\theta}_{\text{FEFI}}$ was constructed to produce zero imputation variance. The implementation of the fractional imputation procedure, as described in (11), could require the use of a large number of donors for each recipient. Therefore, we outline a procedure with a fixed number of donors per recipient that is fully efficient for the grand total, but not necessarily fully efficient for subpopulations. The procedure assigns donors to produce small between-recipient variance of imputed values and modifies the weights of donors to attain full efficiency for the total.

Suppose that M donors are to be assigned to each recipient. We suggest donors be assigned to recipients to approximate the distribution of all respondents in the cell. One possible selection method is to select a stratified sample for each recipient. A second possibility is to use systematic sampling with probability proportional to the weights to select donors for each recipient. Initial fractions w_{ij0}^* are assigned to the donated values. For systematic sampling with equal weights, the initial w_{ij0}^* is M^{-1}.

After the donors are assigned, the initial fractions, w_{ij0}^* are adjusted so that the sum of the weights gives the fully efficient estimator of the mean of y, and such that the estimated cumulative distribution function based on the weights approximates the fully efficient estimator of the cumulative distribution function. The modification of weights using regression has been suggested by Fuller (1984, 2003). Chen, Rao and Sitter (2000) discussed an efficient imputation method that changes the imputed values rather than the weights. Let $\mathbf{z}_{gj} = (z_{gj1}, z_{gj2}, \ldots, z_{gj\alpha})$ be a vector defined by

$$
\begin{aligned}
z_{gj1} &= y_j \\
z_{gj2} &= 1 \quad \text{if } y_j \le L_2 \\
&= 0 \quad \text{otherwise} \\
&\vdots \\
z_{gj\alpha} &= 1 \quad \text{if } L_{\alpha-1} < y_j \le L_\alpha \\
&= 0 \quad \text{otherwise,}
\end{aligned}
$$

where $L_2, L_3, \ldots, L_\alpha$ divide the range of observed y in cell g into $\alpha - 1$ sections. The number of sections that can be used depends on the numbers and type of observations in the cell, the number of recipients and the number of donors per recipient. If the number of donors per recipient is large, it is possible to adjust the set of weights for each recipient so that the sum of w_{ij}^* over i is one for every j and the sum of $w_{ij}^* y_i$ over i is the fully efficient estimator for every j. In most cases the weights will be adjusted so that the sum of the w_{ij}^* over i is one for every j and the cell means of the imputed values are equal to the fully efficient estimator.

Let $\bar{\mathbf{z}}_{\text{FE, g}}$ denote the fully efficient estimator for cell g. Using regression procedures, the modified w_{ij}^*, modified to give the fully efficient cell mean of $\mathbf{z}$, are

$$
w_{ij}^* = w_{ij0}^* + (\bar{\mathbf{z}}_{\text{FE},g} - \bar{\mathbf{z}}_g^*)\, \mathbf{S}_{\mathbf{zz}g}^{-1} w_{ij0}^* (\bar{\mathbf{z}}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})', \qquad (18)
$$

where

$$
\mathbf{S}_{\mathbf{zz}g} = \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* (\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j}) d_{ij},
$$

$$
\bar{\mathbf{z}}_{g\cdot j} = \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},
$$

$$
\bar{\mathbf{z}}_g^* = \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},
$$

$$
b_j = \left(\sum_{s \in A_{Lg}} w_s \right)^{-1} w_j,
$$

A_{Lg} is the set of indexes of recipients in cell g, $\mathbf{z}_{g[i]j} = \mathbf{z}_{gi}$ is the value imputed from donor i for recipient j, and $\bar{\mathbf{z}}_{g\cdot j}$ is the weighted mean of the imputed values for recipient j using the initial w_{ij0}^*.

To estimate the variance, replicates are created so that the weights on the donors reflect the effect of the deletion of an element on the fully efficient estimator. We use the words "deletion" and "delete" to identify the element chosen for principal weight modification for replication variance estimation.

Let $w_i^{(k)}$ be the weight assigned to element i for the k^{th} replicate for variance estimation of the full sample estimator. Then the replicate for the fully efficient mean of y for cell g is

$$\bar{\mathbf{z}}_g^{(k)} = \left[\sum_{i \in A_{Rg}} w_i^{(k)} \right]^{-1} \sum_{i \in A_{Rg}} w_i^{(k)} \mathbf{z}_i. \qquad (19)$$

Replicate fractions are assigned to donors in cell g so that the replicate estimate of the cell mean is $\bar{\mathbf{z}}_g^{(k)}$. Initial fractional weights $w_{ij0}^{*(k)}$ are assigned where $w_{ij0}^{*(k)}$ is small, but positive, if i is a deleted unit for replicate k. The final fractional weights $w_{ij}^{*(k)}$ are computed using the procedure of (18) with $\bar{\mathbf{z}}_g^{(k)}$ replacing $\bar{\mathbf{z}}_{\mathrm{FE},g}$ and $w_{ij0}^{*(k)}$ replacing w_{ij0}^{*}. The procedure simulates the effect of deleting a single element on the fully efficient estimator.

5. An Artificial Example

In this section, we present an example with artificial data to illustrate the implementation of the proposed method. Suppose that two study variables, x and y, are observed in a sample of size $n = 10$ obtained by simple random sampling. Variable x is a categorical variable with three categories, say 1, 2, and 3, and variable y is a continuous variable. Both variables have item nonresponse and there is a set of imputation cells for each variable. Table 5.1 shows the sample observations, where nonresponse is denoted by M in the table. We use a weight of one to simplify the presentation. Divide by ten to obtain weights for the mean.

Table 5.1
An Illustrative Data Set

Observation	Weight	Cell for x	Cell for y	x	y
1	1	1	1	1	7
2	1	1	1	2	M
3	1	1	2	3	M
4	1	1	1	M	14
5	1	1	2	1	3
6	1	2	1	2	15
7	1	2	2	3	8
8	1	2	1	3	9
9	1	2	2	2	2
10	1	2	1	M	M

Because the x variable is a categorical variable with three categories, using three fractions for fractional imputation gives fully efficient estimators for the distribution of the $x-$variable. Thus the weights in Table 5.2 for the three imputed values of x for observation four are the fractions for the three categories in $x-$cell one.

If a subset of donors is to be used for each recipient, a controlled method of selecting donors, such as systematic sampling, is suggested. In our simple illustration we could easily use fractional imputation with all four y responses in cell 1, but to illustrate the regression adjustment we use only three. See Table 5.2.

Several approaches are possible for the situation in which two items are missing, including the definition of a third set of imputation cells for such cases. Because of the small size of our illustration, we impute under the assumption that x and y are independent within cells. Thus we impute four values for observation ten. For each of the two possible values of x we impute two possible values for y. One of the pair of imputed $y-$values is chosen to be less than the mean of responses and one is chosen to be greater than the mean. See the imputed values for observation 10 in Table 5.2.

Table 5.2
Fractional Weights for Means

Observation	Weight	Donor for y	Cell for x	Cell for y	x	y
1	1.0000		1	1	1	7
2	0.2886	1	1	1	2	7
2	0.3960	6	1	1	2	15
2	0.3154	8	1	1	2	9
3	0.3333	5	1	2	3	3
3	0.3333	7	1	2	3	8
3	0.3334	9	1	2	3	2
4	0.5000		1	1	1	14
4	0.2500		1	1	2	14
4	0.2500		1	1	3	14
5	1.0000		1	2	1	3
6	1.0000		2	1	2	15
7	1.0000		2	2	3	8
8	1.0000		2	1	3	9
9	1.0000		2	2	2	2
10	0.2247	8	2	1	2	9
10	0.2753	4	2	1	2	14
10	0.2095	1	2	1	3	7
10	0.2905	6	2	1	3	15

Initial fractions of one third are assigned to the three imputed values for observations three and four, and initial fractions of one fourth are assigned to the four imputed values for observation ten. The fractional weights are then adjusted using the regression method of equation (18) to give the FEFI mean of y as the estimator, where the fully efficient estimator for the mean of y is

$$\bar{y}_{\mathrm{FE}} = \sum_{g=1}^{2} \frac{n_g}{n} \bar{y}_{Rg} = 8.4833.$$

We restrict the weights for observation 10 so that the estimated fractions for the two categories of x are the cell fractions. Then, because the weighted mean for the categorical variable is controlled for each individual, the vector $\mathbf{z}$ contains only the $y-$variable. Table 5.2 gives the final fractional weights computed with the regression weighting.

An analyst can use the data set of Table 5.2 and any full-sample computer program to compute estimates of functions of y and x, such as the mean of y for the x categories. The fractional data set is fully efficient for any function of the $x-$variable and is also fully efficient for the mean of the $y-$variable.

For jackknife variance estimation, we repeat the weight calculation for each replicate. The replicate estimates of the cell means of y are given in Table 5.3 and the replicate

estimates of the category fractions for x are given in Table 5.4. The values in Table 5.3 and in Table 5.4 are used as the control totals $\bar{\mathbf{z}}_{\mathrm{FE},g}$ in the regression weighting. We used $w_{ij0}^{*(k)} = 3^{-1}$ as the initial value of the replication fractions for observation two and $w_{ij0}^{*(k)} = 4^{-1}$ for observation ten.

Table 5.5 contains the jackknife weights for the fractionally imputed data set of Table 5.2. The replicate weights are used in the same way as replicates for a full sample. They are appropriate, with the caveats of the next section, for any statistic for which the full sample jackknife is appropriate. Thus the procedure is particularly appealing for a general purpose data set, because no additional computations are required of the analyst.

The fully efficient estimator of the mean of y is obtained by treating the respondents as the second phase of a two phase sample. A two-phase variance estimator is

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2}\frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{\mathrm{FE}})^2 + \sum_{g=1}^{2}\left(\frac{n_g}{n}\right)^2\frac{1}{r_g}s_{Rg}^2 = 3.043,$$

where s_{Rg}^2 is the within cell sample variance for cell g. If we use the replication weights in Table 5.5, the replication variance estimate for the mean of y is

$$\hat{V}_{\mathrm{JK}}(\bar{y}_{\mathrm{FI}}) = \sum_{k=1}^{10} 0.9\,(\bar{y}_{\mathrm{FI}}^{(k)} - \bar{y}_{\mathrm{FI}})^2 = 3.078.$$

The difference between the linearized variance estimator and the jackknife variance estimator is

$$\sum_{g=1}^{2}\left(\frac{r_g}{r_g - 1}\frac{n-1}{n} - 1\right)s_{Rg}^2.$$

Thus, the jackknife variance estimator slightly overestimates the true variance in this example.

Table 5.3
Jackknife Replicates of Cell Mean of $y-$variable

Cell	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	12.67	11.25	11.25	10.33	11.25	10.00	11.25	12.00	11.25	11.25
2	4.33	4.33	4.33	4.33	5.00	4.33	2.50	4.33	5.50	4.33

Table 5.4
Jackknife Replicates of Cell Mean of the Dummy Variables of $x-$variable

Cell	Level of x	Replicate									
		1	2	3	4	5	6	7	8	9	10
	1	0.33	0.67	0.67	0.50	0.33	0.50	0.50	0.50	0.50	0.50
1	2	0.33	0.00	0.33	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	3	0.33	0.33	0.00	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	0.50	0.50	0.50	0.50	0.50	0.33	0.67	0.67	0.33	0.50
	3	0.50	0.50	0.50	0.50	0.50	0.67	0.33	0.33	0.67	0.50

Table 5.5
Jackknife Weights for Fractional Imputation

Obs.	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	0	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111
2	0.1664	0	0.3206	0.4205	0.3206	0.4563	0.3206	0.2392	0.3206	0.2724
2	0.6559	0	0.4400	0.3002	0.4400	0.2500	0.4400	0.5540	0.4400	0.5075
2	0.2888	0	0.3505	0.3904	0.3505	0.4048	0.3505	0.3179	0.3505	0.3312
3	0.3706	0.3706	0	0.3706	0.3226	0.3706	0.5018	0.3706	0.2867	0.3706
3	0.3697	0.3697	0	0.3697	0.5018	0.3697	0.0090	0.3697	0.6004	0.3697
3	0.3708	0.3708	0	0.3708	0.2867	0.3708	0.6003	0.3708	0.2240	0.3708
4	0.3703	0.7407	0.7407	0	0.3703	0.5556	0.5556	0.5556	0.5556	0.5556
4	0.3704	0	0.3704	0	0.3704	0.2777	0.2777	0.2777	0.2777	0.2777
4	0.3704	0.3704	0	0	0.3704	0.2778	0.2778	0.2778	0.2778	0.2778
5	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111	1.1111
6	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111
7	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111
8	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111
9	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111
10	0.1624	0.2777	0.2777	0.3061	0.2777	0.2286	0.3474	0.3013	0.1520	0
10	0.3931	0.2778	0.2778	0.2494	0.2778	0.1417	0.3934	0.4395	0.2185	0
10	0.0932	0.2778	0.2778	0.3231	0.2778	0.4400	0.1483	0.0746	0.3171	0
10	0.4623	0.2778	0.2778	0.2324	0.2778	0.3008	0.2220	0.2957	0.4235	0

6. Simulation Studies

6.1 Study Parameters

To study the properties of the imputation procedure we conducted a Monte Carlo study. The sample is a stratified sample with two elements per stratum and two imputation cells, where the cells cut across the strata. Cell one is 20% of the population in strata 1–25 and 80% of the population in strata 26–50. The probability of response is 0.7 for cell one and 0.5 for cell two. Two variables are considered. The variable D is always observed and defines a subpopulation. The probability that $D = 1$ is 0.25 for cell one and 0.40 for cell two. The variable y is subject to nonresponse with constant within-cell response probabilities. The variable D is independent of y and of the response probability. The variable y is normally distributed, where the parameters for a population of 50 strata are given in Table 5.1. In the data generating model of Table 6.1, there are no stratum effects. The parameters of interest are: $\theta_1 =$ mean of y, $\theta_2 =$ mean of y for $D = 1$, $\theta_3 =$ fraction of Y's less than two, $\theta_4 =$ fraction of Y's less than one.

Table 6.1
Parameter Set A

		Cell One		Cell Two	
Strata	Element Weight	Mean	Variance	Mean	Variance
1–25	0.01	0.4	0.36	1.6	0.36
26–50	0.01	0.4	0.36	1.6	0.36

6.2 Estimation Procedures

In the simulation $M = 5$ and $M = 3$ donors were used per recipient. Systematic samples were selected to serve as donors for each recipient. If the number of respondents in the cell is less then M, every respondent was used as a donor for every recipient and the w_{ij}^* are proportional to the original w_i of the respondents. If there are more than M respondents in a cell, the donors are ordered by size and numbered from one to r_g. Then the donors are placed in the order $1, 3, 5, \ldots, r_g, r_{g-1}, r_{g-3}, \ldots, 2$ for r_g odd and the order $1, 3, 5, \ldots, r_{g-1}, r_g, r_{g-2}, \ldots, 2$ for r_g even. The cumulated sums of the weights are formed and m_g systematic samples of size M are selected, where $m_g = n_g - r_g$. The cumulative sums are normalized so that the grand sum is one, a random number, R_{Ng}, between zero and $0.2m_g$ is selected and the m_g samples are the systematic samples of size M defined by the donor associated with $R_{Ng} + 0.2(s-1) + (t-1)m_g^{-1}$, $s = 1, 2, 3, 4, 5$ for recipients $t = 1, 2, \ldots, m_g$. The initial imputation fraction for each donor is $w_{ij}^* = M^{-1}$.

The initial imputation fractions are modified using the regression procedure of (18). The donors in a cell were ordered from smallest to largest and the cumulative sum of the weights formed. Let

$$S_{g,wt} = \sum_{i=1}^{t} w_{[i]}, i \in A_{Rg}, \tag{20}$$

where $w_{[i]}, i = 1, 2, \ldots, r_g$, is the weight of $y_{g,(i)}$ and $y_{g,(1)} \le \ldots \le y_{g,(n)}$ are the ordered y–values in cell g. To define the boundaries of groups to be used to create indicator functions, let t_{*s} be the t for which

$$\max \{S_{g,wt} : S_{g,wt} \le 0.2sS_{gw}\}$$

for $s = 1, 2, 3, 4$, where S_{gw} is the total of the weights of the donors in cell g. Define

$$\begin{aligned} z_{gi,s+1} &= 1 \quad \text{if } y_i \le y_{g,(t_{*s})} \text{ and } i \in A_{Rg} \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{21}$$

for $s = 1, 2, 3, 4$ and let $\mathbf{z}_{gj} = (y_{gj1}, z_{gj2}, \ldots, z_{gj5})$. The regression modified imputed estimator of the mean for each of the five variables in the z–vector is the fully efficient estimator of the respective mean.

The k–deleted FE estimator of the cell mean of $\mathbf{z}$ is defined in (19). The initial fractional weight for donor k to element j is set at $w_{kj0}^{*(k)} = 0.01w_{kj}^*$. This initial weight assures that the final weight will be small, but permits regression adjustment. The final $w_{ij}^{*(k)}$ are computed using the regression procedure of (18) using the initial weight $w_{ij0}^{*(k)}$.

6.3 Monte Carlo Results

The Monte Carlo results for 5,000 samples generated by the parameters of Table 6.1 are given in Table 6.2 and Table 6.3. Results are given for the full sample, for fractional imputation with 5 donors, fractional imputation with three donors, and for multiple imputation (MI) using the Approximate Bayesian Bootstrap (ABB) of Rubin and Schenker (1986) with $M = 5$ and ABB with $M = 3$. Both the FI and MI procedures are unbiased for all four parameters of Table 6.2. The last column of Table 6.2 gives the Monte Carlo variance of the estimator divided by the Monte Carlo variance of the FI procedure with $M = 5$, expressed in percent. The FI procedure is five to ten percent more efficient than MI with $M = 5$ and 9 to 13 percent more efficient than MI with $M = 3$.

Under the model, the mean of the observed values is not the best estimator of the domain mean. In this example, the FI estimator is about as efficient as the full sample estimator. The effect of a smaller number of observations is balanced by the use of a superior estimator of the mean for the domain. Under the model, the domain indicator is independent of the y values, given the cell. Therefore it is efficient to use all values in the cell as donors, not just respondents in the domain.

The properties of the variance estimators are given in Table 6.3. The column headed "Relative Mean" gives the Monte Carlo estimated mean of the estimated variances

divided by the Monte Carlo estimated variance, where the Monte Carlo estimated variance is given in Table 6.2. Both variance estimation procedures appear to be nearly unbiased for the variance of the mean. The relative variance of the MI variance estimator for $M = 5$ is nearly twice that of the FI variance estimator for $M = 5$. For $M = 3$, the MI variance estimator is more than three times that for FI. The MI variance estimator has a large variance because the variance due to missing observations is estimated with four degrees-of-freedom for $M = 5$ and with two-degrees-of freedom for $M = 3$.

The MI variance estimator for the domain mean is seriously biased. This property was first identified by Fay (1991, 1992) and studied by Meng (1994) and Wang and Robins (1998). The FI variance estimator for the domain mean also has a positive bias, though much smaller than that of MI. The bias in the FI variance estimator can be reduced by increasing M, but the bias of MI has little relationship to M.

All variance estimators for the variance of $\hat{\theta}_4$ are slightly negatively biased. We believe FI is slightly biased for $\hat{\theta}_4$ because, although we use the z–vector, the weights are slightly smoothed by the regression procedure. MI is known to have a small sample bias. See Kim (2002).

Table 6.2
Mean and Variance of the Point Estimators Under Setup A (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand. Var.
Mean	Complete Sample	1.00	0.00570	67
(θ_1)	FI(3)	1.00	0.00849	100
	ABB(3)	1.00	0.00926	109
	FI(5)	1.00	0.00849	100
	ABB(5)	1.00	0.00903	106
Domain Mean	Complete Sample	1.14	0.02020	99
(θ_2)	FI(3)	1.14	0.02050	100
	ABB(3)	1.14	0.02230	109
	FI(5)	1.14	0.02040	100
	ABB(5)	1.14	0.02170	106
$\Pr(Y < 2)$	Complete Sample	0.87	0.00104	51
(θ_3)	FI(3)	0.87	0.00202	100
	ABB(3)	0.87	0.00228	113
	FI(5)	0.87	0.00202	100
	ABB(5)	0.87	0.00223	110
$\Pr(Y < 1)$	Complete Sample	0.50	0.00208	66
(θ_4)	FI(3)	0.50	0.00313	100
	ABB(3)	0.50	0.00342	109
	FI(5)	0.50	0.00313	100
	ABB(5)	0.50	0.00329	105

Table 6.3
Relative Mean, t–statistic and Relative Variance for the Variance Estimators Under Setup A (5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)**	t–statistic*	Relative Variance (%)
Mean	FI(3)	100.1	0.05	5.66
(θ_1)	ABB(3)	99.6	–0.19	19.25
	FI(5)	100.1	0.03	5.65
	ABB(5)	98.2	–0.89	9.95
Domain Mean	FI(3)	115.9	7.54	13.88
(θ_2)	ABB(3)	127.9	12.72	28.88
	FI(5)	106.6	3.14	11.62
	ABB(5)	128.4	13.43	20.03
$\Pr(Y < 2)$	FI(3)	103.9	1.86	13.90
(θ_3)	ABB(3)	100.8	0.36	48.42
	FI(5)	101.7	0.82	12.07
	ABB(5)	98.5	–0.67	25.10
$\Pr(Y < 1)$	FI(3)	98.5	–0.75	4.67
(θ_4)	ABB(3)	96.3	–1.80	18.51
	FI(5)	97.6	–1.20	4.45
	ABB(5)	96.7	–1.65	10.17

* Statistic for hypothesis that the estimated variance is unbiased.
** Monte Carlo mean of variance estimates divided by Monte Carlo variance of estimates, in percent.

In a second set of parameters, denoted by *C*, the means were as follows:

$$\begin{aligned}&\text{Cell 1 of strata } 1-25; \quad \mu = 0.4\\ &\text{Cell 1 of strata } 26-50; \quad \mu = 3.0\\ &\text{Cell 2 of strata } 1-25; \quad \mu = 1.6\\ &\text{Cell 2 of strata } 26-50; \quad \mu = 2.2.\end{aligned}$$

All other parameters are the same as in parameter set A. The properties of the estimators are given in Table 6.4. Both FI and MI produce unbiased estimates of the means and of the domain mean. As with parameter set A, the FI procedure is eight to twelve percent more efficient than MI for $M = 5$ and 14 to 16 percent more efficient for $M = 3$.

The assumptions required for MI variance estimation are not satisfied for parameter set C. Therefore the MI estimated variance is seriously biased for all parameters. See Table 6.5. The bias in the MI estimated variance with $M = 5$ is about 17% for the variance of the overall mean and nearly 50% for the domain mean. The bias of the MI variance of the mean for a binomial variable is smaller than the bias for the mean of the continuous variable because the stratification effect is smaller for the binomial variable.

The properties of the estimated variances for the FI procedures are similar to those for setup A. There is a positive bias for the variance of the domain mean of about 23% for $M = 3$ and about 6% for $M = 5$.

The variance of the MI estimated variance is 2.4 to 3.5 times the variance of the FI estimated variance for $M = 5$ and 3 to 7 times for $M = 3$, demonstrating the clear superiority of the FI variance estimator for this configuration.

Table 6.4
Mean and Variance of the Point Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand.Variance
Mean	Complete Sample	2.10	0.00500	48
(θ_1)	FI(3)	2.10	0.01050	100
	ABB(3)	2.10	0.01220	116
	FI(5)	2.10	0.01050	100
	ABB(5)	2.10	0.01150	110
Domain Mean	Complete Sample		0.02530	102
(θ_2)	FI(3)	2.01	0.02510	101
	ABB(3)	2.01	0.02850	115
	FI(5)	2.01	0.02480	100
	ABB(5)	2.01	0.02710	109
$\Pr(Y < 2)$	Complete Sample		0.00127	45
(θ_3)	FI(3)	0.45	0.00281	100
	ABB(3)	0.45	0.00322	115
	FI(5)	0.45	0.00280	100
	ABB(5)	0.45	0.00314	112
$\Pr(Y < 1)$	Complete Sample		0.00107	54
(θ_4)	FI(3)	0.15	0.00199	100
	ABB(3)	0.15	0.00226	114
	FI(5)	0.15	0.00199	100
	ABB(5)	0.15	0.00214	108

Table 6.5
Relative Mean, *t*−statistic and Relative Variance for the Variance Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)	*t*−statistic*	Relative Variance (%)
Mean	FI(3)	100.9	0.41	6.42
(θ_1)	ABB(3)	116.7	7.31	40.14
	FI(5)	100.8	0.39	6.42
	ABB(5)	117.1	7.99	22.29
Domain Mean	FI(3)	122.7	10.78	16.23
(θ_2)	ABB(3)	144.4	19.79	46.05
	FI(5)	106.1	2.95	11.95
	ABB(5)	148.7	22.51	32.49
$\Pr(Y < 2)$	FI(3)	104.4	2.18	6.63
(θ_3)	ABB(3)	114.7	6.54	42.32
	FI(5)	101.8	0.89	6.42
	ABB(5)	112.1	5.74	20.67
$\Pr(Y < 1)$	FI(3)	102.3	1.13	11.08
(θ_4)	ABB(3)	101.3	0.58	39.14
	FI(5)	99.9	−0.04	10.05
	ABB(5)	102.2	1.04	23.60

* Statistic for hypothesis that the estimated variance is unbiased.

7. Summary

In fractional imputation, several donors are used for each missing value and each donor is given a fraction of the weight of the nonrespondent. If all donors are used, the procedure is fully efficient, under the model, for all functions of a **y**–vector. It is shown that the use of fractional imputation with a small number of imputations per nonrespondent can give a fully efficient estimator of the mean. Estimates of other parameters, such as estimates of the cumulative distribution are nearly fully efficient.

Fractional imputation permits the construction of general purpose replicates for variance estimation. A single set of replicates can be used for variance estimation for imputed variables, variables observed on all respondents, and under model assumptions, for functions of the two types of variables. The replicates give estimates of the variances of domain means with much smaller biases than those of multiple imputation. The bias goes to zero as M increases and, in the simulation, is modest for $M = 5$. The replication variance estimator is easily implemented with replication software such as Wesvar.

Fractional imputation with a fixed number of donors per recipient is slightly more efficient for the mean than multiple imputation with the same number of donors. Fractional imputation gives variance estimates with smaller bias and much smaller variance than multiple imputation estimators with the same number of imputations.

8. Acknowledgements

This research was partially supported by a subcontract between Westat and Iowa State University under Contract No. ED-99-CO-0109 between Westat and the Department of Education and by Cooperative Agreement 13-3AEU-0-80064 between Iowa State University, the U.S. National Agricultural Statistics Service and the U.S. Bureau of the Census. We thank Jean Opsomer and Damiao Da Silva for useful comments.

References

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Ford, B.M. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

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

Fuller, W.A. (1996). *Introduction to Statistical Time Series*. 2nd Edition. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds. R.L. Chambers and C.J. Shinner). Wiley, Chichister, England, 307-322.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics Part A – Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, 89, 470-477.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Navarro, A. and Fuller, W.A. (2005). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, to appear.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.

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

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New York: John Wiley & Sons, Inc.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-573.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

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 Sitter, R.R. (1995). Variance estimation under two-phase sampling with applications to imputation for missing data *Biometrika*, 82, 453-460.

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

Rubin, D.B., and Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.

Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Tollefson, M., and Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.

Wang, N., and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 151-159
Statistics Canada, Catalogue No. 12-001-XPB

Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods

J. Michael Brick, Michael E. Jones, Graham Kalton and Richard Valliant [1]

Abstract

Complete data methods for estimating the variances of survey estimates are biased when some data are imputed. This paper uses simulation to compare the performance of the model-assisted, the adjusted jackknife, and the multiple imputation methods for estimating the variance of a total when missing items have been imputed using hot deck imputation. The simulation studies the properties of the variance estimates for imputed estimates of totals for the full population and for domains from a single-stage disproportionate stratified sample design when underlying assumptions, such as unbiasedness of the point estimate and item responses being randomly missing within hot deck cells, do not hold. The variance estimators for full population estimates produce confidence intervals with coverage rates near the nominal level even under modest departures from the assumptions, but this finding does not apply for the domain estimates. Coverage is most sensitive to bias in the point estimates. As the simulation demonstrates, even if an imputation method gives almost unbiased estimates for the full population, estimates for domains may be very biased.

Key Words: Adjusted jackknife; Domain estimation; Model-assisted variance estimation; Multiple imputation; Nonresponse.

1. Introduction

Imputation is frequently used in survey research to assign values for missing item responses, thereby producing complete data sets for public use or general analysis. It is well-recognized that treating imputed values as observed values results in downwardly biased variance estimates for the survey estimates. As a result, confidence intervals have lower than nominal levels. The biases in the variance estimates tend to increase with the item nonresponse rate and can be substantial when that rate is high.

Three methods of variance estimation that have been developed for use with imputed data are studied here: a model-assisted method (Särndal 1992), an adjusted jackknife method (Rao and Shao 1992), and multiple imputation (Rubin 1987). Each method has been evaluated theoretically and by simulation methods, primarily under conditions consistent with the assumptions of the methods. This paper uses simulation to compare the three methods under the same experimental conditions in which some of the assumptions required by the methods do not hold. The goal is to examine the relative performances of the methods in situations that are likely to occur in practice. Other simulation studies of variance estimation methods with imputed data have generally been more limited. Even the more extensive simulation study by Lee, Rancourt, and Särndal (2001) was based on small populations and it did not include multiple imputation.

A single-stage disproportionate stratified sample selected from a real population data set is used to evaluate these variance estimation methods in a realistic setting. The imputed values are assigned using a hot deck imputation method, one of the most popular methods of imputation in survey research. Since hot deck imputation is a form of regression imputation (Kalton and Kaspryzk 1986), restricting the simulation study to the hot deck is not a crucial feature for examining the implications for variance estimation. We study estimation for both full population and domain totals. For the domain estimates, the domain indicator is assumed to be known for all sample members.

Three different combinations of missing data mechanisms and hot deck cell formation are used in the simulations to assess the performance of the variance estimation methods under conditions that violate the assumptions of the methods to varying degrees. The three variance estimation methods we study all assume that data are randomly missing in each hot deck cell and the model-assisted (MA) and multiple imputation (MI) methods also assume that a simple model with common mean and variance holds in each cell. Studying the robustness of the variance estimation methods is an important feature of the simulation because in practice the assumptions underlying the methods will almost never be fully satisfied.

The next section briefly describes three variance estimation methods with hot deck imputed data. The third section outlines the study population, the sample design used in the simulations, and the methods used to generate the missing

1. J. Michael Brick, Michael E. Jones and Graham Kalton, Westat, 1650 Research Boulevard, Rockville, MD 20850; Richard Valliant, University of Michigan, 1218 Lefrak Hall, College Park, MD 20742.

data and implement the hot deck imputations. The fourth section gives the results of the simulations. The last section gives some conclusions about the methods and their applicability.

2. Description of the Variance Estimation Methods

We denote the full sample by *A*, the subset that responds to an item by A_R, and the subset that does not respond by A_M. For the imputations the units are divided into hot deck cells indexed by $g = 1, \ldots, G$, where the subset of n_{Rg} respondents in cell g is A_{Rg}, and the subset of nonrespondents is A_{Mg}. For each unit with a missing value, the hot deck method consists of randomly selecting a respondent from within the same hot deck cell to be the donor of the imputed value.

With hot deck imputation, donors are often selected within a cell by simple random sampling with replacement (srswr), by simple random sampling without replacement, or by sampling with probabilities proportional to the survey weights with replacement (ppswr). Since the simulation results obtained using the srswr and the ppswr methods are very similar, only the results for the ppswr method–termed the weighted hot deck–are presented here. The imputed estimator of a population total is $\hat{\theta}_I = \sum_{i \in A_R} w_i y_i + \sum_{i \in A_M} w_i y_i^*$, where w_i is the survey weight, y_i is the reported value and y_i^* is the imputed value for unit *i* in the nonrespondent set.

2.1 Model-Assisted Variance Estimation

The model-assisted (MA) approach with hot deck imputation assumes that data are randomly missing within the hot deck cells and that a model for the generation of the *y*'s holds. A natural model for use with hot deck imputation is that the y_i's are independently and identically generated within the hot deck cells, *i.e.*, $y_{gi} \overset{\text{iid}}{\sim} (\mu_g, \sigma_g^2)$ for cell *g*. Inferences from the model-assisted approach depend on the validity of the model assumptions.

Särndal (1992) decomposed the total variance of the imputed estimator into three components denoted by V_{SAM}, V_{IMP}, and V_{MIX}. The estimators used for these components in the simulations are those given in Brick, Kalton, and Kim (2004). The MA variance estimator is the sum of the component estimates: $\hat{V}_{\text{MA}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}$. The $\hat{V}_{\text{IMP}}$ and $\hat{V}_{\text{MIX}}$ estimators require an estimator of the element variance in each hot deck cell. Since the simulations showed little difference between weighted and unweighted estimators only the weighted estimator of σ_g^2 is discussed, that is $\hat{\sigma}_g^2 = n_{Rg}(n_{Rg} - 1)^{-1} \sum_{A_{Rg}} w_i (y_i - \bar{y}_{Rg})^2 \times (\sum_{A_{Rg}} w_i)^{-1}$, with $\bar{y}_{Rg} = \sum_{A_{Rg}} w_i y_i (\sum_{A_{Rg}} w_i)^{-1}$.

2.2 Adjusted Jackknife Variance Estimation

The Rao and Shao (1992) adjusted jackknife (AJ) variance estimator for a stratified sample with imputations and ignorable finite population correction factors (*fpc*'s) is

$$\hat{V}(\hat{\theta}_I) = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \frac{n_h - 1}{n_h} (\hat{\theta}_{Ih}^{(k)} - \hat{\theta}_I)^2,$$

where n_h is the number sampled in stratum *h*,

$$\hat{\theta}_{Ih}^{(k)} = \sum_{g=1}^{G} \left\{ \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} + \sum_{(hj) \in A_{Mg}} w_{hj}^{(k)} (y_{hj}^* + \hat{y}_{Rg}^{(k)} - \bar{y}_{Rg}) \right\}$$

is the adjusted estimator when unit *k* is omitted,

$$\hat{y}_{Rg}^{(k)} = \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} \Big/ \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)},$$

$$\bar{y}_{Rg} = \sum_{(hi) \in A_{Rg}} w_{hi} y_{hi} \Big/ \sum_{(hi) \in A_{Rg}} w_{hi},$$

$w_{hi}^{(k)}$ is the weight for unit *hi* adjusted to account for the omission of unit *k*. The notation $(hi) \in B$ denotes unit *i* in stratum *h* is part of set *B*. This procedure requires the computation of $\sum n_h$ replicate estimates, $\hat{\theta}_{Ih}^{(k)}$. A commonly used strategy to reduce the computations is to combine units into variance strata (*e.g.*, see Rust and Rao 1996). Let h^* denote a combined variance stratum and *k* a group of sample units within the combined stratum. All sampled units are assigned to one of the groups. Then, the grouped adjusted jackknife variance estimator is

$$\hat{V}_{AJ} = \sum_{h^*} \sum_{k=1}^{n_{h^*}} \frac{n_{h^*(k)}}{n_{h^*}} (\hat{\theta}_{Ih^*}^{(k)} - \hat{\theta}_I)^2,$$

where n_{h^*} is the number of sample units in combined variance stratum h^*, $n_{h^*(k)}$ is the number of units retained in stratum h^* when units in group *k* are deleted and, corresponding to $\hat{\theta}_{Ih}^{(k)}$, $\hat{\theta}_{Ih^*}^{(k)}$ is the adjusted imputed estimate for the full population when units in group *k* in stratum h^* are deleted. The retained units from design stratum *h* that are in combined variance stratum h^* are assigned replicate weights of $w_{h^*i}^{(k)} = n_{h^*} (n_{h^*(k)})^{-1} w_{h^*i}$.

The AJ method assumes a uniform response probability model within each hot deck cell but, unlike the MA method, it does not require distributional assumptions. Under the uniform response probability model without distributional assumptions, a weighted hot deck is needed to produce unbiased imputed estimates.

In developing the theory for the AJ method, Rao and Shao (1992) assume that *fpc*'s are ignorable. However, the *fpc*'s are not negligible in some strata in the simulations, ranging from about 0.05 to 0.24. Shao and Steel (1999) and Lee, Rancourt, and Särndal (1995) provide methods for accounting for nonnegligible *fpc*'s. The Lee, Rancourt, and Särndal (1995) *fpc* adjustment was applied in the simulations because of its ease of implementation. Without the

fpc adjustment, the AJ variance estimator substantially overestimated the variances in the simulations.

2.3 Multiple Imputation

Multiple imputation (MI) is described in detail in Rubin (1987) and Little and Rubin (2002). The summary here relates to its application with hot deck imputation. As with the model-assisted approach, within the hot deck cells responses are assumed to be missing randomly and the y's are assumed to be independent random variables with a common mean and variance. For each unit that has a missing value, M values are imputed, creating M completed data sets.

To avoid underestimation of variances with the MI method, the hot deck method needs to be modified. Rubin and Schenker (1986) proposed the approximate Bayesian bootstrap (ABB) for simple random sampling with hot deck imputation for use with the MI method. The ABB was modified for the simulations to accommodate sampling donors by ppswr. In the simulations a donor pool for the ABB was created in each cell by selecting respondents with replacement with probabilities proportional to w_i. (There is no literature that discusses the application of ABB methods with unequal weights. In hindsight, an unweighted ABB might have been preferable. The use of an unweighted ABB with a ppswr hot deck yields unbiased point estimates of population totals under the response probability model).

3. Design of the Simulation Study

3.1 Description of the Study Population and Sample Design

The sampling frame for the simulations is a subset of the file of public school districts extracted from the 1999–2000 Common Core of Data (CCD) compiled by the U.S. National Center for Education Statistics. The final frame consists of 11,941 districts.

The sample design used in the simulations is a stratified simple random sample of 1,020 school districts. Twelve strata were created by cross-classifying four categories of number of students (district size) by three categories of the percentage of students at or below the poverty level (poverty status). The strata and number of districts in the frame are given in Table 1. The table also gives the stratum sample sizes and sampling rates used in the simulations.

The table also contains the stratum means and standard deviations for the two study variables, the number of students in the district and the number of districts that include pre-kindergarten as the lowest grade. These study variables were chosen because they are typical of many estimates computed from this type of design.

In addition to the full population estimates we computed the two study estimates for two domains, defined as districts located in the Northeast region and those in nonmetropolitan areas. The means for these domains are substantially different from the full population means for both study variables.

3.2 Missing Data Mechanisms and Imputation Methods

By construction, information on the two study variables is available for all districts in the sampling frame. To create missing values, response indicators were assigned to sampled units within "response cells". In some cases the response cells are the sampling strata, termed STR cells, whereas in other cases they are what are termed HD cells. The HD cells were defined by the cross-classification of four geographic regions and a fourfold categorization of the number of full time equivalent teachers in the district. The HD cells are somewhat correlated with the sampling strata, but each cell contains units from more than one stratum.

Table 1
Stratum Definitions, Population Counts, Sample Sizes, Sampling Rates, Means and Standard Deviations of Number of Students and Proportions of Districts with Pre-Kindergarten

Stratum	District size	Poverty status	N_h	n_h	Sampling rate	Number of students		Proportion with pre-kindergarten
						Mean	Std. dev.	
1	1	1	615	32	0.0520	270.0	155.0	0.44
2	1	2	1,147	59	0.0514	263.3	175.0	0.49
3	1	3	1,292	66	0.0511	243.5	142.5	0.49
4	2	1	1,720	111	0.0645	1,607.2	837.0	0.44
5	2	2	2,305	149	0.0646	1,429.7	784.1	0.52
6	2	3	1,893	122	0.0644	1,427.8	788.8	0.63
7	3	1	692	75	0.1084	4,695.3	1,360.6	0.35
8	3	2	579	63	0.1088	4,728.5	1,365.0	0.51
9	3	3	527	57	0.1082	4,591.8	1,380.3	0.63
10	4	1	342	83	0.2427	16,003.4	12,670.2	0.51
11	4	2	449	110	0.2450	17,577.3	14,246.7	0.58
12	4	3	380	93	0.2447	19,331.8	16,142.7	0.68
Total			11,941	1,020		3,237.9	6,770.5	0.52

Within a given response cell, sampled units were assigned at random to be missing or nonmissing at a specified rate. For each type of response cell, three schemes for assigning rates of missingness were chosen. In two of the schemes, the rates of missingness varied across the response cells, whereas in the other scheme the rate was constant across the cells.

The simulations were conducted by first drawing a stratified simple random sample using the stratum sample sizes in Table 1. Once the sample was selected, response status (respondent/nonrespondent) was randomly assigned to each sampled unit according to the given response scheme. For the MA and AJ methods, the weighted hot deck imputation procedures described earlier were used to impute for missing values. For the MI method, a donor pool was first created using the weighted ABB, and weighted hot decks were then used to impute for each of the $M = 5$ imputed data sets. The estimated total numbers of students and districts with pre-kindergarten were computed for the simulated sample with imputed values, and variance estimates were computed for these estimates using the three variance estimation methods. (If the estimated variance could not be computed in a particular simulation run or the sample size in a cell was less than 2, then that sample was deleted. The maximum number of deleted samples across all the simulations of 10,000 runs each was 2 for the MA method and 28 for the AJ (only one run had 28 AJ samples deleted; the next largest number was 3). The AJ method was based on three combined variance strata and 40 groups of units per stratum for a total of 120 replicates. The three combined strata, formed from strata having about the same *fpc*, consisted of strata 1–6, 7–9, and 10–12. As a check of the grouping, we verified that the grouped jackknife variance procedure gave essentially the same average variance estimates and confidence interval coverage rates as the ungrouped jackknife in the case of complete response. The entire process was repeated 10,000 times for each response scheme.

A feature of the design of the simulation is that the means for the two domains considered often differ substantially from the full population means by strata and HD cells. A key point for the domain estimates is that imputations were made by selecting donors from all the respondents in a hot deck cell, without specifically recognizing the domain as might be done in practice for some domains. After imputations were made for the full sample, the estimated total for a domain was estimated by $\hat{\theta}_I = \sum_{i \in A_R} \delta_i w_i y_i + \sum_{j \in A_M} \delta_j w_j y_j^*$ where $\delta_i = 1$ if unit i is in the domain and 0 if not.

Three of the four possible combinations of response mechanism (STR or HD cells) and hot deck cell formation (STR or HD cells) were studied in the simulations. We refer to these combinations as STR/STR, HD/HD, and STR/HD, where the first set of letters identifies the response mechanism and the second set identifies the type of hot deck cell. The three sets of response rates were 0.2 to 0.6 spaced evenly across the response cells, a constant 0.7 in all cells, and 0.6 to 0.9 spread evenly across the cells. The three combinations of response/hot deck cells with the three sets of response rates generated nine separate simulation schemes for each estimate.

3.3 Assumptions for Models of Response and Population Structure

There are two models involved in the simulations. The population model assumes that the y values within each hot deck cell are independent and have the same expected value. The response model assumes that there is a uniform response probability within each hot deck cell. If both models hold, then the use of either an unweighted or a weighted hot deck will lead to an unbiased estimate of the overall population total. However, if only the response model is assumed, then the use of a weighted hot deck is needed to produce an unbiased estimate of the overall population total. Since the weighted hot deck is used in the simulations, only the response probability model needs to be satisfied for unbiased point estimation of the overall population total. The response probability model holds for all the STR/STR and HD/HD combinations and for the STR/HD combination with a constant response rate; however, it does not hold for the other two STR/HD combinations. The AJ theory for variance estimation of population totals was developed assuming only the response probability model. The MA and MI theories assume that both models hold.

Reliance on only the response probability model and the weighted hot deck to produce unbiased estimates of population totals does not in general extend to estimates of domain totals. When domains cut across hot deck cells, it is necessary to invoke a population model that assumes that the expected value of the domain values is the same as that of the nondomain values in each hot deck cell. However, if the hot deck cells are defined such that each domain comprises the full population in a subset of the hot deck cells, then the situation for point and variance estimation is the same as stated above for overall population totals.

The simulation schemes were generally constructed so that the hot deck cells do not incorporate the domains in order to reflect the practical consideration that it is essentially impossible to incorporate all domains in an imputation scheme. Specifically, in the simulations the districts in the Northeast (NE) region and districts in nonmetropolitan statistical areas (NMSA) are unrelated to the stratum definitions in Table 1 (which are used as hot deck cells in some cases). Also, districts in the NMSA domain can be found in all HD cells. However, the NE

domain is a subset of four of the HD cells. Thus, the definition of the HD cells is more consistent with estimating NE domain totals than NMSA domain totals.

3.4 Summary Statistics

The relative bias of a point estimate is estimated by $\text{relbias}(\hat{\theta}_I)=\text{bias}(\hat{\theta}_I)/\theta_N$, where $\text{bias}(\hat{\theta}_I)=\sum_s(\hat{\theta}_{Is}-\theta_N)/10{,}000$, $\hat{\theta}_{Is}$ is the estimate from sample s, and θ_N is the finite population parameter. The empirical variance of $\hat{\theta}_I$ is $\text{Var}(\hat{\theta}_I)=\sum_s(\hat{\theta}_{Is}-\bar{\theta}_I)^2/10{,}000$, where $\bar{\theta}_I=\sum_s\hat{\theta}_{Is}/10{,}000$. The average variance estimate for a particular method is $v=\sum_s v_s/10{,}000$, where v_s is the estimated variance for simulation run s.

The percentages of intervals that include θ_N are based on the nominal 95 percent confidence intervals $(\hat{\theta}_I \pm t\hat{V}^{1/2})$ computed for each of the 10,000 simulations for each simulation scheme. An issue to consider here is the precision of the variance estimates from a disproportionate stratified sample design and its impact on whether normal approximation or t intervals should be used to calculate confidence intervals. We found that the use of the t–distribution did not have a substantial effect for most cases with the MA and AJ methods, and we have therefore used a multiplier of 1.96 for confidence intervals based on these methods. Rubin and Schenker (1986) suggest using a t–distribution with λ degrees of freedom for confidence intervals with the MI method, where

$$\lambda=(M-1)\left(1+\frac{M}{M+1}\frac{U}{B}\right)^2.$$

Since using 1.96 with the MI method yielded intervals that had severe undercoverage, the t–distribution with λ degrees of freedom is used for the MI confidence intervals.

4. Simulation Results

This section presents the main results from the simulations, beginning with the performance of the three methods of variance estimation for estimates from the full population, followed by the results for the domain estimates. Key outcomes are summarized here graphically, but tables with full details are available in Brick, Jones, Kalton, and Valliant (2004).

4.1 Full Population Estimates

Figure 1 shows the results of the simulations for estimating the total number of students and the number of districts offering pre-kindergarten from the 10,000 samples for each of the nine simulation schemes. The figure gives the relative bias of the imputed estimator, the average variance estimate as a percentage of the empirical variance, and the confidence interval coverage rate.

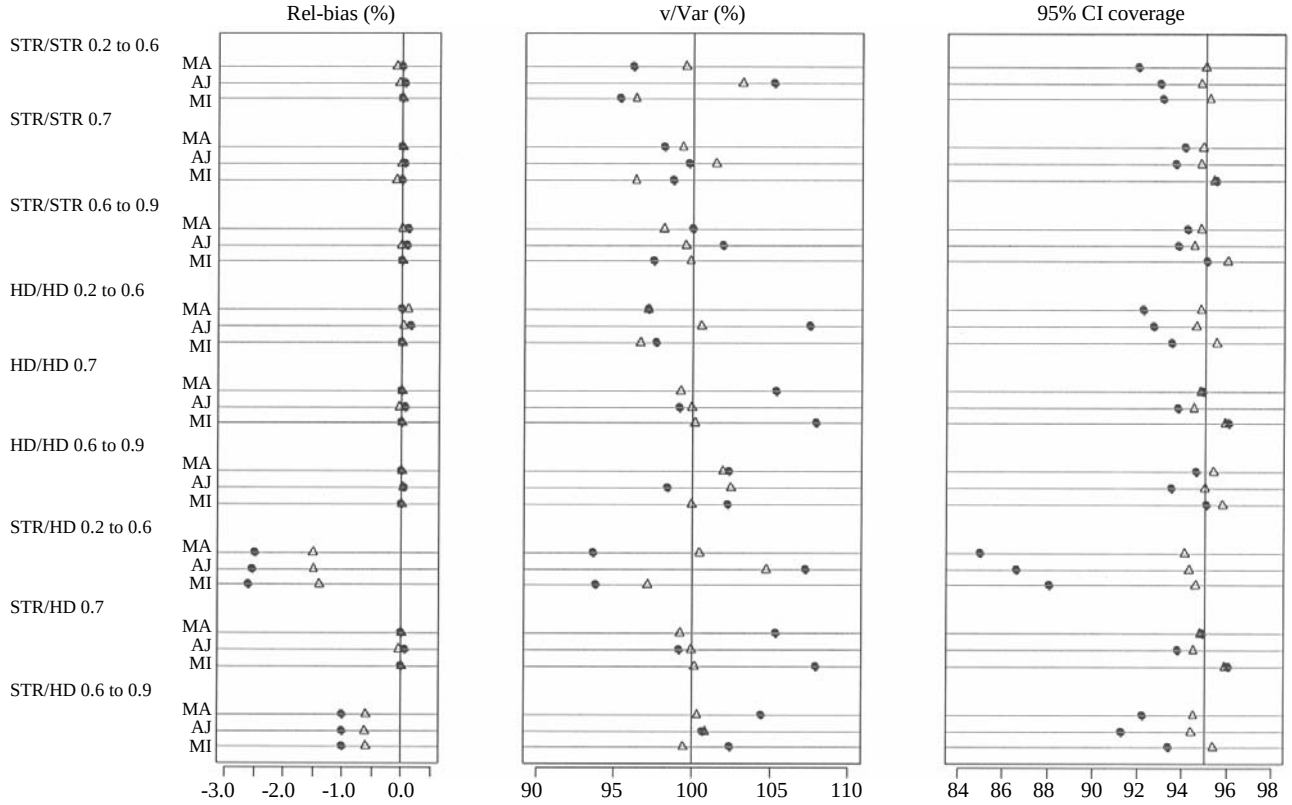


Figure 1. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (•) and number of districts with pre-kindergarten (Δ).

The point estimates are theoretically unbiased with weighted hot deck imputation if all units in a hot deck cell have the same response probability. As noted earlier, this condition holds for the STR/STR and HD/HD combinations and also for the STR/HD combination with a uniform overall response probability. The graph of relative biases in Figure 1 is consistent with this theoretical result within the bounds of simulation error. While the relative biases of the point estimates in the other two STR/HD schemes are small (always less than 3%), they still may be important if the standard errors of the estimates are also small. Cochran (1977, page 12) shows that when the ratio of the bias to the standard error is relatively large, then the coverage rate can be much lower than the nominal level. For the full population estimates with this sample size the ratios never exceed 0.4, but much larger ratios occur for domain estimates, as discussed later.

The graph of the ratios of the average variance estimates to the empirical variances (v/Var in the figures) for the three methods shows that these estimates have relatively small biases in most cases, within a range of plus or minus 8 percent around the simulated true variance. While the ratios for all the methods vary across the nine schemes, the MI ratios are slightly more variable than the other two.

A primary reason for computing variances is to produce confidence intervals. The right-hand panel in Figure 1 shows that the coverage rates for the confidence intervals for the estimates are generally close to the nominal 95 percent level, especially for the pre-kindergarten statistic. The coverage rates for both statistics and all the methods and schemes are between 91% and 96%, with the exception of the number of students for the STR/HD 0.2 to 0.6 scheme. The coverage rates of 88% or less for all three methods in this case, with its extremely high rate of nonresponse, are due to the relatively large bias in the point estimate. Overall, all three variance estimation methods produce confidence intervals with coverages that are vast improvements over those for intervals based on naïve variance estimates (Brick *et al*. 2004).

The confidence interval coverage rates for the MA and AJ methods are essentially equivalent. The MI coverage rates are generally slightly greater than those for the MA and AJ methods. The MI coverage rates are slightly closer to the nominal level for the number of students. Most of the differences are small.

For all three variance estimation methods, the upper and lower confidence interval coverage rates were similar. For the number of students, which is a highly skewed variable, the coverage rates in the two tails are unequal due to correlation between the estimated total and the standard error estimates. The asymmetric tail coverages are also associated with lower overall coverage rates.

The MA and AJ methods yield confidence intervals that have nearly the same average length across the schemes and variables. Because the MI method uses t–distribution values, its intervals range from 10 to 20 percent longer than the MA and AJ intervals when the response rates are low. With the higher response rates, the MI intervals range from about the same to 5 percent longer than the intervals from the two other methods. The MI confidence intervals could, of course, be shortened by increasing M (Rubin 1987, Chapter 4), even though $M = 5$ is typical for applications.

4.2 Domain Estimates

Estimating characteristics for domains that are not explicitly incorporated in the imputation scheme can be problematic when the missing data rate is not trivial. Kalton and Kaspryzk (1986) and Rubin (1996) along with many others have discussed this point and urged the inclusion of as many variables as possible in the imputation process. However, given the many preplanned and ad hoc domain analyses that are carried out with survey data, it is unrealistic to assume that all domains can be accounted for in an imputation scheme. For this reason, the design of the simulations intentionally did not include the domains explicitly in the definition of the hot deck cells. In the case of multiple imputation, issues of variance estimation for domain estimates have received much attention (*e.g.*, Fay 1992; Meng 1994; Rubin 1996).

In the simulations we estimate the totals for two domains: school districts in the NE and those in NMSA. Figures 2 and 3 present the results of the simulations for the NE domain and for the NMSA domain, respectively, in the same format as used before. Note that the scales for Figures 2 and 3 differ from each other and are very different from those used for the full population estimates.

For the NE domain, the point estimates have large positive biases for the STR/STR combinations. Hot deck cells based on STR are not related to region, and, as a result, NE districts with missing data have donors from other regions, which have different characteristics. In contrast, the inclusion of region in the construction of the HD imputation cells removes the bias of the point estimates in the HD/HD combinations and the STR/HD combination with uniform overall response probability, and reduces the bias in the other STR/HD combinations.

All three methods of variance estimation require unbiased point estimates and theory for the methods does not provide guidance on how the methods will perform under the conditions we study. The variance estimates are approximately unbiased for all three variance estimation methods when the domain point estimates are unbiased or have only small biases. However, Figure 2 shows that for the STR/STR combination, where the point estimates are

seriously biased, the variance estimates usually overestimate the empirical variances.

Figure 2 shows that the coverage rates for the HD/HD and STR/HD schemes–for which the point estimates have no or small relative biases–are between 92 percent and 96 percent for all but one of these schemes and variance estimation methods. The exception is the STR/HD combination with response rates between 0.2 and 0.6, which has coverage rates as low as 86 percent for the number of students.

For the STR/STR schemes, Figure 2 shows that all the methods tend to cover at greater than the nominal level for the number of students and less than the nominal level for the number of districts with pre-kindergarten. The difference in the coverage rates for the two variables is due to the sizes of the relative bias of the point estimates and of the variance estimates.

Turning to the NMSA domain estimates in Figure 3, note that metropolitan status is not explicitly included in the definitions of either STR or HD, although it is clearly correlated with size and, thus, with STR. The point estimates for the number of students in the NMSA domain for all the schemes have substantial positive biases. The MA confidence intervals consistently cover at the nominal level or higher, primarily due to the extreme positive biases of the variance estimates. The AJ intervals cover at close to the nominal level for the HD/HD and STR/HD schemes, but undercover in the three STR/STR schemes. The patterns for the MI coverages are similar to those of the AJ, except that the MI intervals appreciably undercover in the HD/HD scheme with 0.2 to 0.6 response rates.

The point estimates of the number of districts with pre-kindergarten in the NMSA domain have moderate negative relative biases for all nine schemes. The confidence intervals for all three methods of variance estimation are close to the nominal level, without the overcoverage found in the NE domain estimates.

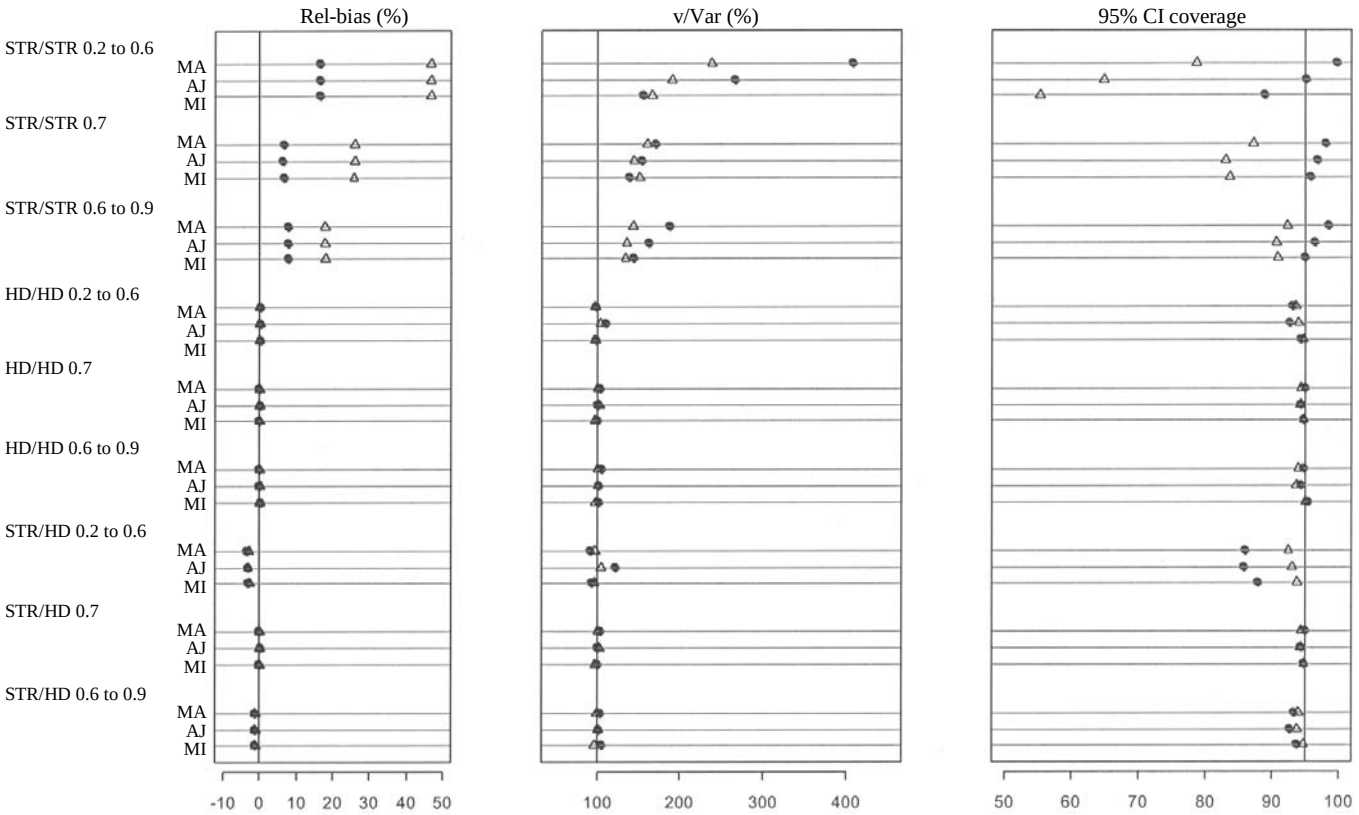


Figure 2. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (•) and number of districts with pre-kindergarten (Δ) in the Northeast.

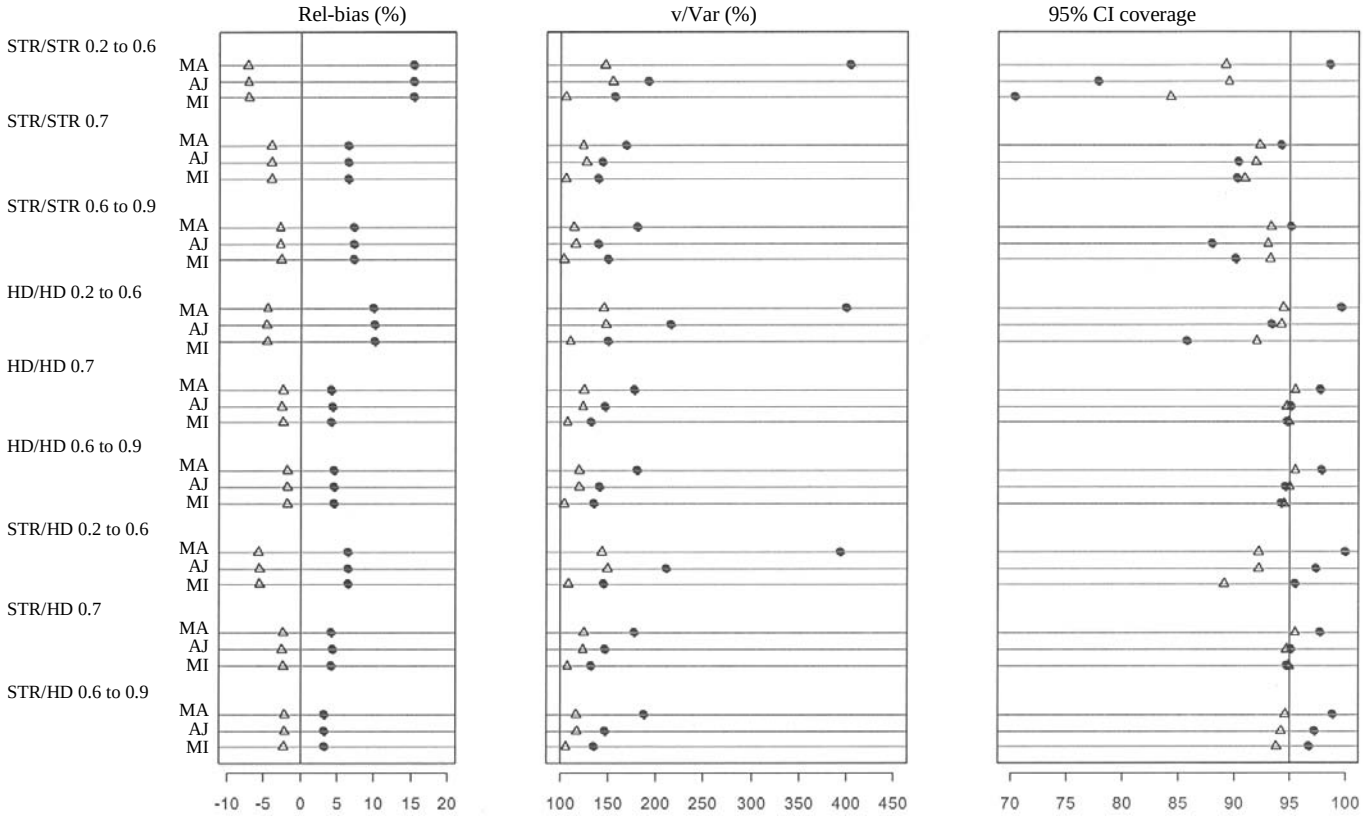


Figure 3. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (•) and number of districts with pre-kindergarten (Δ) in nonmetropolitan areas.

5. Conclusions

The simulations examined the performance of three variance estimators for imputed totals from a single-stage stratified sample design under different response mechanisms with weighted hot deck imputation. The circumstances reflected what can be expected in practice in the sense that the assumptions of the methods were violated in different ways. All three methods were substantial improvements over the naïve variance estimator. All three methods performed very well with unbiased point estimates. When the point estimates had large biases, none of the methods produced confidence intervals with the nominal coverage levels. Poor coverage rates for biased point estimates are not unexpected since the same result holds with no missing data. When the point estimates had relatively small biases, the actual coverage rates for the three variance estimation methods sometimes exceeded and sometimes fell short of the nominal levels. In this case the tendency of all three methods to overestimate the variance often resulted in coverage rates close to the nominal level. Low response rates were associated with undercoverage, largely due to the greater biases in the point estimates.

The differences in the coverage rates of the three methods were generally too small and inconsistent to support claims that any one method is superior in general. With very low response rates, the average lengths of the confidence intervals for the MI method were appreciably longer than those for the MA and AJ methods, but using a larger number of sets of imputations with the MI method would rectify that problem. It should, however, be noted that these simulations only address single stage sampling. Differences in confidence interval lengths between methods may exist in cluster samples. This possibility awaits further investigation.

The results of this study give practitioners of hot deck imputation empirical evidence that all of the variance estimation methods perform well in single stage samples provided that the point estimate is unbiased, even when other assumptions are violated. Estimates for domains that are not taken into account in the imputation scheme are susceptible to large biases. When the point estimates are seriously biased, the methods may produce confidence intervals that cover at far less than the nominal rate. Analysts of imputed data sets should examine whether the imputation method that has been used is likely to give approximately unbiased estimates, especially for domain

estimates. If not, they may need to re-impute the missing items to give less biased point estimates. Advice to imputers to take advantage of as many explanatory variables as feasible in the imputation process is not new, but the evidence from the simulations demonstrates its importance.

Acknowledgements

The authors would like to thank the National Center for Education Statistics, Institute for Education Sciences for supporting this research, and in particular Marilyn Seastrom. We also would like to thank the referees for their constructive comments.

References

Brick, J.M., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.

Brick, J.M., Jones, M., Kalton, G. and Valliant, R. (2004). A simulation study of three methods of variance estimation with hot deck imputation for stratified samples. Prepared under contract No. RN95127001 to the National Center for Education Statistics. Rockville, MD: Westat, Inc.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley & Sons Inc.

Fay, R.E. (1992). When are imputations from multiple imputation valid. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Lee, H., Rancourt, E. and Särndal, C.-E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Statistical Society of Canada Survey Methods Section*, 111-115.

Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A Little), Chapter 21, New York: John Wiley & Sons Inc.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons Inc.

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input. (With discussion). *Statistical Science*, 9, 538-573.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

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

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with nonignorable nonresponse. *Journal of the American Statistical Association*, 81, 361-374.

Rust, K., and Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medicine*, 5, 381-397.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite estimation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 161-168
Statistics Canada, Catalogue No. 12-001-XPB

Does Weighting for Nonresponse Increase the Variance of Survey Means?

Roderick J. Little and Sonya Vartivarian [1]

Abstract

Nonresponse weighting is a common method for handling unit nonresponse in surveys. The method is aimed at reducing nonresponse bias, and it is often accompanied by an increase in variance. Hence, the efficacy of weighting adjustments is often seen as a bias-variance trade-off. This view is an oversimplification – nonresponse weighting can in fact lead to a reduction in variance as well as bias. A covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias – it needs to be related to the probability of response, and it needs to be related to the survey outcome. If the latter is true, then weighting can reduce, not increase, sampling variance. A detailed analysis of bias and variance is provided in the setting of weighting for an estimate of a survey mean based on adjustment cells. The analysis suggests that the most important feature of variables for inclusion in weighting adjustments is that they are predictive of survey outcomes; prediction of the propensity to respond is a secondary, though useful, goal. Empirical estimates of root mean squared error for assessing when weighting is effective are proposed and evaluated in a simulation study. A simple composite estimator based on the empirical root mean squared error yields some gains over the weighted estimator in the simulations.

Key Words: Missing data; Nonresponse adjustment; Sampling weights; Survey nonresponse.

1. Introduction

In most surveys, some individuals provide no information because of noncontact or refusal to respond (*unit* nonresponse). The most common method of adjustment for unit nonresponse is weighting, where respondents and nonrespondents are classified into adjustment cells based on covariate information known for all units in the sample, and a nonresponse weight is computed for cases in a cell proportional to the inverse of the response rate in the cell. These weights often multiply the sample weight, and the overall weight is normalized to sum to the number of respondents in the sample. A good overview of nonresponse weighting is Oh and Scheuren (1983). A related approach to nonresponse weighting is post-stratification (Holt and Smith 1979), which applies when the distribution of the population over adjustment cells is available from external sources, such as a Census. The weight is then proportional to the ratio of the population count in a cell to the number of respondents in that cell.

Nonresponse weighting is primarily viewed as a device for reducing bias from unit nonresponse. This role of weighting is analogous to the role of sampling weights, and is related to the design unbiasedness property of the Horvitz-Thompson estimator of the total (Horvitz and Thompson 1952), which weights units by the inverse of their selection probabilities. Nonresponse weighting can be viewed as a natural extension of this idea, where included units are weighted by the inverse of their inclusion probabilities, estimated as the product of the probability of selection and the probability of response given selection; the inverse of the latter probability is the nonresponse weight. Modelers have argued that weighting for bias adjustment is not necessary for models where the weights are not associated with the survey outcomes, but in practice few are willing to make such a strong assumption.

Sampling weights reduce bias at the expense of increased variance, if the outcome has a constant variance. Given the analogy of nonresponse weights with sampling weights, it seems plausible that nonresponse weighting also reduces bias at the expense of an increase in the variance of survey estimates. The idea of a bias-variance trade-off arises in discussions of nonresponse weighting adjustments (Kalton and Kasprzyk 1986, Kish 1992, Little, Lewitzky, Heeringa, Lepkowski and Kessler 1997). Kish (1992) presents a simple formula for the proportional increase in variance from weighting, say L, under the assumption that the variance of the observations is approximately constant:

$$L = cv^2, \tag{1}$$

where cv is the coefficient of variation of the respondent weights.

Equation (1) is a good approximation when the adjustment cell variable is weakly associated with the survey outcome. However, since it approximates variance rather than mean squared error, it does not measure the potential nonresponse bias reduction that is the main objective of weighting, and it does not apply to outcomes

1. Roderick J. Little, University of Michigan, U.S.A. E-mail: rlittle@umich.edu; Sonya Vartivarian, Mathematica Policy Research, Inc. 600 Maryland Ave SW, Suite 550, Washington, D.C. 20024-2512. E-mail: SVartivarian@Mathematica-MPR.com.

that are associated with the adjustment cell variable, where nonresponse weighting can in fact reduce the variance. The fact that nonresponse weighting can reduce variance is implicit in the formulae in Oh and Scheuren (1983), and is noted in Little (1986) when adjustment cells are created using predictive mean stratification. It is also seen in the related method of post-stratification for nonresponse adjustment (Holt and Smith 1979).

Variability of the weights per se does not necessarily translate into estimates with high variance: an estimate with a high value of L can have a smaller variance than an estimate with a small value of L, as is shown in the simulations in section 3. Also, the situations where nonresponse weighting is most effective in reducing bias are precisely the situations where the weighting tends to reduce, not increase, variance, and Equation (1) does not apply. This differs from the case of sampling weights, and is related to "super-efficiency" that can result when weights are estimated from the sample rather than fixed constants; see, for example, Robins, Rotnitsky and Zhao (1994).

We propose a simple refinement of Equation (1), namely Equation (14) below, that captures both bias and variance components whether or not the adjustment cell variable is associated with the outcome, and hence is a more accurate gauge of the value of weighting the estimates, and of alternative adjustment cell variables. In multipurpose surveys with many outcomes, the standard approach is to apply the same nonresponse weighting adjustment to all the variables, with the implicit assumption that the value of nonresponse bias reduction for some variables outweighs the potential variance increase for others. Our empirical estimate of mean squared error allows a simple refinement of this strategy, namely to restrict nonresponse weighting to the subset of variables for which nonresponse weighting reduces the estimated mean squared error. This composite strategy is assessed in the simulation study in section 3, and shows some gains over weighting all the outcomes. As noted in section 4, there are alternative approaches that have even better statistical properties, but these lead to different weights for each variable and hence are more cumbersome to implement and explain to survey users.

2. Nonresponse Weighting Adjustments for a Mean

Suppose a sample of n units is selected. We consider inference for the population mean of a survey variable Y subject to nonresponse. To keep things simple and focused on the nonresponse adjustment question, we assume that units are selected by simple random sampling. The points made here about nonresponse adjustments also apply in general to complex designs, although the technical details become more complicated.

We assume that respondents and nonrespondents can be classified into C adjustment cells based on a covariate X. Let M be a missing-data indicator taking the value 0 for respondents and 1 for nonrespondents. Let n_{mc} be the number of sampled individuals with $M = m$, $X = c$, $m = 0, 1$; $c = 1, \ldots, C$, $n_{+c} = n_{0c} + n_{1c}$ denote the number of sampled individuals in cell c, $n_0 = \sum_{c=1}^{C} n_{0c}$ and $n_1 = \sum_{c=1}^{C} n_{1c}$ the total number of respondents and nonrespondents, and $p_c = n_{+c}/n$, $p_{0c} = n_{0c}/n_0$ the proportions of sampled and responding cases in cell c. We compare two estimates of the population mean μ of Y, the unweighted mean

$$\bar{y}_0 = \sum_{c=1}^{C} p_{0c}\bar{y}_{0c}, \qquad (2)$$

where $\bar{y}_{0c}$ is the respondent mean in cell c, and the weighted mean

$$\bar{y}_w = \sum_{c=1}^{C} p_c\bar{y}_{0c} = \sum_{c=1}^{C} w_c p_{0c}\bar{y}_{0c}, \qquad (3)$$

which weights respondents in cell c by the inverse of the response rate $w_c = p_c / p_{0c}$. The estimator (3) can be viewed as a special case of a regression estimator, where missing values are imputed by the regression of Y on indicators for the adjustment cells. We compare the bias and mean squared error of (2) and (3) under the following model, which captures the important features of the problem. We suppose that conditional on the sample size n, the sampled cases have a multinomial distribution over the $(C \times 2)$ contingency table based on the classification of M and X, with cell probabilities

$$\Pr(M = 0, X = c) = \phi\pi_{0c}; \Pr(M = 1, X = c) = (1-\phi)\pi_{1c},$$

where $\phi = \Pr(M = 0)$ is the marginal probability of response. The conditional distribution of X given $M = 0$ and n_0 is multinomial with cell probabilities $\Pr(X = c \mid M = 0) = \pi_{0c}$, and the marginal distribution of X given n is multinomial with index n and cell probabilities

$$\Pr(X = c) = \phi\pi_{0c} + (1-\phi)\pi_{1c} = \pi_c,$$

say. We assume that the conditional distribution of Y given $M = m$, $X = c$ has mean μ_{mc} and constant variance σ^2. The mean of Y for respondents and nonrespondents are

$$\mu_0 = \sum_{c=1}^{C} \pi_{0c}\mu_{0c}, \ \mu_1 = \sum_{c=1}^{C} \pi_{1c}\mu_{1c},$$

respectively, and the overall mean of Y is $\mu = \phi\mu_0 + (1-\phi)\mu_1$.

Under this model, the conditional mean and variance of $\bar{y}_w$ given $\{p_c\}$ are respectively $\sum_{c=1}^{C} p_c\mu_{0c}$ and $\sigma^2\sum_{c=1}^{C} p_c^2/n_{0c}$. Hence the bias of $\bar{y}_w$ is

$$b(\bar{y}_w) = \sum_{c=1}^{C} \pi_c(\mu_{0c} - \mu_c),$$

where π_c and μ_c are the population proportion and mean of Y in cell c. This can be written as

$$b(\bar{y}_w) = \tilde{\mu}_0 - \mu, \tag{4}$$

where $\tilde{\mu}_0 = \sum_{c=1}^{C} \pi_c \mu_{0c}$ is the respondent mean "adjusted" for the covariates, and $\mu = \sum_{c=1}^{C} \pi_c \mu_c$ is the true population mean of Y. The variance of $\bar{y}_w$ is the sum of the expected value of the conditional variance and the variance of its conditional expectation, and is approximately

$$V(\bar{y}_w) = (1+\lambda)\sigma^2 / n_0 + \sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2 / n, \tag{5}$$

where $\lambda = \sum_{c=1}^{C} \pi_{0c}((\pi_c / \pi_{0c} - 1)^2)$ is the population analog of the variance of the nonresponse weights $\{w_c\}$, which is the same as L in Equation (1) since the weights are scaled to average to one. The formula for the variance of the weighted mean in Oh and Scheuren (1983), derived under the quasi-randomization perspective, reduces to (5) when the within-cell variance is assumed constant, and finite population corrections and terms of order $1/n^2$ are ignored. The mean squared error of $\bar{y}_w$ is thus

$$\text{mse}(\bar{y}_w) = b^2(\bar{y}_w) + V(\bar{y}_w). \tag{6}$$

The mean squared error of the unweighted mean (2) is

$$\text{mse}(\bar{y}_0) = b^2(\bar{y}_0) + V(\bar{y}_0), \tag{7}$$

where:

$$b(\bar{y}_0) = b(\bar{y}_w) + \mu_0 - \tilde{\mu}_0, \tag{8}$$

is the bias and

$$V(\bar{y}_0) = \sigma^2 / n_0 + \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 / n_0, \tag{9}$$

is the variance. Hence the difference (say Δ) in mean squared errors is

$$\begin{aligned} &\Delta = \text{mse}(\bar{y}_0) - \text{mse}(\bar{y}_w) = B + V_1 - V_2, \text{ where} \\ &B = (\mu_0 - \tilde{\mu}_0)^2 + 2(\mu_0 - \tilde{\mu}_0)(\tilde{\mu}_0 - \mu), \\ &V_1 = \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 / n_0 - \sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2 / n, \\ &V_2 = \lambda\sigma^2 / n_0 \end{aligned} \tag{10}$$

Equation (10) and its detailed interpretation provide the main results of the paper; note that positive terms in (10) favor the weighted estimator $\bar{y}_w$.

(a) The first term B represents the impact on MSE of bias reduction from adjustment on the covariates. It is order one and increasingly dominates the MSE as the sample size increases. If $\mu \le \tilde{\mu}_0 < \mu_0$ or $\mu_0 < \tilde{\mu}_0 \le \mu$, then weighting has reduced the bias of the respondent mean, and both of the components of B are positive. In particular, if the missing data are missing at random (Rubin 1976, Little and Rubin 2002), in the sense that respondents are a random sample of the sampled cases in each cell c, then $\tilde{\mu}_0 = \mu$ and weighting eliminates the bias of the unweighted mean. The bias adjustment is

$$\mu_0 - \tilde{\mu}_0 \simeq \sum_{c=1}^{C} \pi_{0c}(1 - w_c)(\mu_{0c} - \mu_0),$$

ignoring differences between the weights and their expectations. This is zero to $O(1)$ if either nonresponse is unrelated to the adjustment cells (in which case $w_c \approx 1$ for all c, or the outcome is unrelated to the adjustment cells (in which case $\mu_{0c} \approx \mu_0$ for all c). Thus a substantial bias reduction requires adjustment cell variables that are related both to nonresponse and to the outcome of interest, a fact that has been noted by several authors. It is often believed that conditioning on observed characteristics of nonrespondents will reduce bias, but note that this is not guaranteed; it is possible for the adjusted mean to be further on average from the true mean than the unadjusted mean, in which case weighting makes the bias worse.

(b) The effect of weighting on the variance is represented by $V_1 - V_2$.

(c) For outcomes Y that are unrelated to the adjustment cells, $\mu_{0c} = \mu_0$ for all c, $V_1 = 0$, and weighting increases the variance, since V_2 is positive. The variance part of equation (10) then reduces to the population version of Kish's formula (1). Adjustment cell variables that are good predictors of nonresponse hurt rather than help in this situation, since they increase the variance of the weights without any reduction in bias; but there is no bias-variance trade-off for these outcomes, since there is no bias reduction.

(d) If the adjustment cell variable X is unrelated to nonresponse, then λ is $O(1/n)$ and hence V_2 has a lower order of variability than V_1. The term V_1 tends to be positive, since $\sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 \simeq \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \tilde{\mu}_0)^2$, and the divisor n in the second term is larger than the divisor n_0 in the first term. Thus weighting in this case tends to have no impact on the bias, but reduces variance to the extent that X is a good predictor of the outcome. This contradicts the notion that weighting increases variance. The above-mentioned "super-efficiency" that results from estimating nonresponse weights from the sample is seen by the fact that if the data are missing completely at random, then the "true" nonresponse weight is a constant for all responding units. Hence weighting by "true" weights

leads to (2), which is less efficient than weighting by the "estimated" weights, which leads to (3).

(e) If the adjustment cell variable is a good predictor of the outcome and also predictive of nonresponse, then V_2 is again small because of the reduced residual variance σ^2, and V_1 is generally positive by a similar argument to (d). The term $\sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2$ may deviate more from $\sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2$ because the weights are less alike, but this difference could be positive or negative, and the different divisors seem more likely to determine the sign and size of V_1. Thus, weighting tends to reduce both bias and variance in this case.

(f) Equation (9) can be applied to the case of post-stratification on population counts, by letting n represent the population size rather than the sample size. Assuming a large population, the second term in V_1 essentially vanishes, increasing the potential for variance reduction when the variables forming the post-strata are predictive of the outcome. This finding replicates previous results on post-stratification (Holt and Smith 1979; Little 1993).

A simple qualitative summary of the results (a) – (f) of section 2 is shown in Table 1, which indicates the direction of bias and variance when the associations between the adjustment cells and the outcome and missing indicator are high or low. Clearly, weighting is only effective for outcomes that are associated with the adjustment cell variable, since otherwise it increases the variance with no compensating reduction in bias. For outcomes that are associated with the adjustment cell variable, weighting increases precision, and also reduces bias if the adjustment cell variable is related to nonresponse.

Table 1
Effect of Weighting Adjustments on Bias and Variance of a Mean, by Strength of Association of the Adjustment Cell Variables with Nonresponse and Outcome

	Association with outcome	
Association with nonresponse	Low	High
Low	Cell 1 Bias: --- Var: ---	Cell 3 Bias: --- Var: ↓
High	Cell 2 Bias: --- Var: ↑	Cell 4 Bias: ↓ Var: ↓

It is useful to have estimates of the MSE of $\bar{y}_0$ and $\bar{y}_w$ that can be computed from the observed data. Let $s_{0c}^2 = \sum_{i \in c}(y_i - \bar{y}_{0c})^2/(n_{0c} - 1)$ denote the sample variance of respondents in cell c, $s^2 = \sum_{c=1}^{C}(n_{0c} - 1)s_{0c}^2/(n_{0c} - C)$ the pooled within-cell variance, and $s_0^2 = \sum_{i=1}^{n_0}(y_i - \bar{y}_0)^2/(n_0 - 1)$, the total sample variance of the respondent values. We use the following approximately unbiased expressions, under the assumption that the data are MAR:

$$\hat{\text{mse}}(\bar{y}_0) = \hat{B}^2(\bar{y}_0) + \hat{V}(\bar{y}_0), \tag{11}$$

where $\hat{V}(\bar{y}_0) = s_0^2/n_0$ and

$$\hat{B}^2(\bar{y}_0) = \max\{0, (\bar{y}_w - \bar{y}_0)^2 - V_d\}$$

$$V_d = (n_1/n)^2 \left(\begin{array}{l} \sum_{c=1}^{C} p_{1c}(\bar{y}_{0c} - \bar{y}_0^{(1)})^2/n_1 \\ + \sum_{c=1}^{C} p_{0c}(\bar{y}_{0c} - \bar{y}_0)^2/n_0 \\ + s^2 \sum_{c=1}^{C} (p_{1c} - p_{0c})^2/n_{0c} \end{array} \right), \tag{12}$$

where $\bar{y}_0^{(1)} = \sum_{c=1}^{C} p_{1c}\bar{y}_{0c}$, and V_d estimates the variance of $(\bar{y}_w - \bar{y}_0)$ and is included in (12) as a bias adjustment for $(\bar{y}_w - \bar{y}_0)^2$ as an estimate of $B^2(\bar{y}_0)$, similar to that in Little *et al.* (1997). Also

$$\hat{\text{mse}}(\bar{y}_w) = \hat{V}(\bar{y}_w) = (1 + L)s^2/n_0 + \sum_{c=1}^{C} p_c(\bar{y}_{0c} - \bar{y}_w)^2/n. \tag{13}$$

Subtracting (11) from (13), the difference in MSE's of $\bar{y}_w$ and $\bar{y}_0$ is then estimated by

$$D = Ls^2/n_0 - (s_0^2 - s^2)/n_0 + \sum_{c=1}^{C} p_c(\bar{y}_{0c} - \bar{y}_w)^2/n - \hat{B}^2(\bar{y}_0). \tag{14}$$

This is our proposed refinement of (1), which is represented by the leading term on the right side of (14).

3. Simulation Study

We include simulations to illustrate the bias and variance of the weighted and unweighted mean for sets of parameters representing each cell in Table 1. We also compare the analytic MSE approximations in Equations (6) and (7) and their sample-based estimates (11) and (13) with the empirical MSE over repeated samples.

3.1 Superpopulation Parameters

The simulation set-up for the joint distribution of X and M is described in Table 2. The sample is approximately uniformly distributed across the adjustment cell variable X, which has $C = 10$ cells. Two marginal response rates are chosen, 70%, corresponding to a typical survey value, and 52%, a more extreme value to accentuate differences in methods. Three distributions of M given X are simulated to model high, medium and low association.

The simulated distributions of the outcome Y given $M = m, X = c$ are shown in Table 3. These all have the form

$$[Y \mid M = m, X = c] \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

Table 2
Percent of Sample Cases in Adjustment Cell *X* and Missingness Cell *M*

a. Overall Response Rate = 52%

	Association Between *M* and *X*	*X*	1	2	3	4	5	6	7	8	9	10
1.	High	*M* = 0	0.55	1.00	4.01	4.52	5.04	5.55	6.06	6.58	9.14	9.96
		M = 1	8.69	9.00	6.01	5.53	5.04	4.54	4.04	3.54	1.02	0.20
2.	Medium	*M* = 0	2.77	3.50	4.01	4.52	5.04	5.55	6.06	6.58	7.11	7.62
		M = 1	6.47	6.50	6.01	5.53	5.04	4.54	4.04	3.54	3.05	2.54
3.	Low	*M* = 0	4.62	5.15	5.21	5.28	5.34	5.40	5.45	5.52	5.58	5.64
		M = 1	4.62	4.85	4.81	4.77	4.73	4.69	4.65	4.60	4.57	4.52

b. Overall Response Rate = 70%

	Association Between *M* and *X*	*X*	1	2	3	4	5	6	7	8	9	10
1.	High	*M* = 0	0.55	3.00	6.51	7.04	7.55	8.07	8.59	9.11	9.64	9.96
		M = 1	8.69	7.00	3.51	3.02	2.52	2.02	1.52	1.01	0.51	0.20
2.	Medium	*M* = 0	4.44	5.30	5.81	6.33	6.85	7.37	7.88	8.40	8.93	9.45
		M = 1	4.80	4.70	4.21	3.72	3.22	2.72	2.22	1.72	1.22	0.71
3.	Low	*M* = 0	6.19	6.85	6.91	6.98	7.05	7.11	7.17	7.24	7.31	7.37
		M = 1	3.05	3.15	3.11	3.07	3.02	2.98	2.93	2.88	2.84	2.79

Table 3
Parameters for $[Y \mid M = m,\ X = c] \sim N(\beta_0 + \beta_1 c, \sigma^2)$

	Association Between *Y* and *X*	β_1	σ^2	ρ^2
1.	High	4.75	46	≈ 0.80
2.	Medium	3.70	122	≈ 0.48
3.	Low	0.00	234	0.00

Three sets of values of (β_1, σ^2) are simulated to model high, medium and low associations between *Y* and *X*. The intercept β_0 is chosen so that the overall mean of *Y* is $\mu = 26.3625$ for each scenario.

A thousand replicate samples of size n = 400 and n = 2,000 were simulated for each combination of parameters in Tables 2 and 3. Samples where $n_{0c} = 0$ for any c were excluded, since the weighted estimate cannot be computed; in practice some cells would probably be pooled in such cases. The numbers of excluded simulations are shown in Table 4.

Table 4
Numbers of Replicates Excluded Because of Cell with no Respondents

		Response Rate	
Association of *M* and *X*	Association of *Y* and *X*	52%	70%
High	High	134	113
	Medium	120	117
	Low	131	104
Medium	Low	1	0

3.2 Comparisons of Bias, Variance and Root Mean Squared Error, and their Estimates

Summaries of empirical bias and root MSE's (RMSE's) are reported in Table 5. The empirical RMSE's of the weighted mean can be compared with the following estimates, which are displayed in Table 5, averaged over the 1,000 replicates: The estimated RMSE based on Kish's rule of thumb Equation (1), namely:

$$\hat{\text{mse}}_{\text{Kish}}(\bar{y}_w) = (1 + L)s_Y^2 / n_0,$$

$$\text{where } s_Y^2 = \sum_{i=1}^{n_0} (y_i - \bar{y}_0)^2 / (n_0 - 1); \qquad (15)$$

The analytical RMSE from Equations (6) and (7); and the estimated RMSE from Equations (11) and (13).

Following the suggestion of Oh and Scheuren (1983), we include in the last two columns of Table 5 the average empirical bias and RMSE of a composite mean that chooses between $\bar{y}_w$ and $\bar{y}_0$, picking the estimate with a lower sample-based estimate of the MSE. The empirical bias relative to the population parameter is reported for all estimators. We also include the bias and RMSE of the mean before deletion of cases due to nonresponse.

Table 5a shows results for simulations with a response rate of 52%. Rows are labeled according to the four cells in Table 1, with medium and high associations combined. For each row, the lower of the RMSE's for the unweighted and weighted respondent means is bolded, indicating superiority for the corresponding method.

The first four rows of Table 5a correspond to cell 4 in Table 1, with medium/high association between *Y* and *X* and

medium/high association between M and X. In these cases $\bar{y}_w$ has much lower RMSE than $\bar{y}_0$, reflecting substantial bias of $\bar{y}_0$ that is removed by the weighting.

The next two rows of Table 5a corresponding to cell 3 of Table 1, with medium/high association between Y and X and low association between M and X. In these cases $\bar{y}_0$ is no longer seriously biased, but $\bar{y}_w$ has improved precision, particularly when the association of Y and X is high. These are cases where the variance is reduced, not increased, by weighting. The analytic estimates of RMSE and sample-based estimates are close to the empirical RMSE estimates, while Kish's rule of thumb overestimates the RMSE, as predicted by the theory in section 2.

The next two rows of Table 5a correspond to cell 2 of Table 1, where the association between Y and X is low and the association between M and X is medium or high. In these cases, $\bar{y}_w$ has higher MSE than $\bar{y}_0$. These cases illustrate situations where the weighting increases variance, with no compensating reduction in bias. The last row corresponds to cell 1 of Table 1, with low associations between M and X and between Y and X. The unweighted mean has lower RMSE in these cases, but the increase in RMSE from weighting is negligible. For the last three rows of Table 5a, RMSE's from Kish's rule of thumb are similar to those from the analytical formula in section 2 and empirical estimates based on this formulae, and all these estimates are close to the empirical RMSE.

The last two columns of Table 5a show empirical bias and RMSE of the composite method that chooses $\bar{y}_w$ or $\bar{y}_0$ based on the estimated RMSE. For the simulations in the first 6 rows, the composite estimator is the same as $\bar{y}_w$, and hence detects and removes the bias of the unweighted mean. For simulations in cell 1 (the last row) the composite estimator performs like $\bar{y}_w$ or $\bar{y}_0$, as expected since $\bar{y}_w$ and $\bar{y}_0$ perform similarly in this case. For simulations in cell 2 that are not favorable to weighting, the composite estimator has lower RMSE than $\bar{y}_w$, but considerably higher than that of $\bar{y}_0$, suggesting that for the conditions of this simulation the empirical MSE affords limited ability to pick the better estimator in individual samples.

Nevertheless, the composite estimator is the best overall estimator of the three considered in this simulation.

Table 5b shows results for the 70% response rate. The pattern of results is very similar to that of Table 5a. As expected, differences between the methods are smaller, although they remain substantial in many rows of the table.

Table 5a

Summaries of Estimators Based on 1,000 Replicate Samples for $C = 10$ Adjustment Cells, Restricted to Sample Replicates with $n_{0c} > 0$ for all c. Response Rate of 52%. Values are Multiplied by 1,000

Association with Adjustment Cells Based on X				Unweighted Mean				Weighted Mean					Before Deletion Mean		Composite Mean	
Cell	(M, X)	(Y, X)	n	emp. bias	emp. rmse	analytical rmse[1]	est. rmse[2]	emp. bias	emp. rmse	Kish rmse[3]	analytical rmse[4]	est. rmse[5]	emp. bias	emp. rmse	emp. bias	emp. rmse
4	High	High	400	6,955	7,024	7,055	6,974	0	**1,057**	1,410	956	988	−38	795	0	1,057
			2,000	7,008	7,020	7,006	7,015	−2	**424**	608	427	434	12	342	−2	424
4	High	Medium	400	5,376	5,471	5,536	5,404	−33	**1,264**	1,510	1,216	1,297	−21	776	−33	1,264
			2,000	5,424	5,441	5,466	5,466	−41	**561**	650	545	559	−30	338	−41	561
4	Medium	High	400	3,664	3,794	3,809	3,754	−4	**816**	1,071	835	842	6	741	−4	816
			2,000	3,703	3,731	3,700	3,712	7	**369**	473	373	374	4	337	7	369
4	Medium	Medium	400	2,838	3,006	3,042	2,991	−18	**938**	1,095	954	970	−9	747	−18	938
			2,000	2,864	2,900	2,898	2,893	−2	**426**	483	426	428	6	335	−2	426
3	Low	High	400	476	1,148	1,113	1,178	40	**823**	1,050	823	828	30	764	40	823
			2,000	376	587	614	595	−11	**361**	465	368	368	−3	333	−11	361
3	Low	Medium	400	350	1,106	1,095	1,134	13	**927**	1,063	925	939	−16	762	13	927
			2,000	287	565	563	559	−20	**429**	470	413	414	−22	353	−20	429
2	High	Low(0)	400	56	**1,070**	1,056	1,275	96	1,658	1,613	1,518	1,631	28	793	83	1,410
			2,000	−11	**464**	473	567	−26	698	698	679	699	−19	337	−25	620
2	Medium	Low(0)	400	9	**1,042**	1,053	1,077	−27	1,122	1,112	1,097	1,125	21	772	−12	1,074
			2,000	−4	**474**	471	480	−11	491	491	491	493	11	340	−9	481
1	Low	Low(0)	400	−30	**1,038**	1,050	1,055	−30	1,053	1,064	1,050	1,076	−30	752	−30	1,040
			2,000	−2	**472**	469	469	−1	474	470	469	471	−8	343	−1	472

[1] Computed using Equation (7)
[2] Computed using Equation (11)
[3] Computed using Equation (15)
[4] Computed using Equation (6)
[5] Computed using Equation (13)

Table 5b
Summaries of Estimators based on 1,000 Replicate Samples for $C = 10$ Adjustment Cells, Restricted to Sample Replicates with $n_{0c} > 0$ for all c. Response Rate of 70%. Values are Multiplied by 1,000

Association with Adjustment Cells based on X				Unweighted Mean				Weighted Mean					Before Deletion Mean		Composite Mean	
Cell	(M, X)	(Y, X)	n	emp. bias	emp. rmse	analytical rmse[6]	est. rmse[7]	emp. bias	emp. rmse	Kish rmse[8]	analytical rmse[9]	est. rmse[10]	emp. bias	emp. rmse	emp. bias	emp. rmse
4	High	High	400	4,692	4,810	4,893	4,860	−133	**1,129**	1,192	889	894	−129	998	−133	1,129
			2,000	4,827	4,841	4,839	4,854	−20	**400**	529	398	405	−5	334	−20	400
4	High	Medium	400	3,581	3,716	3,855	3,733	−133	**1,266**	1,250	1,075	1,097	−128	917	−127	1,284
			2,000	3,763	3,784	3,778	3,777	−9	**501**	554	481	490	11	343	−9	501
4	Medium	High	400	2,666	2,812	2,878	2,837	−58	**803**	910	794	796	−49	772	−58	803
			2,000	2,732	2,760	2,767	2,761	−6	**353**	406	355	355	−9	333	−6	353
4	Medium	Medium	400	2,104	2,282	2,315	2,291	−28	**833**	924	854	861	−43	751	−28	833
			2,000	2,146	2,180	2,170	2,165	13	**370**	411	382	382	10	334	13	370
3	Low	High	400	217	906	954	980	−81	**797**	911	790	793	−77	771	−81	797
			2,000	312	513	506	502	2	**365**	405	353	353	4	349	2	365
3	Low	Medium	400	251	922	942	960	15	**804**	916	845	852	26	727	15	804
			2,000	224	454	472	471	−14	**370**	408	378	379	−15	327	−14	370
2	High	Low(0)	400	0	**952**	915	1,131	35	1,445	1,349	1,298	1,358	1	807	26	1,292
			2,000	−11	**416**	409	485	−41	608	598	580	599	−4	347	−31	535
2	Medium	Low(0)	400	22	**911**	910	920	24	942	936	930	946	2	757	21	925
			2,000	23	**418**	407	411	20	425	416	416	417	15	344	19	420
1	Low	Low(0)	400	1	**914**	914	912	2	917	916	914	926	−5	751	1	914
			2,000	4	**402**	408	408	4	403	409	408	410	6	331	4	402

[6] Computed using Equation (7)
[7] Computed using Equation (11)
[8] Computed using Equation (15)
[9] Computed using Equation (6)
[10] Computed using Equation (13)

4. Discussion

The results in sections 2 and 3 have important implications for the use of weighting as an adjustment tool for unit nonresponse. Surveys often have many outcome variables, and the same weights are usually applied to all these outcomes. The analysis of section 2 and simulations in section 3 suggests that improved results might be obtained by estimating the MSE of the weighted and unweighted mean and confining weighting to cases where this relationship is substantial. A more sophisticated approach is to apply random-effects models to shrink the weights, with more shrinkage for outcomes that are not strongly related to the covariates (*e.g.*, Elliott and Little 2000). A flexible alternative to this approach is imputation based on prediction models, since these models allow for interval-scaled as well as categorical predictors, and allow interactions to be dropped to incorporate more main effects. Multiple imputation (Rubin 1987) can be used to propagate uncertainty.

When there is substantial covariate information, one attractive approach to generalizing weighting class adjustments is to create a propensity score for each respondent based on a logistic regression of the nonresponse indicator on the covariates, and then create adjustment cells based on this score. Propensity score methods were originally developed in the context of matching cases and controls in observational studies (Rosenbaum and Rubin 1983), but are now quite commonly applied in the setting of unit nonresponse (Little 1986; Czajka, Hirabayashi, Little and Rubin 1987; Ezzati and Khare 1992). The analysis here suggests that for this approach to be productive, the propensity score has to be predictive of the outcomes. Vartivarian and Little (2002) consider adjustment cells based on joint classification by the response propensity and summary predictors of the outcomes, to exploit residual associations between the covariates and the outcome after adjusting for the propensity score. The requirement that adjustment cell variables predict the outcomes lends support to this approach.

The analysis presented here might be extended in a number of ways. Second order terms in the variance are ignored here, which if included would penalize weighting adjustments based on a large number of small adjustment cells. Finite population corrections could be included, although it seems unlikely that they would affect the main conclusions. It would be of interest to see to what extent the results can be generalized to complex sample designs involving clustering and stratification. Also, careful analysis of the bias and variance implications of nonresponse weighting on statistics other than means, such as subclass means or regression coefficients, would be worthwhile. We

expect it to be important that adjustment cell variables predict the outcome in many of these analyses too, but other points of interest may emerge.

Acknowledgements

This research is supported by grant SES-0106914 from the National Science Foundation. We thank an associate editor and three referees for useful comments on earlier drafts.

References

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A. and Rubin, D.B. (1987). Evaluation of a new procedure for estimating income aggregates from advance data. In *Statistics of Income and Related Administrative Record Research*: *1986-1987*, U.S. Department of the Treasury, 109-136.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Ezzati, T., and Khare, M. (1992). Nonresponse adjustments in a National Health Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 339-344.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663-685.

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical, A*, 142, 33-46.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

Little, R.J.A. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.

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

Little, R.J.A., Lewitzky, S. Heeringa, S., Lepkowski, J. and Kessler, R.C. (1997). An assessment of weighting methodology for the national comorbidity study. *American Journal of Epidemiology*, 146, 439-449.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley & Sons, Inc.

Oh, H.L., and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*, 2, Theory and Bibliographies, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), Academic Press, New York, 143-184.

Robins, J.M., Rotnitsky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

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

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

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

Vartivarian, S., and Little, R.J.A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 169-182
Statistics Canada, Catalogue No. 12-001-XPB

Variance-Covariance Functions for Domain Means of Ordinal Survey Items

Alistair James O'Malley and Alan Mark Zaslavsky [1]

Abstract

Estimates of a sampling variance-covariance matrix are required in many statistical analyses, particularly for multilevel analysis. In univariate problems, functions relating the variance to the mean have been used to obtain variance estimates, pooling information across units or variables. We present variance and correlation functions for multivariate means of ordinal survey items, both for complete data and for data with structured non-response. Methods are also developed for assessing model fit, and for computing composite estimators that combine direct and model-based predictions. Survey data from the Consumer Assessments of Health Plans Study (CAHPS®) illustrate the application of the methodology.

Key Words: Variance function; Correlation function; Hierarchical model; Ordinal response; Nonresponse; Skip pattern.

1. Introduction

Survey data are often used to obtain measures for comparisons across estimation domains. In our motivating example, surveys are conducted to elicit reports on experiences with health plans (entities administering health care) from enrolled members; similarly a survey might assess schools by administering tests to a sample of students.

An essential part of the analysis of survey data is the calculation of sampling variances, or the sampling-covariance matrix of a multivariate estimator. The standard survey sampling approach is to compute variances directly for each estimator in each domain. Direct variance estimates may be unstable when the number of respondents to an item is small because the sample size for a domain is small, because the item is applicable to only a fraction of respondents (such as users of specialized equipment in health surveys), or because we are interested in means for a small subgroup (such as those with chronic illnesses).

By modeling variance estimates as functions of the unit (domain) means, we can pool information across units to obtain more stable estimates. Although modeling may introduce bias, for small units this is offset by the reduction in sampling variation. One may also consider generalizing variance estimates across items in addition to or instead of domains. This will be appropriate when there are groups of items for which the same mean-variance relationship is likely to hold. However, when there are many more domains than items, the greatest potential gain is from generalizing across domains rather than across items.

A *Generalized Variance Function* (GVF) is a mathematical model describing the relationship between the variance or relative variance of a survey estimator and its expectation. When multiple estimates are produced from the same sample, Wolter (1985, chapter 5) proposes the model

$$V / M^2 = \theta_0 + \theta_1 / M,$$

where M and V denote the expected value and variance of the estimator respectively. Such a form might be suitable for variables such as income or wealth for which a nearly constant coefficient of variation might be plausible because the mean and standard deviation are proportional to the length of the reference period. Modeling the coefficient of variation is thus most suited to situations where the variables are similar in content but have different scales with unrestricted ranges (*e.g.*, income collected monthly and yearly). In our problem the items are ordinal and so a model of the coefficient of variation is not a natural choice. Other proposed GVFs also have simple forms (Woodruff 1992; Otto and Bell 1995).

If a suitable GVF can be found, it can simplify calculations and make variance estimates more stable. Furthermore, summarizing sampling variance estimates in the form of a function also facilitates presentation of large volumes of statistics (Wolter 1985, pages 201-202). Finally, modeling variances as functions of means facilitates iterative re-estimation of sampling variances in hierarchical modeling. In practice the decision to use variance functions in a hierarchical modeling context depends on the goodness of the fit of the GVF; only with a sufficiently good fit is use of the GVF worthwhile.

Past work on GVFs is relatively sparse. Wolter (1985, chapter 5) gave an overview but provided only a few references, as did Valliant, Dorfman and Royall (2000,

1. Alistair James O'Malley and Alan Mark Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899, U.S.A. E-mail: omalley@hcp.med.harvard.edu and zaslavsk@hcp.med.harvard.edu.

pages 344–348). Valliant (1992a, 1992b) used GVFs to smooth time-dependent indices in time series analysis. Woodruff (1992) used GVFs for variance estimation of employment change in the Current Employment Survey, and Wolter (1985, pages 208–217) illustrates the use of GVFs on data from the Current Population Survey. GVFs are also used in the National Health Interview Survey (Valliant *et al.* 2000, page 344).

Huff, Eltinge, and Gershunskaya (2002) and Cho, Eltinge, Gershunskaya and Huff (2002) considered GVFs for the United States Current Employment Survey and Consumer Expenditure Survey. Eltinge (2002) uses GVFs to estimate a full sampling covariance matrix when samples are too small to produce stable estimates for all areas, estimating the components of the mean squared error (MSE) of the GVF model. Otto and Bell (1995) fit GVFs to median income, per capita income, and age-group poverty rates in the Current Population Survey, assuming an autoregressive dependence between rates over time and a Wishart distribution for the sampling covariance matrices.

Our research extends previous research on GVFs in four directions. First, we use the GVF to generalize across domains rather than items. Thus, we do not assume that different items have the same GVF, although it might be reasonable to fit models of the same form for items with similar response categories. Second, we develop GVFs for the full covariance matrix, which must be estimated for joint inference on multiple outcomes. Thirdly, we focus on the relationship between means and variances of items with the ordinal response formats often used in survey questionnaires, rather than on homoscedastic continuous responses. Finally, we explicitly allow for patterns of nonresponse due to structured skip patterns. While structured item nonresponse can be ignored (except for its effect on sample size) in univariate estimation, it must be considered explicitly to model bivariate relationships because it affects the sampling covariance of item means. Furthermore, because the number of responses varies across items, we cannot model the sampling covariances using a Wishart distribution, which has only a single parameter for sample size.

We first describe direct estimation of variances and covariances, including the case when data are missing due to skip patterns. In section 3 we introduce models for generalized variance and covariance functions (GVCFs) and lay out our strategies for model fitting and evaluation and for combining direct estimates and model predictions. In section 4, we apply our methods to a major health care survey. In section 5, we conclude by describing applications and extensions of our methods.

2. Direct Estimates of Sampling Variances of Domain Means

We index observations by domain h, items (indices i and j), and respondents (indices k and l); $y_{h,ik}$ and $r_{h,ik}$ denote the outcome and response indicator of subject k in domain h on item i. We suppress the index for item when referring to all items for a respondent or domain, and have no need for the subscript for respondent when discussing the means, variances, and correlations of items.

Direct estimation of the sampling covariance matrix of domain means (henceforth, "variance estimation") begins by expressing the means as functions of totals of the outcomes and response indicators. We replace $y_{h,ik}$ with 0 for missing observations so that totals are defined in the presence of skip patterns. Following the notation of Särndal, Swenson and Wretman (1992, pages 24–28; 36–42), let U_h and S_h describe the population and sample respectively for the h^{th} domain, $Y_{h,i} = \sum_{U_h} y_{h,ik}$, $R_{h,i} = \sum_{U_h} r_{h,ik}$, $\hat{Y}_{h,i} = \sum_{S_h} \breve{y}_{h,ik}$, and $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$, where $\breve{y}_{h,ik} = y_{h,ik} / \pi_{h,k}$, $\breve{r}_{h,ik} = r_{h,ik} / \pi_{h,k}$, and $\pi_{h,k} = \text{pr}(k \in S_h)$.

The vector of mean outcomes for the population of elements within domain h is

$$M_h = f(Y_h, R_h) = \left(\frac{Y_{h,1}}{R_{h,1}}, \ldots, \frac{Y_{h,I}}{R_{h,I}} \right),$$

where $Y_h = (Y_{h,1}, \ldots, Y_{h,I})$ and $R_h = (R_{h,1}, \ldots, R_{h,I})$. An estimator is

$$f(\hat{Y}_h, \hat{R}_h) = \left(\frac{\hat{Y}_{h,1}}{\hat{R}_{h,1}}, \ldots, \frac{\hat{Y}_{h,I}}{\hat{R}_{h,I}} \right).$$

A first order Taylor series expansion of $f(\hat{Y}_h, \hat{R}_h)$ about $f(Y_h, R_h)$ produces the approximation

$$\text{var}(f(\hat{Y}_h, \hat{R}_h)) \approx V_h = f'(Y_h, R_h)\, \text{var}(\hat{Y}_h, \hat{R}_h)\, f'(Y_h, R_h)^T,$$

where $f'(Y_h, R_h)$ is the Jacobian of $f(Y_h, R_h)$. Often it is computationally easier to first calculate $u_{h,k} = f'(Y_h, R_h) z_{h,k}$, where $z_{h,k} = (y_{h,k}, r_{h,k})$, and then evaluate the variance as

$$\begin{aligned} V_h &= \text{var}\left(\sum_{S_h} \breve{u}_{h,k} \right) \\ &= \text{var}\left(\sum_{U_h} \breve{u}_{h,k} I_{h,k} \right) \\ &= \sum_{k,l \in U_h} \Delta_{h,kl} \breve{u}_{h,k} \breve{u}_{h,l}^T, \end{aligned}$$

where $I_{h,k} = 1$ if $k \in S_h$ (indicating that the k^{th} member of domain h is sampled) and 0 otherwise, $\Delta_{h,kl} = \pi_{h,kl} - \pi_{h,k}\pi_{h,l}$, and $\pi_{h,kl} = \text{pr}(k, l \in S_h)$. An estimator for V_h is

$$\hat{V}_h = \sum_{k,l \in S_h} \breve{\Delta}_{h,kl}\, \breve{u}_{h,k}\, \breve{u}_{h,l}^T, \qquad (1)$$

where $\breve{\Delta}_{h,kl} = \Delta_{h,kl} / \pi_{h,kl}$.

To describe evaluation of $\hat{V}_h$ one need only consider one diagonal (*i.e.*, variance) element and one off-diagonal (*i.e.*, covariance) element. The sub-matrix of the Jacobian formed by the i^{th} and j^{th} items is given by

$$f'(Y_h, R_h) = \begin{pmatrix} \dfrac{1}{R_{h,i}} & 0 & -\dfrac{Y_{h,i}}{R_{h,i}^2} & 0 \\ 0 & \dfrac{1}{R_{h,j}} & 0 & -\dfrac{Y_{h,j}}{R_{h,j}^2} \end{pmatrix}.$$

For, $z_{h,k} = (y_{h,ik}, y_{h,jk}, r_{h,ik}, r_{h,jk})$, it follows that

$$u_{h,k} = f'(Y_h, R_h)\, z_{h,k} = \begin{pmatrix} \dfrac{1}{R_{h,i}}(y_{h,ik} - M_{h,i}\, r_{h,ik}) \\ \dfrac{1}{R_{h,j}}(y_{h,jk} - M_{h,j}\, r_{h,jk}) \end{pmatrix},$$

where $M_{h,i} = Y_{h,i} / R_{h,i}$ is the mean outcome of the i^{th} item in domain *h*. Hence,

$$\hat{V}_{h,ii} = \frac{1}{R_{h,i}^2} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} (\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,il} - M_{h,i}\, \breve{r}_{h,il}) \quad (2)$$

and

$$\hat{V}_{h,ij} = \frac{1}{R_{h,i}\, R_{h,j}} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} (\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik}) \times (\breve{y}_{h,jl} - M_{h,j}\, \breve{r}_{h,jl}). \qquad (3)$$

To evaluate (2) and (3), we make a further approximation by substituting $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$ and $\hat{M}_{h,i} = \sum_{S_h} \breve{y}_{h,ik} / (\sum_{S_h} \breve{r}_{h,ik})$ for $R_{h,i}$ and $M_{h,i}$.

When sampling rates are small, or if we wish to make predictions for a large super-population (*e.g.*, all potential enrollees in a health plan, not just those currently enrolled), $\breve{\Delta}_{h,kl} = 1 - \pi_{h,k} \approx 1$ if $k = l$, $\breve{\Delta}_{h,kl} \approx 0$ if $k \neq 1$, and the sampling design approaches sampling with replacement. Under the sampling with replacement design, approximately unbiased estimators are

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{h,i}^2} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})^2 \qquad (4)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{h,i}\, \hat{R}_{h,j}} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,jk} - \hat{M}_{h,j}\, \breve{r}_{h,jk}). \qquad (5)$$

These estimators can be generalized to accommodate clustering.

With equal-probability sampling within domains, (4) and (5) reduce to

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{S_{h,i}}^2} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})^2 \qquad (6)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{S_{h,i}} \hat{R}_{S_{h,j}}} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})(y_{h,jk} - \hat{M}_{h,j}\, r_{h,jk}), \qquad (7)$$

where $\hat{R}_{S_{h,i}}$ is the number of respondents to item *i* in domain *h*.

3. Models for Variance Functions

In this section we propose specifications for models for variances and for sample correlations with complete responses or with structured skipped responses. We then discuss model fitting and evaluation strategies. We assume that these domains are nonoverlapping strata, so the sampling errors for different domains are independent.

We transform the ordinal ratings to the $[0,1]$ interval by the transformation $p_{h,i} = (B_{h,i} - M_{h,i}) / (B_{h,i} - A_{h,i})$, where $A_{h,i}$ and $B_{h,i}$ are the minimum and maximum response categories for item *i* in domain *h* respectively. We focus on modeling variances for large values of $M_{h,i}$ (small values of $p_{h,i}$) because in our motivating example mean outcomes are typically near the high end of the scale.

3.1 Variance Functions

To account for the variable number of respondents over domains and items, and differing scales, we normalize the variance estimators in (6) for sample size and re-scale:

$$\tilde{V}_{h,ii} = \frac{\hat{R}_{S_{h,i}}\, \hat{V}_{h,ii}}{(B_{h,i} - A_{h,i})^2}.$$

With unequal probability sampling within domains, a normalization factor could be used that accounts for the weights. One possible normalization is to multiply $\hat{V}_{h,ii}$ by $\hat{R}^*_{S_{h,i}} = (\sum \breve{r}_{h,ik})^2 / (\sum \breve{r}^2_{h,ik})$, where $\breve{r}_{h,ik}$ is the response indicator for item *i* for the k^{th} subject in the h^{th} domain, in place of $\hat{R}_{S_{h,i}}$. This approximation, proposed in Kish (1965), has a model based justification (Gabler, Haeder and Lahiri 1999). It works well if the sampling probabilities vary modestly in the sample, but can lead to inefficiency if the variation is excessive (Korn and Graubard 1999, page 173; Spencer 2000).

Because the items in our example have ordinal scales, the variance must go to 0 as $p_{h,i} \to 0$ or $p_{h,i} \to 1$. An obvious predictor with this property is the variance function of the Bernoulli distribution, $p_{h,i}(1 - p_{h,i})$. This holds exactly for

dichotomous items, and might be a useful approximation for items with three or more categories.

As alternatives to the Bernoulli variance model we considered models with a variety of polynomial and other functions of the means as predictors. Of all the models considered, the quadratic family of models were found to fit as well as any. We focused on the following quadratic models.

$$\text{Model V1:}\quad \tilde{V}_{h,ii} = \beta_{1i}\, p_{h,i}, \tag{8}$$

$$\text{Model V2:}\quad \tilde{V}_{h,ii} = \beta_{2i}\, p_{h,i}(1-p_{h,i}), \tag{9}$$

$$\text{Model V3:}\quad \tilde{V}_{h,ii} = \beta_{1i}\, p_{h,i} + \beta_{2i}\, p_{h,i}(1-p_{h,i}). \tag{10}$$

Thus we consider a linear variance model V1, a binomial-like model V2, and a general quadratic variance model V3. All models correctly ensure $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 0$, but only V2 ensures that $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 1$. The rationale behind V1 is that relationships are often approximately linear over small intervals. Both V1 and V2 are submodels of the two-parameter quadratic V3. We also considered models for $\log(\tilde{V}_{h,ii})$, but these models did not fit as well.

The model V3 is equivalent to the model suggested by Wolter (1985, chapter 5); the equivalence is seen by expressing the right-hand side of V3 in terms of $p_{h,i}$ and $p_{h,i}^2$, and then dividing both sides by $p_{h,i}^2$ to obtain the relative variance. However, parameter estimates obtained by fitting the two forms of the model may be different depending on the modeling assumptions used.

3.2 Correlation Functions with Complete Data

Because correlations are independent of the scale of the data, we model the correlations and derive the sampling covariances, rather than modeling the covariances directly. We model the sample correlations

$$\hat{\rho}_{h,ij} = \frac{\hat{V}_{h,ij}}{(\hat{V}_{h,ii}\,\hat{V}_{h,jj})^{1/2}},$$

via the unrestricted transformed values $Z_{h,ij} = \log\{(1+\hat{\rho}_{h,ij})/(1-\hat{\rho}_{h,ij})\}$. Unlike the variance models, models for correlations may include an unrestricted intercept, since there is no natural restriction on the correlation when $p_{h,i}$ or $p_{h,j}$ approaches 0 or 1.

Because $\hat{\rho}_{h,ij}$ is a function of the first and second moments of items *i* and *j*, it seemed reasonable to first focus on linear and quadratic models for $Z_{h,ij}$. As with variance functions, we found that a more extensive range of models (*e.g.*, models with logarithms of the means as predictors) did not substantially improve model fit. We ultimately focused on the following nested series of models.

$$\text{Model C1:}\quad Z_{h,ij} = \alpha_{0ij}, \tag{11}$$

$$\text{Model C2:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{3ij}\; p_{h,i}\; p_{h,j}, \tag{12}$$

$$\text{Model C3:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij}(p_{h,i} + p_{h,j}) + \alpha_{3ij}\; p_{h,i}\, p_{h,j}, \tag{13}$$

$$\text{Model C4:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij}\, p_{h,i} + \alpha_{2ij}\, p_{h,j} + \alpha_{3ij}\; p_{h,i}\, p_{h,j}, \tag{14}$$

$$\text{Model C5:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij}\, p_{h,i} + \alpha_{2ij}\, p_{h,j} + \alpha_{3ij}\; p_{h,i}\, p_{h,j} + \alpha_{4ij}\; p_{h,i}^2 + \alpha_{5ij}\; p_{h,j}^2, \tag{15}$$

Model C3 is model C4 with the constraint $\alpha_{1ij} = \alpha_{2ij}$.

3.3 Predicting Covariances with Structured Missing Data

When the data have skip patterns, the sample correlations of the ratings for the set of respondents who answered both items can be modeled by (11)–(15), as in the complete response case. The corresponding sample covariances can be easily estimated by using the fitted variance functions to re-scale the predicted correlations. However, because the sampling covariance reflects the variability in the whole sampling process, not just the variability within the sub-population of respondents who answered both items, the relationship between sample covariance and sampling covariance is more complicated than if the data were complete. In this section we derive the relationship between the sample covariance for the set of respondents who answered both items and the sampling covariance. This allows correlation models such as (11)–(15) to be applied to data with skip patterns.

There are four distinct data patterns for any pair of items: response to both items, one response and one skipped item (two patterns), and both items skipped. We extend our notation by introducing a superscript representing the response status of a second item. Let $\hat{Y}_{h,ij}^1 = \sum_{S_h} \breve{y}_{h,ik}\breve{r}_{h,jk}$, $\hat{Y}_{h,ij}^0 = \sum_{S_h} \breve{y}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{R}_{h,ij}^1 = \sum_{S_h} \breve{r}_{h,ik}\breve{r}_{h,jk}$, $\hat{R}_{h,ij}^0 = \sum_{S_h} \breve{r}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{M}_{h,ij}^1 = \hat{Y}_{h,ij}^1 / \hat{R}_{h,ij}^1$, $\hat{M}_{h,ij}^0 = \hat{Y}_{h,ij}^0 / \hat{R}_{h,ij}^0$. Then

$$\hat{M}_{h,i} = \frac{\hat{R}_{h,ij}^1\,\hat{M}_{h,ij}^1 + \hat{R}_{h,ij}^0\,\hat{M}_{h,ij}^0}{\hat{R}_{h,i}}.$$

In the equal probability sampling case, substitution of the above expression for $\hat{M}_{h,i}$ into (7) yields

$$\tilde{V}_{h,ij} = \frac{\hat{R}_{h,ij}^1}{\hat{R}_{h,i}\,\hat{R}_{h,j}}\left\{\hat{C}_{h,ij}^1 + \frac{\hat{R}_{h,ij}^0\,\hat{D}_{h,ij}\,\hat{R}_{h,ji}^0\,\hat{D}_{h,ji}}{\hat{R}_{h,i}\,\hat{R}_{h,j}}\right\}, \tag{16}$$

where $\hat{D}_{h,ij} = \hat{M}^1_{h,ij} - \hat{M}^0_{h,ij}$. Here $\hat{C}^1_{h,ij} = \sum_S (\breve{y}_{h,ik} - \hat{M}^1_{h,ij}\,\breve{r}_{h,ik})(\breve{y}_{h,ik} - \hat{M}^1_{h,ji}\,\breve{r}_{h,jk}) / \hat{R}^1_{h,ij}$ is the normalized sample covariance of the ratings for the set of respondents who answered both items (which can be predicted using correlation and variance functions, and in the case of unequal probability sampling applying a normalization factor). When the sampling probabilities are not equal, Equation (16) holds exactly only if $\sum_S \breve{r}_{h,jk}(\breve{y}_{h,ik} - \hat{M}^1_{h,ik}\,\breve{r}_{h,ik}) = 0$. Therefore, (16) may be expected to provide a good approximation if the sampling probabilities for one item are not highly correlated with the residuals for another item. In general, the appropriateness of using (16) for unequal probability sampling designs should be checked.

The estimated mean differences $\hat{D}_{h,ij}$ determine the contribution of the response pattern to the sampling covariance. Either $\hat{D}_{h,ij}$ or $\hat{D}_{h,ij}\,\hat{D}_{h,ji}$ may be modeled in the process of obtaining smoothed estimates of $\tilde{V}_{h,ij}$. In our application, the $\hat{D}_{h,ij}$ were typically small. Because the second term of (16) is a product of two factors of small magnitude ($\hat{D}_{h,ij}$ and $\hat{D}_{h,ji}$), the contribution of $\hat{D}_{h,ij}$ to (16) was small and it sufficed to use a simple model for $\hat{D}_{h,ij}$, such as a constant for each item pair. However, unique constants should be estimated for each pair of items.

3.4 Model Fitting and Evaluation

We estimate the parameters of the variance or correlation function using iteratively reweighted least squares regression. Weighting is important when the number of responses varies greatly across domains, as in our motivating example.

In this section we index domains (h) and respondents (k) but not items as the same methodology applies to each variance and correlation model. Exact computations are derived for the equal probability sampling case, and approximations are noted for the unequal probability sampling case. Generically, the direct estimators $\tilde{f}_h$, true values f_h, and model predictions $\hat{f}_h$ are related through the hierarchical model

$$\text{Level I}: \quad \tilde{f}_h = f_h + \in_h, \tag{17}$$

$$\text{Level II}: \quad f_h = \hat{f}_h + e_h, \tag{18}$$

where $\in_h \sim [0, \sigma^2_h / \hat{R}_{S_h}]$, $e_h \sim [0, \tau^2]$, and $[\mu, \sigma^2]$ indicates a distribution with expectation μ and variance σ^2 but unspecified form. In the unequal probability sampling case we replace $\hat{R}_{S_h}$ with $\hat{R}^*_{S_h}$. Here $\in_h$ represents sampling error and e_h represents model error. Marginally, $\tilde{f}_h = \hat{f}_h + e_h + \in_h$ so in the regression we weight the observation for domain h by $w_h = (\tau^2 + \sigma^2_h / \hat{R}_{S_h})^{-1}$, the inverse of the marginal variance. With equal-probability sampling, the variance of the direct estimate of $\sigma^2_h = E[\tilde{f}_h - f_h]^2$ is given by

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{1}{\hat{R}_{S_h} - 1}\left\{\frac{1}{\hat{R}_{S_h}}\sum_{k\in S}(y_{h,k} - \hat{M}_h r_{h,k})^4 - (1 - \frac{3}{\hat{R}_{S_h}})\tilde{f}^2_h\right\}$$
if f is a variance (19)

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{4}{\hat{R}_{S_h} - 3} \text{ if } f \text{ is a transformed correlation.} \tag{20}$$

In the equal probability sampling case Equation (19) is exact and does not depend on parametric assumptions (Seber 1977, page 14). The asymptotic approximation (20) to the variance of the transformed correlation Z_h (Freund and Walpole 1987, page 477) deteriorates as sample sizes decrease, and fails altogether for $\hat{R}_{S_h} \le 3$. However, domains with small sample sizes have little impact on the fitted models; we exclude domains with $\hat{R}_{S_h} \le 3$ from correlation modeling.

When the sampling probabilities are not equal, the large sample counterpart to (19), given by

$$\hat{\sigma}^2_h(\tilde{f}_h) = \sum_{k\in S}\left\{\begin{array}{l}\dfrac{(\breve{y}_{h,k} - \hat{M}_h\breve{r}_{h,k})^2}{\sum_{l\in S}\breve{r}^2_{h,l}} - \dfrac{2w_h}{\sum_{l\in S}\breve{r}_{h,l}} \\ \times(\breve{y}_{h,k} - \hat{M}_h\breve{r}_{h,k}) - \dfrac{\tilde{f}^2_h}{\sum_{l\in S}\breve{r}^2_{h,l}}\breve{r}^2_{h,k}\end{array}\right\}^2,$$

where $w_h = (\sum_S \breve{y}_{h,l}\breve{r}_{h,l}) / \sum_S \breve{r}^2_{h,l}) - \hat{M}_h$, may be used. In the equal probability sampling case, $w_h = 0$ and the above expression reduces to a non-bias corrected version of (19). If the sampling probabilities are not equal, we suggest replacing (20) with the design-effect-corrected estimator

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{4}{\hat{R}^*_{S_h} - 3}.$$

The model error variance τ^2 is estimated as:

$$\hat{\tau}^2 = \max\left\{0, \hat{\text{MSE}} - \frac{\sum \hat{R}_{S_h}\hat{\sigma}^2_h(\tilde{f}_h)}{\sum \hat{R}_{S_h}}\right\},$$

where $\hat{\text{MSE}} = \sum_h q_h(\tilde{V}_h - \hat{f}_h)^2$, $q_h = N\,\hat{R}_{S_h} / \sum_h \hat{R}_{S_h}$, and $N = \sum_h I(\hat{R}_{S_h} > 0)$. The weights are then re-estimated as $\hat{w}_h = (\tau^2 + \hat{\sigma}^2_h(\tilde{f}_h) / \hat{R}_{S_h})^{-1}$, and the GVCF models are refit, iterating to convergence. We again suggest replacing $\hat{R}_{S_h}$ with $\hat{R}^*_{S_h}$ if the sampling probabilities are not equal.

We compared the predictive accuracy of models using $R^2 = 1 - \hat{\text{MSE}} / \hat{\text{MSV}}$, where $\hat{\text{MSE}}$ is the mean squared error of the regression, and $\hat{\text{MSV}}$ is the sample size weighted average of the sampling variances of the direct estimators (variances or transformed correlations) for each

domain. Note that we could have $R^2 < 0$ for a very poorly fitting model.

3.5 Combined Estimators

For domains with small samples, direct survey variance estimates often are too imprecise to be useful, while estimates for larger domains in the same study may be quite reliable. Fay and Herriot (1979) and Ghosh and Rao (1994) demonstrated that shrinking direct estimates towards a model-based smoothed value can lead to substantial gains in precision. They proposed composite or empirical Bayes estimators that are weighted averages of direct and model-based estimators. That is, instead of either using the direct estimates or estimates obtained from generalized variance/ covariance modeling, we use a weighted average of the two estimators to potentially obtain even better estimates.

Such weighted estimators can be constructed for domain variances using the model specified in (17) and (18). A natural approach is to weight the direct model-based estimators inversely proportional to the corresponding sampling and model error variances respectively (denoted σ_h^2 and τ^2 respectively for domain *h*). The resulting estimator for domain *h* (for variances and transformed correlations) is:

$$\tilde{f}_h = \frac{\hat{\tau}^2\,\tilde{f}_h^{\,\mathrm{dir}} + \hat{\sigma}_h^2\,\tilde{f}_h^{\,\mathrm{mod}}}{\hat{\tau}^2 + \hat{\sigma}_h^2} = \tilde{f}_h^{\,\mathrm{dir}} + \frac{\hat{\sigma}_h^2}{\hat{\tau}^2 + \hat{\sigma}_h^2}\left(\tilde{f}_h^{\,\mathrm{mod}} - \tilde{f}_h^{\,\mathrm{dir}}\right),$$

where $\tilde{f}_h^{\,\mathrm{dir}}$ and $\tilde{f}_h^{\,\mathrm{mod}}$ denote the direct and model-based estimators. This generic formula applies to the variance estimates for all items, and correlation estimates for all pairs of items. The right-most expression has the form of an empirical Bayes estimator.

If the direct and model-based variance estimators are independent, the variance of the resulting combined estimator is $\tau^2\sigma_h^2/(\tau^2 + \sigma_h^2) \le \min\{\tau^2, \sigma_h^2\}$. Thus the composite is as least as precise as either of its two component estimators, improving on ad hoc selection between direct and model-based predictions. This is a useful strategy especially when model-based predictions improve on direct estimates for some, but not all domains.

4. Example: CAHPS® Data Set

The Consumer Assessments of Health Plans Study (CAHPS®) survey (Goldstein, Cleary, Langwell, Zaslavsky and Heller 2001) was designed primarily to elicit consumer ratings and reports on health plans. Plan mean scores (perhaps after recoding) on the various survey items are calculated and reported to consumers, health plans, and purchasers. Each analytic domain consists of the enrollees of a health plan (or geographically defined portion of one) in a year; most of the plans are sampled in multiple years. The stratum is the reporting unit (plan or portion thereof) in a given year; reporting units corresponded to plans with the exception of a few large plans that had multiple reporting units. Therefore, there are many units for variance and covariance function estimation.

We illustrate our methods with a CAHPS data set for beneficiaries of U.S. Medicare managed care plans, a system of private but government-funded entities serving from 5.7 to 6.9 million elderly or disabled beneficiaries in each year during our study period (1997 to 2001). Our data represent 381 reporting domains each sampled in 1 to 5 years for a total of 932 distinct reporting unit by year domains with 705,848 responses. Because samples are drawn independently each year, patients may be sampled in multiple years. However, repeated sampling is rare and can be overlooked for our analysis. Therefore, the domains are strata with equal probability element sampling performed within each. Note that in CAHPS analyses no corrections are made for finite-population sampling since the data are collected to guide choices for future years rather than to record experiences of the specific population in a particular year.

CAHPS items use a variety of ordinal response formats with either 11, 4, 3, or 2 response options. Overall ratings of doctor, specialist, care, and plan are measured on a 0 to 10 scale from "worst possible" to "best possible". Other items use a 4–point ordinal "frequency" scale (never/sometimes/ usually/always), or a 3–point ordinal "problem" scale (not a problem/somewhat a problem/a big problem), or are dichotomous (no/yes). Many items are answered only by respondents who used particular services or had particular needs, as determined by screener items. For example, an item about whether advice was obtained successfully by telephone is only answered by those who first reported that they attempted to obtain advice in that way.

4.1 Descriptive Statistics

Table 1 presents response distributions and domain mean distributions by item type. Missing observations due to structured skip patterns often occurred in blocks, with as many as 11 items skipped on the basis of a single screening question. Very little nonresponse (less than 2% on almost all items) was not due to a structured skip pattern. In this analysis we treat all types of nonresponse identically.

Item response rates were lowest (as low as 4%) for problem items, several of which dealt with specialty services such as therapy or home health care needed by relatively few respondents. Some of the frequency and yes/no items also had low response rates. The greatest variation in the proportions of skipped items was evident among the yes/no items: 96.7% for a "complaint or problem

with plan" to 12.5% for "get prescription through plan". Domain mean outcomes are in general concentrated towards the higher end of their scales, indicating that most responses were favorable.

Table 1
Distribution of Responses and Ratings Evaluated over Items of the Same Type (n = 705,848 Respondents)

Statistic	Numerical	How Often	Problem	Yes/No
Number of items	4	11	11	9
Percentage responding				
Mean	74.97	62.56	30.32	57.26
Minimum	50.90	27.70	4.00	12.50
Maximum	95.00	74.50	64.40	96.70
Item means				
Mean	8.76	3.57	2.70	1.78
Minimum	8.57	3.09	2.49	1.62
Maximum	8.88	3.84	2.86	1.97
Distribution of ratings (across items in group)				
0	0.5			
1	0.4	2.0	5.7	19.5
2	0.4	6.3	12.1	80.5
3	0.7	23.9	82.2	
4	0.9	67.8		
5	4.6			
6	3.0			
7	6.2			
8	16.1			
9	17.8			
10	49.5			

Items are on a 0–10 numerical scale from "worse possible" to "best possible", a 4–point 1–4 ordinal "frequency" scale (never/sometimes/usually/always), a 3–point 1–3 ordinal "problem" scale (not a problem/somewhat a problem/a big problem), or are dichotomous 1–2 items (no/yes).

The domain mean, minimum, and maximum values across all items of the same type are also presented in Table 1. These illustrate that the 0–10 items have the smallest total variation (after rescaling to the common 0–1 range), while the 1–2 items have the largest total variation across domains and items. This is also illustrated in Figure 1, where we observe that the distribution of the 1–2 items varies substantially across items whereas the distributions of the 0–10 items are more homogeneous.

Table 2 presents statistical summary measures for the means and standard deviations of the domain mean ratings, evaluated across items of the same type. This complements Figure 1 by summarizing the difference in distributions of items within a given scale. Items with more response categories are concentrated towards the top of the scale and hence have smaller variance. For example, the mean standard deviation of the 1–2 items (0.36) is twice that of the rescaled 0–10 items (0.172). With the exception of the 0–10 items, the distributions of domain mean ratings vary greatly across items of the same type. For instance, the standard deviation of the means of 1–2 items across items is 0.30 compared to a rescaled standard deviation of 0.03 for the 0–10 items.

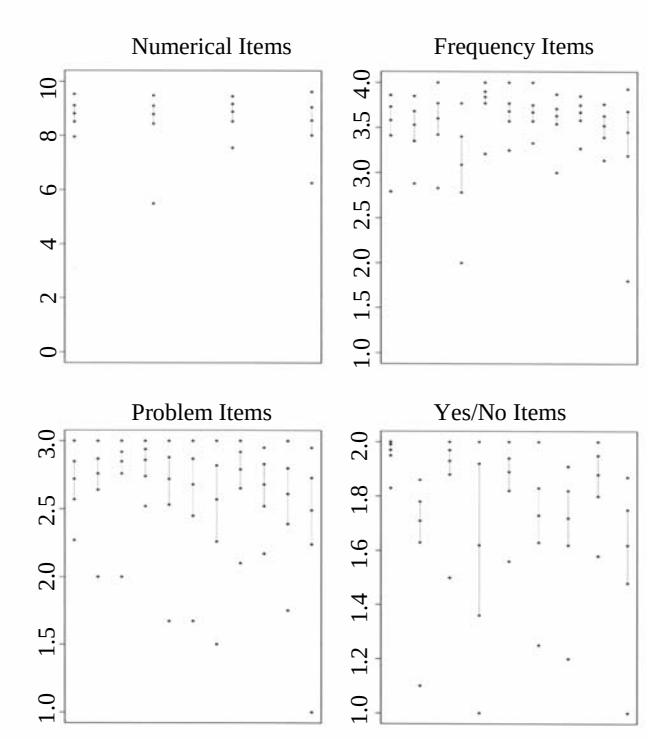


Figure 1. Five-point Summary of the Domain Sample Means for Each Item. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

Table 2
Summary Statistics of Domain Means and Standard Deviations Evaluated Over Domains and Items

Type	Summary Statistics for: Item Means				Item SDs	
	Min	Max	Mean	SD	Mean	SD
Numerical 0–10	6.82	9.52	8.76	0.30	1.72	0.26
Frequency 1–4	2.86	3.90	3.57	0.12	0.66	0.09
Problem 1–3	1.88	2.99	2.70	0.14	0.57	0.13
Yes/No 1–2	1.34	1.96	1.78	0.08	0.36	0.06

Note: Columns 2 through 5 give the minimum, maximum, mean, and standard deviation of the domain item means across items of a given type. Columns 6 and 7 give the mean and standard deviation of the domain item standard deviations across items of a given type.

Sample correlations also varied greatly across the pairs of items (Figure 2), although most were positive. Correlations between items of the same type most often were higher than those between items of different types. The numerical 0–10 ratings had the largest correlations (mean = 0.49), and generally ratings with more categories tended to have higher correlations than ratings with fewer categories. Although most of the pairs of 1–4 items had mean correlations near to 0.5, one item was negatively correlated with the others (revealed by the cluster of mean correlations below 0); this arose from reverse coding an item whose overall sample mean was not in the top half of the scale. The distributions

of the correlations of pairs of 1–2 items were centered near 0, indicating that pairs of items of this type often have negative correlations. Complete item wordings and additional summary statistics appear in Zaslavsky, Beaulieu, Landon and Cleary (2000) and Zaslavsky and Cleary (2002).

Models fitted to the variances and correlations are presented in the remainder of this section. Extensive checking of the best-fitting models indicated that the residuals did not follow any discernible pattern.

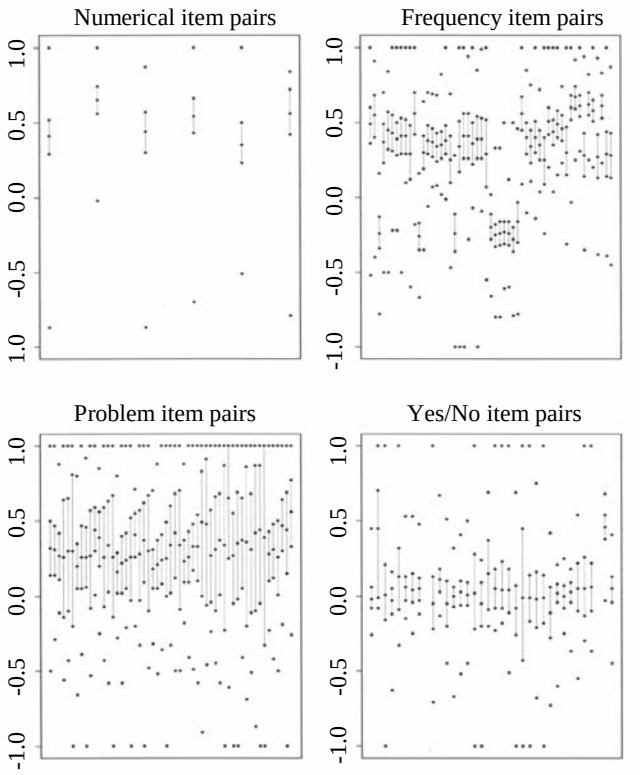


Figure 2. Five-point summary of the domain sample correlations between items with the same type. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

4.2 Variance Functions

In preliminary investigations not reported here, we fit two models within groups of items with the same response scale, one with common and one with different regression parameters for each item, to the data set comprising all of the items. Comparisons of the overall fits of the models (using criteria such as Mallow's C_p, R^2, adjusted R^2) and tests of the significance of effect-item interactions demonstrated that allowing parameters to vary across items significantly improved model fit. For instance, for the rescaled numerical ratings, weighted by domain sample size, the two models' root mean squared errors were 0.446 versus 0.402, and values of R^2 were 0.783 versus 0.825. Based on this we decided to fit separate models for each item.

The variance functions (8–10) were fitted to each item except the yes/no items, which follow the binomial variance function in the equal-probability sampling case. The iterative procedure described in section 3.4 converged almost precisely in exactly two iterations. This is because the weights for the observations change only with the estimate of τ^2, and so very little change in the weights occurs after the first iteration.

Table 3 presents the average sampling variation, average model error variation, and R^2, for each model averaged over items of each response scale. Sampling variation, computed using (19), does not depend on the model.

Table 3
Goodness-of-fit Statistics for Variance Functions

Rating Scale	0–10		1–4		1–3	
Sampling Variation	0.1460		0.3511		3.1703	
	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model V1	0.020	0.741	0.066	0.824	0.069	0.916
Model V2	0.043	0.710	0.036	0.835	0.000	0.940
Model V3	0.016	0.750	0.024	0.847	0.000	0.947
	Prob(ModErr < Sampling Variation)					
Model V1	0.968		0.916		0.996	
Model V2	0.858		0.967		0.996	
Model V3	0.981		0.983		0.996	

ModErr is the variance component for lack of fit, R^2 is as defined in section 3.4, Prob(ModErr < Sampling Variation) is the proportion of domains for which model error is smaller than sampling variation. All ratings are rescaled to a 0–1 scale, and model errors are multiplied by 10^4.

For items with few categories (more closely resembling the binomial), the quadratic component of the variance function tends to dominate the linear component, making models V2 and V3 fit better than V1. Because V2 imposes a constraint at a point far outside the range of the domain means, it does not fit the data as well when there are more categories and the data are consequently further from binomial. The 0–10 items are less dispersed than the 1–4 and 1–3 ratings, enabling the linear model to fit better. The R^2 values for model V3 were close to 0.75 for numerical (0–10) items, 0.85 for the frequency (1–4) items, and 0.95 for the problem (1–3) items.

The lower portion of Table 3 displays for each item the proportion of domains (of those with at least 2 responses to the given item) for which sampling variation is larger than model error variation. For over 90% of domains, model error variation was less than the sampling variation of the direct variance estimate.

Figure 3 illustrates the fit of V3 for two each of the 0–10, 1–4, and 1–3 items. Illustrations for the remaining items are similar, but are not provided due to space limitations. The fitted curves are constrained to 0 at the maximum ratings. To assess the impact this constraint has on the fitted

variance function, we also fit an unrestricted (three parameter) quadratic variance function; these attained values very close to 0 at the maximum rating, and closely approximated the fitted curve from the constrained models, further supporting V3.

Average parameter estimates and their standard deviations over items of the same type are shown in Table 4. The parameters differed substantially across items, supporting the decision to estimate separate regression coefficients. In most cases the coefficients for both the $p_{h,i}$ and $p_{h,i}(1-p_{h,i})$ terms in V3 were significant, indicating that these are needed for generalized variance modeling. In some cases (particularly with the 0–10 items) the coefficient of the $p_{h,i}(1-p_{h,i})$ term was negative, resulting in an estimated variance function that is convex rather than concave (the shape of the binomial variance function). This can happen when the sample means for the ratings are concentrated on a small proportion of the response scale, over which the linear term explains much of the variation in the data. As mentioned earlier, adding higher-order polynomial or logarithmic functions of $p_{h,i}$ did not significantly improve model fit.

Table 4
Average Variance Function Parameter Estimates for Each Type of Item and Standard Deviations Across Items (in Parentheses)

Model	Item Type					
	0–10		1–4		1–3	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
V1	0.236	–	0.354	–	0.569	–
	(0.016)	–	(0.039)	–	(0.068)	–
V2	–	0.271	–	0.421	–	0.711
	–	(0.020)	–	(0.034)	–	(0.069)
V3	0.334	–0.114	0.151	0.241	0.239	0.420
	(0.143)	(0.155)	(0.104)	(0.132)	(0.112)	(0.110)

See Table 1 for a description of the 0–10, 1–4, and 1–3 items.

4.3 Correlation Functions

Models are ordered from simplest (C1, the constant model) to most complex (C5, containing all linear and quadratic terms). As for the variance models, statistical tests found highly significant item interaction effects, implying that separate models should be fit for each pair. We did not expect all pairs of items to have similar correlations, since by intention the items are divided into internally consistent groups, each of which measures a distinct aspect of patient experiences such as interactions with doctor or dealings with customer service agents (Hays, Shaul, Williams, Lubalin, Harris-Kojetin, Sweeny and Cleary 1999).

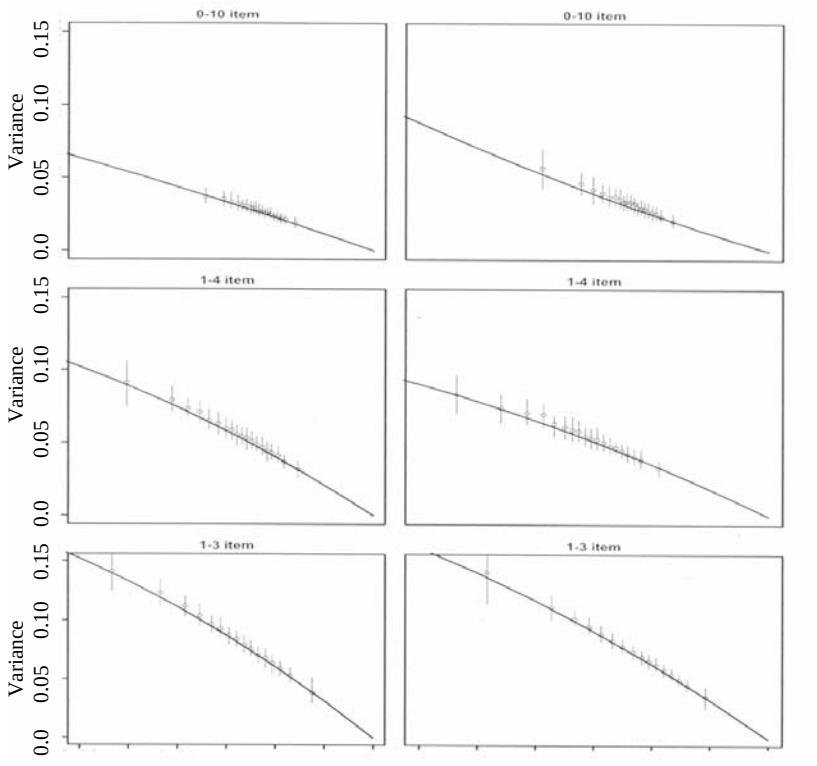


Figure 3. Quadratic Variance Function (V3) of Two Items for each Rating Type. Each point is the average of 60 domains. Vertical lines join the 10th and 90th percentiles of the distribution of the variances. For this and following displays the direction of the transformed horizontal axis has been reversed to agree with that of the original variables.

The fits of the correlation models for pairs of items of the same type are summarized in Table 5. Over the range of models considered, the biggest improvements in model performance (as measured by R^2) occur between model C1 and model C2, and between model C3 and model C4. For example, the average R^2 for the numerical ratings in models C3–C5 are 0.0391, 0.1494, and 0.1508 respectively, and the average R^2 for the 1–4 ratings over C1–C3 are 0, 0.0700, and 0.0789 respectively. This suggests that C2 and C4 are the best models for different pairs of items, a claim that is supported by the hypothesis tests on the significance of the incremental improvements in model fit.

Sampling variation was highest for the 1–3 ratings, at least in part because high rates of non-response due to skipped responses diminished the sample sizes. Model error and R^2 of correlation models for items of different types were similar to those for models for items having the same type.

The R^2 values for the correlation models were between 0.029 and 0.15 for all pairs of items. Although there was no evidence to suggest that C4 was an inappropriate model for the correlations, these results indicate that substantial variation in the correlations is not explained by the item means.

The sampling variances of the direct estimates were often less than the corresponding model error variances (lower part of Tables 5 and 6 especially for the 0–10 items. Under C4, model error variances were smaller for only 13% of domains for the 0–10 ratings, 45% of domains for the 1–4 ratings, and approximately 81% of domains for the 1–3 and 1–2 ratings.

Figure 4 presents the observed correlations and fitted function C4 for an illustrative pair of items from each of the 10 combinations of item types, representing the 595 distinct pairs of items. To illustrate the fitted correlation models, we adjust the observed and fitted correlations to the mean of one item and plot the resulting values in two-dimensional space. This process is repeated for the other item, yielding two plots for each correlation.

Figure 4 illustrates the generally weak relationship of the correlation to the means of the items seen in Tables 5 and 6. Analysis of Tables 5 and 6 reveals that the relationship between the correlation and the mean outcome is weaker for items with fewer categories and with correlations of items of different types. In particular, the 0 –10 numerical ratings are the only group for which there is a clear correlation-mean relationship.

Table 5
Model Fitting Diagnostics for Correlation Functions for Items of the Same Type, Averaged over Pairs of Items of the Same Type

Rating Type	0 – 10		1 – 4		1 – 3		1 – 2	
Sampling Variation	0.0124		0.0178		0.1482		0.0325	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model C1	0.060	0.000	0.028	0.000	0.112	0.000	0.018	0.000
Model C2	0.060	0.013	0.025	0.070	0.103	0.048	0.017	0.014
Model C3	0.057	0.039	0.024	0.079	0.102	0.054	0.017	0.018
Model C4	0.047	0.150	0.023	0.100	0.100	0.068	0.016	0.029
Model C5	0.044	0.151	0.023	0.105	0.096	0.080	0.015	0.034
	Prob(ModErr < Sampling Variation)							
Model C1	0.033		0.339		0.461		0.788	
Model C2	0.033		0.400		0.498		0.795	
Model C3	0.034		0.411		0.502		0.796	
Model C4	0.038		0.435		0.516		0.799	
Model C5	0.065		0.440		0.530		0.802	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

Table 6
Model Fitting Diagnostics for Correlation Functions for C4 by Type of Item.
Averaged over Items of the Same Type

Types	0 – 10		1 – 4		1 – 3		1 – 2	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
0–10	0.047	0.149	0.021	0.104	0.040	0.094	0.013	0.059
1–4			0.023	0.100	0.038	0.076	0.013	0.039
1–3					0.100	0.068	0.028	0.031
1–2							0.016	0.029
Prob(ModErr < Sampling Variation)								
0–10	0.038		0.358		0.523		0.784	
1–4			0.435		0.605		0.790	
1–3					0.516		0.827	
1–2							0.799	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

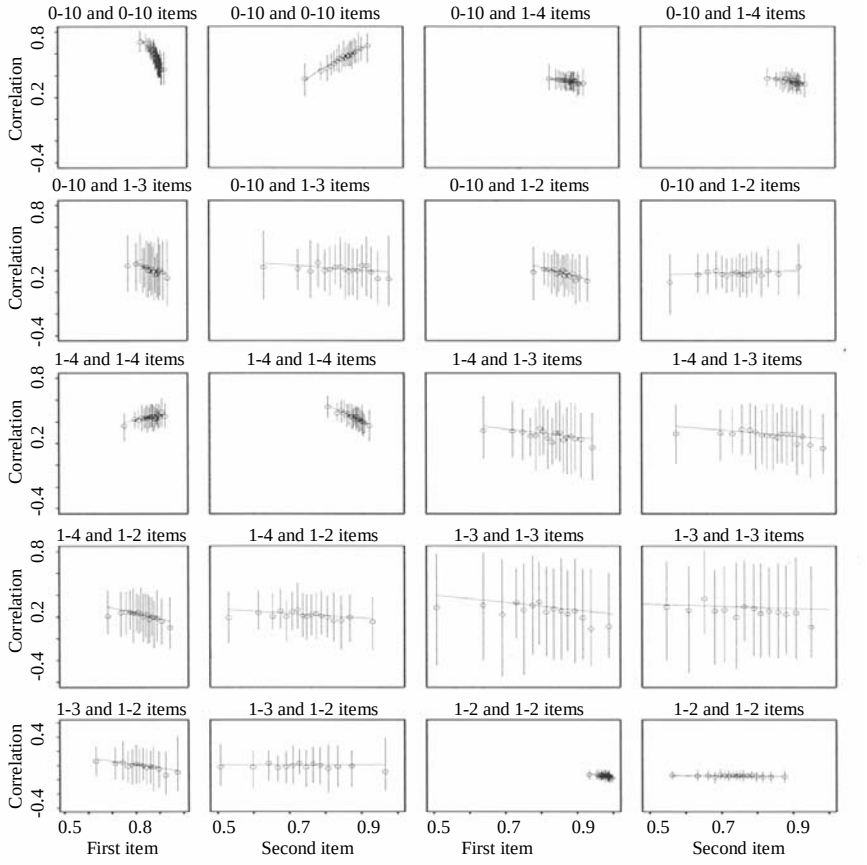


Figure 4. Correlation Functions for One Pair of Items for Each Combination of Rating Types.

Note: The plots for each items involved in the correlation are side by side. Refer to Figure 3 for a description of the contents and axes of the plot.

Although the fitted curves for the correlation functions are nearly flat, the variation in the parameter estimates under model C4 for α_4 are large and were suggestive of instability. The wildly varying parameter estimates are a consequence of collinearity among the predictors in model C4. In many cases the estimated value of α_4 offsets the parameter estimates for the linear predictors, resulting in a fitted curve that is nearly flat.

4.4 Mean Difference Functions

The difference $\hat{D}_{h,ij}$ appeared to depend on both the marginal mean and its square, implying a model analogous to V3 could be appropriate. However, because $\hat{D}_{h,ij}$ typically is small enough that $\hat{D}_{h,ij}\,\hat{D}_{h,ji}$ has minimal impact on (16), we fit a constant model.

4.5 Composite Estimator

Table 7 presents the quantiles of the distribution of weights $\sigma_h^2/(\tau^2+\sigma_h^2)$ for the model-based estimate, used in the composite estimator of section 3.5, averaged over items (or pairs of items) of the same type. The proportion of domains for which the standard error of the model-based predictions was smaller than that of the direct estimates is also presented. As noted previously, the model-based predictions have more weight in the composite variance estimates than in the composite correlation estimates. The average (across items or pairs) median of the weights of the model-based estimator ranged from 0.892 to 1.000 for variances, 0.256 to 0.709 for correlations of items of the same type, and from 0.468 to 0.738 for correlations of items of different types. Also, for both variances and correlations, the weight of the model-based predictions was larger for items with fewer response categories. For example, the model-based estimator had median weights of 0.256, 0.468, 0.540, and 0.647 on the composite estimates of correlations when the numerical 0–10 ratings were paired with the 0–10, 1–4, 1–3, and 1–2 ratings, respectively. However, even for pairs of 0–10 numerical ratings, for which sampling error of the direct estimator exceeded the model error in only 3.81% of domains, these results indicate that the median weight of the model-based estimator was 0.256, a nontrivial amount.

Table 7
Distribution of Weights for the Model-Based Component of the Composite Estimator, Averaged Over Items of Same Type

Model	Item Type		Prob(ModErr <	Quantiles		
	1	2	Sampling Variation)	10%	Median	90%
Variance	0 – 10	–	0.981	0.778	0.892	0.948
	1 – 4	–	0.983	0.948	0.966	0.974
	1 – 3	–	0.996	1.000	1.000	1.000
Correlation	0 – 10	0 – 10	0.038	0.141	0.256	0.335
	0 – 10	1 – 4	0.358	0.301	0.468	0.562
	0 – 10	1 – 3	0.523	0.357	0.540	0.654
	0 – 10	1 – 2	0.784	0.531	0.695	0.767
	1 – 4	1 – 4	0.435	0.324	0.497	0.591
	1 – 4	1 – 3	0.605	0.404	0.587	0.699
	1 – 4	1 – 2	0.853	0.584	0.738	0.805
	1 – 3	1 – 3	0.516	0.349	0.540	0.675
	1 – 3	1 – 2	0.827	0.584	0.737	0.817
	1 – 2	1 – 2	0.799	0.541	0.709	0.780

The distribution of weights is summarized by the 10^{th}, 50^{th}, and 90^{th} percentiles. See Table 3 for definition of ModErr.

4.6 Joint Predictions

Because we modeled the correlations independently for each item, our fitted correlation matrices do not necessarily satisfy the constraint of positive definiteness, which can be important for multivariate inference. In additional work, we have determined that as long as the multivariate analysis is restricted to items of the same type, the fitted correlations from the C2 and C4 models yield positive definite estimates of correlation matrices for almost all domains. However, for analyses including items of different types (*e.g.*, the 0–10 numerical items, and the 1–2 yes/no items), predictions based on C4 predict correlation matrices that are indefinite for many domains, while predictions based on C2 are more stable and almost always yield positive definite predictions. This suggests that while C4 may be slightly superior in terms of univariate model fit, C2 may be more appropriate for multivariate inference.

One way of overcoming the problem of indefinite predicted correlation matrices is to use a weighted average of the predicted correlation matrix for a domain and the estimated average correlation matrix (EACM) across domains. The EACM may be constructed by weighting the direct estimates (each of which is at least positive semi-definite) by the total sample size for each domain. Then any indefinite predicted correlation matrices are replaced with the weighted average of the predicted correlation matrix and the EACM, where the weight used for each domain is increased until a positive definite matrix results. Like an empirical Bayes estimator, this process stabilizes estimates by effectively shrinking the model coefficients toward those of a simpler (constant) model.

When analyzing all 35 CAHPS items simultaneously the EACM had an average weight across domains of 0.65 with model C4, whereas with model C2 the average weight was only 0.01 since the predicted correlations under C2 were usually positive definite. In analyzing only the 0–10, 1–4, and 1–3 items the EACM had average weights of 0.28 and 0.00 with C4 and C2 respectively, while in analyzing just the 0–10 and 1–4 items the corresponding average weights were 0.06 and 0.00. When analyzing the different types of items separately, the average weight of the EACM with C4 was 0.00 for the 0–10 and 1–4 items, 0.01 for the 1–3 items, and 0.17 for the 1–2 items. The EACM is thus not needed when analyzing the 0–10 and 1–4 items because the predicted correlation matrices were positive definite for every domain.

5. Conclusion

We have presented methodology for estimating variance and covariance functions for domain means of ordinal survey items. Our methodology can also be applied to survey items measured on continuous scales. We introduced a decomposition of the model error that allows the variation due to sampling to be separated from that due to model fit. The decomposition also helps to avoid over-fitting because it estimates the proportion of variation in the data that can be modeled and thus when the current predictors suffice.

The procedure for fitting the variance and correlation models is the same regardless of whether or not the data contain skip patterns. The analytic derivation in section 3.3 shows that if skip patterns are present, mean differences of items by response status of other items are required in order to compute the sampling covariance estimates. However, we argued that these quantities are likely to have minimal impact on the results and that therefore a constant model

could be used, which was supported by our empirical findings.

A quadratic variance function constrained to 0 at the maximum rating, and a model for transformed correlations involving the product but not the squares of the means, best predicted the direct estimates in our applied example. The modeled variance estimates generally had much smaller standard errors than the direct estimates; the same was, however, not true of the correlation estimates. It is interesting and reassuring that our quadratic variance function can be expressed as the widely-used relative variance model of Wolter (1985).

For our ordinal data, the estimates of the domain mean ratings contain minimal information about the correlation between the ratings. Hence, the mean-covariance relationship is principally an artifact of the mean-variance relationship. However, for items with many response categories, the association between correlations and mean outcomes for items of the same type was stronger most notably for pairs of 0–10 items. With the exception of the 0–10 and possibly the 1–4 ratings, the correlations might as well be modeled as constants, which also makes it easier to guarantee positive definiteness of the predicted correlation matrix. However, it is important that the parameters of the correlation model be allowed to vary across pairs of items.

A composite estimator that weights the direct and model-based estimators proportional to their precisions has smaller variance than either estimator alone, especially when the components have close to equal weight. The model-based estimator had the greatest influence on estimates for small domains, for which little information is available. The model-based estimator had the greatest influence on estimates for variances, followed by correlations of items of the same type, and lastly correlations of items of different types. Both model-based and composite estimators can be benchmarked (ratio adjusted) to agree on the average across domains with direct estimates, although this proved to be unnecessary in our example.

GVCFs find several applications in our continuing research. We are developing quasi likelihood-based methods for estimating covariance matrices for the domain means of ordinal survey items, representing the second-level (structural) covariance in a hierarchical model (O'Malley and Zaslavsky 2004). GVCF models are needed to provide estimates of sampling variances and covariances and to modify those estimates as the means are re-estimated during the fitting procedure. If the sampling variability of the GVCF estimates is minimal because the number of domains is large, the GVCF predicted variances and covariances can be treated as known. However, if the sampling error of the GVCF-based estimates is large a model that allows these errors to propagate through the analysis should be used. In related work, Fay and Train (1997) used a binomial model with a design effect for each domain in empirical Bayes estimation of binomial rates. Our research extends this approach to multivariate estimation and more general response formats.

Another application of GVCFs is the computation of variance estimates for linear combinations of item means, facilitating variance estimation for composite scores, like those used in CAHPS reporting. The methods described in section 2 are applicable to variance estimation for any functions of totals, including functions of means, other ratios, or regression coefficients.

There are several ways of extending the GVCF methodology. In addition to summary measures of outcomes, generalized variance and covariance functions (GVCFs) may also depend on other independent variables, in particular those that would better predict correlations. We considered variables summarizing response patterns, such as the proportion of respondents in a domain, but these did not improve the model. GVCFs could also be extended to multi-stage sampling.

Acknowledgements

This work was supported by the U.S. Agency for Healthcare Research and Quality through the Consumer Assessments of Health Plans Study (grant U18 HS09205-06) and by the U.S. Centers for Medicare and Medicaid Services (contract 500-95-007). We thank Paul D. Cleary for his ongoing support of this work, Matt Cioffi for data management, and Elizabeth Goldstein and Amy Heller of the Centers for Medicare and Medicaid Services (CMS), and the other members of the CAHPS-MMC survey implementation team.

References

Cho, M.J., Eltinge, J.L., Gershunskaya, J. and Huff, L.L. (2002). Evaluation of generalized variance function estimators for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM]. Alexandra, VA: American Statistical Association, 534-539.

Eltinge, J. (2002). Use of generalized variance functions in multivariate analysis. In *Proceedings of the Section on Survey Research Methods*, Alexandra, VA: American Statistical Association, 904-913.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fay, R.E., and Train, G.F. (1997). Small domain methodology for estimating income and poverty characteristics for states in 1993. In *Proceedings of the Social Statistics Section*, Alexandra, VA: American Statistical Association, 183-188.

Freund, J.E., and Walpole, R.E. (1987). *Mathematical Statistics*. New Jersey: Prentice-Hall, Inc., 4th Edn.

Gabler, S., Haeder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.

Goldstein, E., Cleary, P.D., Langwell, K.M. Zaslavsky, A.M. and Heller, A. (2001). Medicare Managed Care CAHPS: A tool for performance improvement. *Health Care Financing Review*, 22, 101-107.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Hays, R.D., Shaul, J.A., Williams, V.S.L., Lubalin, J.S., Harris-Kojetin, L.D., Sweeny, S.F. and Cleary, P.D. (1999). Psychometric properies of the CAHPS 1.0 survey measures. *Medical Care*, 37 (Supplement), 22-31.

Huff, L.L., Eltinge, J.L. and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM], Alexandra, VA: American Statistical Association, 1519-1524.

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

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

O'Malley, A.J. and Zaslavsky, A.M. (2004). Implementation of cluster-level covariance analysis for survey data with structured nonresponse. In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 1907-1914.

Otto, M.C., and Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

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

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*, 26, 137-138.

Valliant, R. (1992a). Longitudinal smoothing of price index variances. In *Statistics Canada Symposium*. Ottawa: Statistics Canada. 113-120.

Valliant, R. (1992b). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8, 433-444.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

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

Woodruff, S. (1992). Variance estimation for estimates of employment change in the Current Employment Statistics Survey. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA: American Statistical Association, 626-631.

Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E. and Cleary, P.D. (2000). Dimensions of consumer-assessed quality of Medicare managed-care health plans. *Medical Care*, 38, 162-174.

Zaslavsky, A.M., and Cleary, P.D. (2002). Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey. *Medical Care*, 40, 951-964.

.

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 183-195
Statistics Canada, Catalogue No. 12-001-XPB

Spatio-Temporal Models in Small Area Estimation

Bharat Bhushan Singh, Girja Kant Shukla and Debasis Kundu [1]

Abstract

A spatial regression model in a general mixed effects model framework has been proposed for the small area estimation problem. A common autocorrelation parameter across the small areas has resulted in the improvement of the small area estimates. It has been found to be very useful in the cases where there is little improvement in the small area estimates due to the exogenous variables. A second order approximation to the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP) has also been worked out. Using the Kalman filtering approach, a spatial temporal model has been proposed. In this case also, a second order approximation to the MSE of the EBLUP has been obtained. As a case study, the time series monthly per capita consumption expenditure (MPCE) data from the National Sample Survey Organisation (NSSO) of the Ministry of Statistics and Programme Implementation, Government of India, have been used for the validation of the models.

Key Words: Mixed effects linear model; Spatial autocorrelation; Weight matrix; Best linear unbiased predictor; Empirical best linear unbiased predictor; Kalman filtering; NSSO rounds.

1. Introduction

Local level planning requires reliable data at the appropriate level. The complete enumeration or large sample surveys with adequate sample size is expensive and time consuming. The censuses are usually carried out once in a decade, while the sample surveys are often planned to provide estimates at much higher level. One such large sample survey is socio-economic survey of National Sample Survey Organisation (NSSO). Here the direct survey estimates are available at small area (district) level as most of the districts are stratum in the sampling procedure adopted by the NSSO. However, the estimates are exceedingly unreliable due to unacceptably large standard errors. This requires strengthening of such estimates with the use of information from similar small areas or with the help of some relatable exogenous variables, easily available and related to the variable under study.

Various model based approaches have been suggested to improve the direct estimators. The model-based approach facilitates its validation through the sample data. The simple area specific model suggested is two stage model of Fay and Herriot (1979).

$$y_i = \theta_i + \varepsilon_i,\ E(\varepsilon_i \mid \theta_i) = 0,\ \mathrm{Var}(\varepsilon_i \mid \theta_i) = \sigma_i^2, \qquad (1.1)$$

$$\theta_i = X_i^T\beta + v_i z_i,\ E(v_i) = 0,\ \mathrm{Var}(v_i) = \sigma_v^2,\ i = 1, 2, \ldots, m. \qquad (1.2)$$

Here y_i's are direct survey estimators of θ_i's, the characteristic under study. θ_i's may be population small area means. $X_i = (X_{i1}, \ldots, X_{ip})^T$'s are exogenous variables which are available and assumed to be closely related to θ_i's and z_i's are known positive constants. $\beta(p \times 1)$ is the vector of regression parameters.

The first equation (1.1) is the design model while the second (1.2) is the linking model. The ε_i's are sampling errors. Estimators y_i's are design unbiased and the sampling variances σ_i^2's are known. Further the ε_i's and v_i's are identically and independently distributed random variables. Normality of the random errors and random effects are often assumed. For this model, best linear unbiased predictor (BLUP) on the line of the best linear unbiased estimator (BLUE) has been suggested. The estimate is design consistent and model unbiased (Ghosh and Rao 1994). It is typically the weighted average of the direct survey estimator y_i and the regression synthetic estimator $X_i^T\beta$. The BLUP estimator depends on variance component σ_v^2 which is unknown in pratical applications. Various methods of estimating variance components in general mixed effects linear model are available (Cressie 1992). By replacing σ_v^2 with an asymptotically consistent estimator $\hat{\sigma}_v^2$, an empirical best linear unbiased predictor (EBLUP) has also been obtained.

The main problem associated with the data in the Indian context is the non-availability of administrative or civic registration data at small area level. Often, it is difficult to find out the exogenous variables closely related (multiple correlation coefficient $R^2 > 0.5$) to the variable under study.

In the present paper, the exploitation of spatial autocorrelation amongst the small area units in the form of spatial model, has been considered for improving the small area estimators. Besides this, for the time series data, a spatial temporal model on the line of Kalman filtering has been utilised to further improve the estimators. Time series data on monthly per capital consumption expenditure

1. Bharat Bhushan Singh, Girja Kant Shukla and Debasis Kundu, Department of Mathematics, I.I.T. Kanpur-208016. E-mail: drbbsingh@hotmail.com.

(MPCE) as estimated from a large sample survey carried out by the National Sample Survey Organisation (NSSO) has been studied. In the present paper, we propose suitable models in the framework of mixed effects linear model to provide better estimators of the MPCE at small area level.

Rest of the paper has been organized as follows. In Section 2, we consider a Spatial Model on the line of general mixed effects linear model with the introduction of spatial autocorrelation among the small area units. The BLUP and EBLUP of the mixed effects have been presented. A second order approximation to the MSE of the EBLUP and to the estimator of the MSE has also been obtained. Section 3 deals with the time series extension of Spatial Model in form of Spatial Temporal Model, using the Kalman filtering approach. The BLUP and the EBLUP of the mixed effects along with a second order approximation to the MSE of the EBLUP and to the estimator of the MSE have been discussed. Section 4 presents and analyses estimates of the MPCE from a large sample survey carried out periodically in India. The conclusions of the data analysis are reported in Section 5. All the proofs have been provided in the Appendix.

2. Spatial Model

The small area characteristics usually have the spatial dependence in terms of neighbourhood similarities. Cressie (1990) used conditional spatial dependence among random effects, in the context of adjustment for census undercounts. Here, we use simultaneous spatial dependence (Cliff and Ord 1981) among the random effects which has certain advantage over conditional dependence (Ripley 1981). We have thus tried to explain a portion of the random error unaccounted for and left over by explanatory variables which makes it possible to improve the direct survey estimators. The proposed model is a three stage area specific model (Ghosh and Rao 1994).

$$y = \theta + \varepsilon, \qquad \varepsilon \sim N_m(0, R), \tag{2.1}$$

$$\theta = X\beta + u, \tag{2.2}$$

$$u = \rho W u + v, \qquad v \sim N_m(0, \sigma_v^2 I), \tag{2.3}$$

where θ is a m–component vector (corresponding to number of small areas) for the characteristic under study and y is its direct surevy estimator obtained through small sample data. In the above model, the first equation (2.1) shows the design (sampling) model, the second equation (2.2) shows regression model and the third one (2.3) shows spatial model on the residuals, the later two are linked in the first equation. The above model can be expressed as

$$y = X\beta + Zv + \varepsilon, \qquad Z = (I - \rho W)^{-1}, \tag{2.4}$$

where $X(m \times p)$ is the design matrix of full column rank p, $\beta(p \times 1)$ is a column vector of regression parameters and $Z(m \times m)$ represents the coefficients of random effects v. $W(m \times m)$ is a known spatial weight matrix which shows the amount of interaction between any pair of small areas. The elements of $W \equiv [W_{ij}]$ with $W_{ii} = 0 \quad \forall i$ may depend on the distance between the centers of small areas or on the length of common boundary between them. As a simple alternative, it may have binary values $W_{ij} = 1$ (unscaled) if j^{th} area is physically contiguous to i^{th} area and $W_{ij} = 0$, otherwise. The matrix has been standardised so as to satisfy $\sum_{j=1}^{m} W_{ij} = 1$ for $i = 1, 2, \ldots, m$. The constant ρ is a measure of the overall level of spatial autocorrelation and its magnitude reflects the suitability of W for given y and X. Further v and ε are assumed to be independent of each other. R is a diagonal matrix of order m which may be expressed as $R = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ where σ_i^2's are known sampling variances corresponding to the i^{th} area. The parameter vector $\psi = [\rho, \sigma_v^2]^T$ has two elements.

In this model the strength is borrowed from the similar small areas through two common parameters *viz.* regression parameter β and autocorrelation parameter ρ. Note that the present model is a more general model and the model of Fay and Herriot (1979) can be obtained from this by taking $\rho = 0$.

By adopting the mixed effects linear model approach (Henderson 1975), the best linear unbiased predictor (BLUP) of $\theta = X\beta + Zv$ and the mean squared error (MSE) of the BLUP may be obtained as

$$\begin{aligned}\hat{\theta}(\psi) &= X\hat{\beta}(\psi) + \Lambda(\psi)[y - X\hat{\beta}(\psi)] \\ &= \sigma_v^2 A^{-1}(\psi)\Sigma^{-1}(\psi)y + R\Sigma^{-1}(\psi)X\hat{\beta}(\psi),\end{aligned} \tag{2.5}$$

$$\text{MSE}[\hat{\theta}(\psi)] = E[(\hat{\theta}(\psi) - \theta)(\hat{\theta}(\psi) - \theta)^T] = g_1(\psi) + g_2(\psi), \tag{2.6}$$

$$g_1(\psi) = \Lambda(\psi)R = R - R\Sigma^{-1}(\psi)R, \tag{2.7}$$

$$g_2(\psi) = R\Sigma^{-1}(\psi)X(X^T\Sigma^{-1}(\psi)X)^{-1}X^T\Sigma^{-1}(\psi)R, \tag{2.8}$$

$$\hat{\beta}(\psi) = [X^T\Sigma^{-1}(\psi)X]^{-1}X^T\Sigma^{-1}(\psi)y,$$

$$\Sigma(\psi) = \sigma_v^2 A^{-1}(\psi) + R,$$

$$\Lambda(\psi) = \sigma_v^2 A^{-1}(\psi)\Sigma^{-1}(\psi), \; A(\psi) = (I - \rho W)^T(I - \rho W).$$

Here $\hat{\beta}, \Sigma$ and A, all are the functions of ψ and usually have been expressed as $\hat{\beta}(\psi), \Sigma(\psi)$ and $A(\psi)$ respectively. However, sometimes due to brevity, the suffix ψ has been omitted. The first term, $g_1(\psi)$ in the expression for the MSE, shows the variability of $\hat{\theta}$ when all the parameters are known and is of order $O(1)$. The second term, $g_2(\psi)$, due to estimating the fixed effects β, is of order $O(m^{-1})$ for large m. Further, with $\rho = 0$, the above

model reduces to the standard mixed effects linear regression model while for $X\beta = \mu$, we obtain a purely spatial scheme with only intercept term.

In practice parameter ψ is unknown and is estimated from the data. The maximum likelihood estimator (MLE) of the parameter, ψ is obtained by maximizing the following log likelihood function of ψ

$$l = \text{const} - \frac{1}{2}\log[|\Sigma(\psi)|] - \frac{1}{2}[y - X\hat{\beta}(\psi)]^T \Sigma^{-1}(\psi)[y - X\hat{\beta}(\psi)] \quad (2.9)$$

with respect to the parameter ψ. The empirical best linear unbiased predictor (EBLUP), $\hat{\theta}(\psi)$ and the naive estimator of the MSE are obtained from the equations (2.5) and (2.6) respectively, by replacing the parameter vector ψ by its estimator $\hat{\psi}$.

$$\hat{\theta}(\hat{\psi}) = \hat{\sigma}_v^2 A^{-1}(\hat{\psi})\Sigma^{-1}(\hat{\psi})y + R\Sigma^{-1}(\hat{\psi})X\hat{\beta}(\hat{\psi}), \quad (2.10)$$

$$\text{MSE}[\hat{\theta}(\hat{\psi})] = g_1(\hat{\psi}) + g_2(\hat{\psi}), \quad (2.11)$$

where $\Sigma(\hat{\psi}) = \hat{\sigma}_v^2 A^{-1}(\hat{\psi}) + R$

and $A(\hat{\psi}) = (I - \hat{\rho}W)^T(I - \hat{\rho}W)$.

This expression for the MSE of the EBLUP severely underestimates the true MSE as the variability due to the estimation of the parameters through the data has been ignored. We obtain a second order approximation to the $\text{MSE}[\hat{\theta}(\hat{\psi})]$ in case $\hat{\psi}$ is the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (REMLE) of ψ, with the assumption of large m and by neglecting all the terms of the order $o(m^{-1})$, under the following regularity conditions. The approximation has been worked out along the lines of Prasad and Rao (1990) and Datta and Lahiri (2000) which are heuristic in nature.

Regularity Conditions 1

(a) The elements of X are uniformly bounded such that $X^T\Sigma^{-1}(\psi)X = [O(m)]_{p\times p}$, where $\Sigma(\psi) = [\sigma_v^2 A^{-1}(\psi) + R]$;

(b) m is finite;

(c) $\Lambda(\psi)X = [O(1)]_{m\times p}$, $(\partial[\Lambda(\psi)X])/(\partial\psi_d) = [O(1)]_{m\times p}$, $(\partial^2[\Lambda(\psi)])/(\partial\psi_d\partial\psi_e) = [O(1)]_{m\times m}$ for $d, e = 1, 2$;

(d) $\hat{\psi}$ is the estimator of ψ which satisfies $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y)\ \forall h \in R^p$ and $\forall y$.

These regularity conditions are satisfied in this case. The special standardised form of the weight matrix W satisfies the condition (c) for $|\rho| < 1$ as it has only a finite number of nonzero elements and its row sum is equal to 1. It may be mentioned here that the matrix $\sigma_v^2 A^{-1}\Sigma^{-1}$ has finite number of nonzero elements and the order of $W, (I - \rho W)$, $W(I - \rho W), \Sigma, \Sigma^{-1}$ or any sum or product combination of these and their derivatives mentioned in condition (c) do not increase. The MLE and the REMLE, in addition satisfy the condition (d). A second order approximation to the MSE of the EBLUP has been shown in Theorem A.1 of the Appendix as

$$\text{MSE}[\hat{\theta}(\hat{\psi})] = E[(\hat{\theta}(\hat{\psi}) - \theta)(\hat{\theta}(\hat{\psi}) - \theta)^T] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (2.12)$$

Here the third term $g_3(\psi)$ comes from estimating the unknown parameter vector from the sample data and it is of the same order $O(m^{-1})$ as that of $g_2(\psi)$. Further $g_3(\psi)$ may be expressed as

$$g_3(\psi) = L^T(\psi)[I_\psi^{-1}(\psi) \otimes \Sigma(\psi)]L(\psi), \quad (2.13)$$

where

$$L(\psi) = \underset{1\le d\le 2}{\text{Col}}[L_d(\psi)] = [L_\rho(\psi), L_{\sigma_v^2}(\psi)]^T,$$

$$L_d(\psi) = \frac{\partial\Lambda(\psi)}{\partial\psi_d},\ d = 1, 2. \quad I_\psi(\psi) = E[-\frac{\partial^2 l}{\partial\psi\partial\psi^T}]$$

is the information matrix and $\otimes$ represents Kronecker product. Further $g_3(\psi)$ may also be written as

$$g_3(\psi) = \sum_{d=1}^{2}\sum_{e=1}^{2} L_d(\psi)\Sigma(\psi)L_e^T(\psi)I_{de}^{-1}(\psi) \quad (2.14)$$

where $I_\psi^{-1}(\psi) \equiv (I_{de}^{-1}(\psi))$.

It is common practice to estimate the MSE of the EBLUP by replacing the unknown parameters including components of the variance by their respective estimators. This procedure can lead to severe underestimation of the true MSE (Prasad and Rao 1990, Singh, Stukel and Pfeffermann 1998). We obtain the estimator of the MSE of the EBLUP in Theorem A.2 of the Appendix for large m neglecting all terms of order $o(m^{-1})$. As a result we have the expressions

$$E[g_1(\hat{\psi}) + g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] = g_1(\psi) + o(m^{-1}), \quad (2.15)$$

$$E[g_2(\hat{\psi})] = g_2(\psi) + o(m^{-1})$$

and $E[g_3(\hat{\psi})] = g_3(\psi) + o(m^{-1})$, (2.16)

and finally the estimator of the MSE of $\hat{\theta}(\hat{\psi})$ as

$$\text{mse}[\hat{\theta}(\hat{\psi})] = [g_1(\hat{\psi}) + g_2(\hat{\psi}) + 2g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] + o(m^{-1}), \quad (2.17)$$

where $E[\text{mse}(\hat{\theta}(\hat{\psi}))] = \text{MSE}[\hat{\theta}(\hat{\psi})] + o(m^{-1})$.

Obviously the additional terms, $g_3(\hat{\psi}), g_4(\hat{\psi})$ and $g_5(\hat{\psi})$ are the contributions, due to estimation of unknown parameter vector ψ by $\hat{\psi}$. The expressions for $g_4(\psi)$ and $g_5(\psi)$ up to order $o(m^{-1})$ are given by

$$g_4(\psi) = [b_{\hat{\psi}}^T(\psi) \otimes I_m] \frac{\partial g_1(\psi)}{\partial \psi},$$

$$b_{\hat{\psi}}(\psi) = \frac{1}{2} I_{\hat{\psi}}^{-1}(\psi) \underset{1 \le d \le 2}{\text{Col}} \left[\text{Trace} \left[I_\beta^{-1}(\psi) \frac{\partial I_\beta(\psi)}{\partial \psi_d} \right] \right], \quad (2.18)$$

$$g_5(\psi) = \frac{1}{2} \underset{m}{\text{Trace}} \begin{bmatrix} [I_2 \otimes (R\Sigma^{-1}(\psi))] \\ \frac{\partial^2 \Sigma(\psi)}{\partial \psi \partial \psi^T} [I_\psi^{-1}(\psi) \otimes (\Sigma^{-1}(\psi) R)] \end{bmatrix}. \quad (2.19)$$

Here $b_{\hat{\psi}}(\psi)$ is the bias of $\hat{\psi}$ *i.e.*, $E(\hat{\psi}) - \psi$ up to order $o(m^{-1})$ and $(\partial g_1(\psi))/(\partial \psi)$ is a partitioned matrix $[(\partial g_1(\psi))/(\partial \rho), (\partial g_1(\psi))/(\partial \sigma_v^2)]^T$ of order $(2m \times m)$ having 2 matrices of order $m \times m$ in a column. In the same way $(\partial^2 \Sigma(\psi))/(\partial \psi \partial \psi^T)$ is a partitioned matrix of order $(2m \times 2m)$ having 2 partitions, row and column wise with $(\partial^2 \Sigma(\psi))/(\partial \psi_d \partial \psi_e)$ being a general sub matrix of order $m \times m$ therein. $\text{Trace}(B) = \sum_{d=1}^2 B_{dd}$, where B is a square partitioned matrix with square sub matrices of similar order. In addition $g_4(\psi)$ and $g_5(\psi)$ may also be written as

$$g_4(\psi) = \frac{1}{2} \sum_{d=1}^2 \sum_{e=1}^2 I_{de}^{-1}(\psi) \text{Trace} \left[I_\beta^{-1}(\psi) \frac{\partial I_\beta(\psi)}{\partial \psi_d} \right] \frac{\partial g_1(\psi)}{\partial \psi_e}, \quad (2.20)$$

$$g_5(\psi) = \frac{1}{2} \sum_{d=1}^2 \sum_{e=1}^2 \left[R\Sigma^{-1}(\psi) \frac{\partial^2 \Sigma(\psi)}{\partial \psi_d \partial \psi_e} \Sigma^{-1}(\psi) R I_{de}^{-1}(\psi) \right]. \quad (2.21)$$

The expression (2.17) gives the matrix of the estimator of the MSE of EBLUP, $\hat{\theta}(\hat{\psi})$ and the MSE of the individual small area estimators may be obtained as the respective diagonal element. In case of simple model without the spatial autocorrelation, similar expressions can be obtained. In this case $g_5(\psi)$, however, becomes zero.

3. Spatial Temporal Model

In this section, State Space Models via Kalman filtering have been used to take the advantage of the time series data along with the common regression parameter and common autocorrelation parameter to strengthen the direct survey estimators at any point of time. This is especially advantageous in the case where the past survey estimates are more reliable. The models used in this category are the following

$$y_t = X_t \beta + Z v_t + \varepsilon_t, \varepsilon_t \overset{\text{ind}}{\sim} N_m(0, R_t), Z = (I - \rho W)^{-1}, \quad (3.1)$$

$$v_t = k v_{t-1} + \eta_t, \eta_t \overset{\text{ind}}{\sim} N_m(0, \sigma_v^2 I) \ \ t = 1, 2, \ldots, T \text{ and}$$
$$\varepsilon_t \text{ and } \eta_t \text{ are independent of each other.} \quad (3.2)$$

Here the parameters have usual meaning as explained in the previous section. Weight matrix $W(m \times m)$ and design matrices $X_t(m \times p)$ are known, $Z(m \times m)$ is a matrix of coefficients of random effects and ρ is an unknown autocorrelation coefficient. R_t is a diagonal matrix of order m which may be expressed as $R_t = \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2, \ldots, \sigma_{mt}^2)$ where σ_{it}^2's are known sampling variances corresponding to the i^{th} small area and t^{th} time point. β is unknown vector of fixed effects and $\psi = [\rho, \sigma_v^2, k]^T$ is a vector of three unknown parameters. These parameters are independent of time *t*. It may be noted that the random effects v_t have been allowed to change in accordance with (3.2) and k is temporal autoregressive parameter. For stationarity $|k| < 1$.

The estimators of fixed and random effects and the MSE of these estimators are obtained in stages, starting with assumption of mixed effects linear model approach at time $t = 1$, and by taking $v_1 \sim N_m(0, \sigma_v^2 I)$ (Sallas and Harville 1994). In the standard form we write the model as

$$y_t = U_t \alpha_t + \varepsilon_t, \alpha_t = T\alpha_{t-1} + \zeta_t, T = \text{diag}[I_p, kI_m], \quad (3.3)$$

$$\zeta_t \sim N_{p+m}(0, Q), \quad Q = \text{diag}[0_p, \sigma_v^2 I_m]$$
$$U_t = [X_t, Z], \alpha_t = [\beta_t, v_t]^T. \quad (3.4)$$

Here I_m and 0_m are the unit and zero matrices of order m and by $\text{diag}[I_p, kI_m]$ we mean the matrix

$$\begin{bmatrix} I_{p \times p} & 0_{p \times m} \\ 0_{m \times p} & kI_{m \times m} \end{bmatrix}.$$

In case β is assumed fixed but dependent on time, there is no change in the model except that $T = \text{diag}[0_p, kI_m]$.

The initial estimates of the effects α_t and their variances (based on $t = 1$) are obtained as

$$\hat{\beta}_1 = (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} y_1, \hat{v}_1 = \sigma_v^2 Z^T H_1^{-1} (y_1 - X_1 \hat{\beta}_1),$$
$$H_1 = R_1 \sigma_v^2 A^{-1}, \quad \Sigma_1 = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\Sigma_{11}(p \times p) = (X_1^T H_1^{-1} X_1)^{-1},$$
$$\Sigma_{12}(p \times m) = \Sigma_{21}^T = -\sigma_v^2 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z$$
$$\text{and } \Sigma_{22}(m \times m) = \sigma_v^2 I_m - \sigma_v^4 Z^T H_1^{-1} Z$$
$$+ \sigma_v^4 Z^T H_1^{-1} X_1 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z.$$

The recurring Kalman filtering equations for updation of the estimators at subsequent stages are

$$\Sigma_{t|t-1} = T\Sigma_{t-1}T^T + Q,\ \hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1},\ H_t = R_t + U_t\Sigma_{t|t-1}U_t^T,$$
$$\hat{\alpha}_t = \hat{\alpha}_{t|t-1} + \Sigma_{t|t-1}U_t^T H_t^{-1}(y_t - U_t\hat{\alpha}_{t|t-1}),$$
$$\Sigma_t = \Sigma_{t|t-1} - \Sigma_{t|t-1}U_t^T H_t^{-1}U_t\Sigma_{t|t-1}$$

where $\hat{\alpha}_{t|t-1}$ are the estimators of the effects α_t given the observations $[y_1, y_2, \ldots, y_{t-1}]$ and the $\Sigma_{t|t-1}$ are the mean squared errors of $\hat{\alpha}_{t|t-1}$. H_t are the conditional variance covariance matrix of y_t given $[y_1, y_2, \ldots, y_{t-1}]$. With the help of the above recurring filtering equations, the best linear unbiased predictor (BLUP) of $\theta_t = X_t\beta + Zv_t$, and the mean squared error (MSE) of the BLUP may be obtained as

$$\begin{aligned}\hat{\theta}_t(\psi) &= U_t(\psi)\hat{\alpha}_t(\psi)\\ &= y_t - R_tH_t^{-1}(\psi)[y_t - U_t(\psi)\hat{\alpha}_{t|t-1}(\psi)]\\ &= U_t(\psi)\hat{\alpha}_{t|t-1}(\psi) + \Lambda_t(\psi)e_t(\psi),\end{aligned} \tag{3.5}$$

$$\mathrm{MSE}[\hat{\theta}_t(\psi)] = g_{12t}(\psi) = U_t(\psi)\Sigma_t(\psi)U_t^T(\psi), \tag{3.6}$$

where $\Lambda_t(\psi) = U_t(\psi)\Sigma_{t|t-1}(\psi)U_t^T(\psi)H_t^{-1}(\psi)$
$= I_m - R_tH_t^{-1}(\psi)$
and $e_t(\psi) = y_t - U_t(\psi)\hat{\alpha}_{t|t-1}(\psi)$.

It may be noted that $g_{12t}(\psi)$ is the spatial counterpart of $g_1(\psi) + g_2(\psi)$. As usual in practice, the parameter vector ψ is unknown and its restricted maximum likelihood estimators (REMLE) can be obtained by maximizing the following log likelihood function, based on the sample data covering all time points

$$\begin{aligned}l = \text{const.} &- \frac{1}{2}\log[|X_1^TH_1^{-1}X_1|] - \frac{1}{2}\sum_{t=1}^{T}\log[|H_t|]\\ &- \frac{1}{2}(y_1 - X_1\hat{\beta}_1)^TH_1^{-1}(y_1 - X_1\hat{\beta}_1)\\ &- \frac{1}{2}\sum_{t=2}^{T}(y_t - U_t\hat{\alpha}_{t|t-1})^TH_t^{-1}(y_t - U_t\hat{\alpha}_{t|t-1})\end{aligned} \tag{3.7}$$

with respect to the parameter ψ. With the help of the above, the estimator, $\hat{\psi}$ is obtained and the EBLUP of θ_t and the naive estimator of the MSE of the EBLUP are given by

$$\hat{\theta}_t(\hat{\psi}) = U_t(\hat{\psi})\hat{\alpha}_t(\hat{\psi}) = U_t(\hat{\psi})\hat{\alpha}_{t|t-1}(\hat{\psi}) + \Lambda_t(\hat{\psi})e_t(\hat{\psi}), \tag{3.8}$$

$$\mathrm{MSE}[\hat{\theta}_t(\hat{\psi})] = g_{12t}(\hat{\psi}) = U_t(\hat{\psi})\Sigma_t(\hat{\psi})U_t^T(\hat{\psi}). \tag{3.9}$$

As explained earlier in section 2, the MSE of the EBLUP underestimates the true MSE as it does not take care of the variability due to replacing parameters by their estimates. A second order approximation to the $\mathrm{MSE}[\hat{\theta}_t(\hat{\psi})]$ for large m and neglecting all the terms of order $o(m^{-1})$, has been obtained in Theorem A.3 of the Appendix, under the following regularity conditions satisfied by our model. These conditions are analogous to the regularity conditions 1.

Regularity Conditions 2

(a) The elements of $X_t, t = 1, 2, \ldots, T$ are uniformally bounded such that $X_t^T\Sigma_t^{-1}(\psi)X_t = [O(m)]_{p\times p}$, where $\Sigma_t(\psi) = [\sigma_v^2A^{-1}(\psi) + R_t]$;

(b) m and T are finite;

(c) $\Lambda_t(\psi)U_t(\psi) = [O(1)]_{m\times p}$, $(\partial[\Lambda_t(\psi)U_t(\psi)])/(\partial\psi_d) = [O(1)]_{m\times p}$, $([\partial^2\Lambda_t(\psi)])/(\partial\psi_d\partial\psi_e) = [O(1)]_{m\times m}$, $t = 1, 2, \ldots, T$ and $d, e = 1, 2, 3$;

(d) $\hat{\psi}$ is the estimator of ψ which satisfies $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y)\ \forall h \in R^p$ and $\forall$ y.

The second order approximation to the MSE of the EBLUP is

$$\begin{aligned}\mathrm{MSE}[\hat{\theta}_t(\hat{\psi})] &= E[(\hat{\theta}_t(\hat{\psi}) - \theta_t)(\hat{\theta}_t(\hat{\psi}) - \theta_t)^T]\\ &= g_{12t}(\psi) + g_{3t}(\psi) + o(m^{-1}).\end{aligned} \tag{3.10}$$

Here $g_{3t}(\psi)$ is the bias due to the estimation of the parameters from the sample data and is of the order $O(m^{-1})$ and it is given by

$$g_{3t}(\psi) = L_t^T(\psi)I_\psi^{-1}(\psi)K_\psi(\psi)H_tI_\psi^{-1}(\psi)L_t(\psi) \tag{3.11}$$

where $K_\psi(\psi) \equiv (K_{de}(\psi))$

and $$K_{de}(\psi) = \frac{1}{2}\sum_{i=1}^{T}\mathrm{Trace}\left[H_i^{-1}\frac{\partial H_i}{\partial\psi_d}H_i^{-1}\frac{\partial H_i}{\partial\psi_e}\right]. \tag{3.12}$$

Further

$$L_t(\psi) = \underset{1\le d\le 3}{\mathrm{Col}}\,[L_{td}(\psi)] \text{ and } L_{td}(\psi) = (\partial\Lambda_t(\psi))/(\partial\psi_d)$$
for $d = 1, 2, 3$.

In a proper form, we may write $g_{3t}(\psi)$ as

$$g_{3t}(\psi) = \sum_{d=1}^{3}\sum_{e=1}^{3}L_{td}(\psi)\left[\sum_{f=1}^{3}\sum_{g=1}^{3}I_{df}^{-1}(\psi) \times\sum_{i=1}^{T}\mathrm{Trace}\left(H_i^{-1}\frac{\partial H_i}{\partial\psi_f}H_i^{-1}\frac{\partial H_i}{\partial\psi_g}\right) \times H_tI_{ge}^{-1}(\psi)\right]L_{te}^T(\psi).$$

The expression for the information matrix involved here, may be given as

$$I_{de}(\psi) = E\left[-\frac{\partial^2 l}{\partial\psi_d \partial\psi_e}\right]$$

$$= \frac{1}{2}\sum_{t=1}^{T} \text{Trace}\left[H_t^{-1}\frac{\partial H_t^{-1}}{\partial\psi_d}H_t^{-1}\frac{\partial H_t}{\partial\psi_e}\right] + \sum_{t=1}^{T}\left[\frac{\partial e_t^T}{\partial\psi_d}H_t^{-1}\frac{\partial e_t}{\partial\psi_e}\right]$$

$$-\frac{1}{2}\text{Trace}\left[\begin{array}{l}(X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} \\ \quad\times\left(\frac{\partial^2 H_1}{\partial\psi_d\partial\psi_e} - 2\frac{\partial H_1}{\partial\psi_d}H_1^{-1}\frac{\partial H_1}{\partial\psi_e}\right)H_1^{-1}X_1\end{array}\right]$$

$$-\frac{1}{2}\text{Trace}\left[\begin{array}{l}(X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1}\frac{\partial H_1}{\partial\psi_d}H_1^{-1}X_1 \\ \quad\times(X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1}\frac{\partial H_1}{\partial\psi_e}H_1^{-1}X_1)\end{array}\right].$$

Estimator of the MSE of the EBLUP has also been obtained with the assumption of large m and neglecting all terms of order $o(m^{-1})$ in Theorem A.4 of the Appendix as

$$\text{mse}[\hat{\theta}_t(\hat{\psi})] = [g_{12t}(\hat{\psi}) + g_{3t}(\hat{\psi}) + g_{31t}(\hat{\psi}) - g_{4t}(\hat{\psi}) - g_{5t}(\hat{\psi})] + o(m^{-1}), \quad (3.13)$$

where $g_{31t}(\psi)$, $g_{4t}(\psi)$ and $g_{5t}(\psi)$ are given as

$$g_{31t}(\psi) = L_t^T(\psi)[I_\psi^{-1}(\hat{\psi}) \otimes H_t(\psi)]L_t(\psi), \quad (3.14)$$

$$g_{4t}(\psi) = [b_{\hat{\psi}}^T(\psi) \otimes I_m]\frac{\partial g_{12t}(\psi)}{\partial(\psi)},$$

$$b_{\hat{\psi}} = \frac{1}{2}I_\psi^{-1}(\psi)\underset{1\le d\le 3}{\text{Col}}\left[\text{Trace}\left[I_\beta^{-1}(\psi)\frac{\partial I_\beta(\psi)}{\partial\psi_d}\right]\right], \quad (3.15)$$

$$g_{5t}(\psi) = \frac{1}{2}\underset{m}{\text{Trace}}\left[\begin{array}{l}[I_3 \otimes (R_t H_t^{-1})] \\ \frac{\partial^2 H_t}{\partial\psi\partial\psi^T}[I_\psi^{-1}(\psi) \otimes (H_t^{-1}R_t)]\end{array}\right]. \quad (3.16)$$

4. Analysis of the NSSO Data

National Sample Survey Organisation (NSSO) of the Ministry of Statistics and Programme Implementation (Government of India) conducts quinquennial large sample surveys (QS) on household consumption expenditure and employment, almost every five years in India. The surveys cover more than hundred thousand households spread over a number of villages and urban blocks. In order to fill the gaps in data between the successive QSs, the NSSO conducts annual consumer expenditure survey (CES) in almost every round (equivalent to six months or one year duration). The annual series covers only 10–30 thousand households depending on the number of villages and urban blocks surveyed all over the country. Each round of NSS normally has more than one subject of enquiry. The annual series has a different principal subject of enquiry. However schedule 1.0 of the annual surveys is designed to collect data on household consumption expenditure among other characteristics on employment.

The NSSO adopts two stage stratified sampling design, the first stage units being census villages in the rural sector selected through circular systematic sampling with probability proportional to size (PPS) and the ultimate-stage units being the households selected circular systematically with independent random starts. India has been divided into States and the Districts are the second level administrative units in the States. There is not much difference between the annual and quinquennial surveys excepting that normally in annual series, a small sample of four households per first stage units are surveyed while in the case of quinquennial survey, ten to twelve households per first stage units are surveyed. Besides this, in NSSO surveys, we have two samples *viz*, the first one as central sample surveyed by the investigators of the NSSO, and the second one as state sample surveyed by the State authorities. Regarding the estimation procedure, the first stage units are selected in the form of two independent sub-samples. The estimate of the population mean and its variance based on the two sub-samples are separately obtained. The pooled mean $y_i = (\hat{y}_{1i} + \hat{y}_{2i})/2$ and $R_i = (\hat{y}_{1i} - \hat{y}_{2i})^2/4$ for $i = 1, 2, \ldots, m$, where $\hat{y}_{1i}$, $\hat{y}_{2i}$ are the sub-sample means, estimate respectively the population mean and its variance for a particular district (small area). In case of round 55, first stage units are selected in the form of eight independent sub-samples and the estimate of the population mean and its variance are based on these sub-samples. In view of the problems related to the estimates of R_i's with 1 d.f., the R_i for each small area were analysed and compared over time. In case of any abnormal R_i, it was smoothed out by taking the average of R_i's over neighboring time points and in some cases, over neighboring small areas also. The survey estimates y_i's are the direct estimates, and the smoothed R_i's are the diagonal elements of the sampling variance covariance matrix R, in our model equations (2.1), (2.4) and (3.1), referred in this paper.

In this paper, we have used data from central sample only. The estimates of monthly per capita consumption expenditure (MPCE) and of the standard errors(SE) of the estimators have been obtained under various mixed effects models for the rural 63 districts (small areas) of a large state in India, namely, Uttar Pradesh. We have used data from the six rounds of the NSSO *viz* round 50 (July 1993–June 1994), round 51 (July 1994–June 1995), round 52 (July 1995–June 1996), round 53 (January–December 1997), round 54 (January–June 1998) and round 55 (July 1999–June 2000). Out of these rounds 50 and 55 are based on

quinquennial surveys. The selected exogenous variables used in the models are i) number of households, ii) gross area sown and iii) per capita net area sown in the districts. The agricultural data are available on annual basis while the estimates of the households and the population were obtained through the interpolation techniques based on the 1971, 1981 and 1991 decennial census data. These exogenous variables have been selected from a host of variables ranging from 1991 census to annual agricultural data through the covariate analysis. Different weight matrices such as length of common boundary between a pair of districts, distance between centres of two districts and the binary weights were considered. Binary weights give larger estimate of spatial autocorrelation coefficient, therefore they (standardised by making row sum of the weight matrix as one) have been used for further analysis in this paper. In the whole exercise, maximization of log likelihood function and the estimation of the parameters have been carried out by using the Nelder and Mead simplex method on the software MATLAB.

Various mixed effects models, used for finding out improved estimates of MPCE are given in Table 1. The parameters in the models have usual meaning as shown in sections 2 and 3. Further, in case of each model, sampling variance R or R_t (in case of temporal model) are assumed to be known.

Table 1
Mixed Effects Models

Model−1	Direct Estimates	
Model−2	Regression Model	$y = X\beta + v + \varepsilon$
Model−3	Spatial Model	$y = X\beta + Zv + \varepsilon$
Model−3A	Spatial Model (intercept)	$y = \mu + Zv + \varepsilon$
Model−4	Regression Temporal	$y_t = X_t\beta + v_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$
Model−5	Spatial Temporal	$y_t = X_t\beta + Zv_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$

Table 2 presents the round wise estimates of the parameters for the simple mixed effects regression and spatial models. The value of the multiple correlation coefficients R^2 between MPCE estimates and the auxiliary variables, in case of each round has also been shown here. The figures in bracket show the Standard Errors (SE) of the parameter estimates. Note that $\lambda(=\lambda_1, \lambda_2)$ is the likelihood ratio test (LRT) statistics defined as $-2\log L \sim \chi^2_k$, where L is the ratio of nested likelihoods at the hypothesised parameter values for two competing models under different hypotheses and k is the difference between the number of parameters under two models. Here λ_1 compares regression model and spatial model, under $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ and is distributed as χ^2_1 under H_0, and λ_2 compares spatial model and spatial (intercept) model, under $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ [β does not include intercept term β_0] and is distributed as χ^2_3 under H_0.

On comparison of the simple regression model (Model 2) and spatial model (Model 3) through LRT, we find that under $H_0(\rho = 0)$, the spatial autocorrelation ρ for Model 3 has been found highly significant for the two rounds 52 and 55, obviously for these rounds, use of spatial model results in much improvement in the estimates of MPCE. On the other hand, in case of rounds 50 and 53, and for these only, the regression coefficients β have been found nearly significant for the Model 3 in comparison to Model 3A which shows that the spatial model with intercept term may improve the estimates for these rounds without any help of the exogenous variables.

Table 3 presents the parameter estimates and their SE in case of regression temporal model and spatial temporal model.

For Model 4, unconstrained iterative maximisation process converged the value of k greater than 1, which is inadmissible under the assumption of stationarity. For this

Table 2
Estimates of Parameters for Small Area Estimates of MPCE Under Regression and Spatial Models

Round	R^2	Model 2	Model 3		LRT	Model 3A		LRT
		σ_v^2	ρ	σ_v^2	λ_1	ρ	σ_v^2	λ_2
Rd. 50	0.27	1,724.48	0.30	1,635.70	1.80	0.59	1,724.68	6.64
		(356.19)	(0.18)	(346.45)		(0.13)	(378.66)	
Rd. 51	0.27	3,424.21	0.48	3,156.90	0.66	0.67	3,022.32	4.54
		(820.89)	(0.19)	(815.24)		(0.13)	(824.54)	
Rd. 52	0.17	2,150.54	0.87	714.96	13.46	0.86	768.11	0.90
		(540.23)	(0.07)	(257.15)		(0.07)	(272.27)	
Rd. 53	0.13	6,312.99	−0.39	5,822.99	1.56	0.09	7,141.60	7.66
		(1,397.92)	(0.27)	(1,374.70)		(0.23)	(1,561.72)	
Rd. 54	0.22	3,437.67	0.61	2,793.24	1.30	0.66	2,888.66	3.00
		(806.87)	(0.14)	(742.35)		(0.13)	(768.84)	
Rd. 55	0.31	2,989.73	0.87	1,060.21	20.30	0.86	1,186.58	1.56
		(712.28)	(0.06)	(362.40)		(0.07)	(394.27)	

λ_1 and λ_2 compare models 2,3 and models 3,3A respectively. $\chi^2_{1,.05}$=3.841 for λ_1 and $\chi^2_{3,.05} = 7.815$ for λ_2.

case, estimates were obtained by taking $k=1$ and Model 4 was accordingly modified. Table 3 reports the results for $k=1$ in case of regression temporal model. The spatial temporal model shows higher value of common autocorrelation coefficient and far lower value of the estimate of σ_v^2. A summary of the round wise average estimates of MPCE (based on all the 63 districts), their estimated standard errors (SE) and the coefficient of variation (CV) under each model has been presented in Table 4.

The results of Table 4 have been summarized below.

The Direct survey estimates are less precise and all the models involving mixed effects improve it. The estimates for the rounds 50 and 55 (based on large samples) are more precise than the estimates based on other rounds. Spatial model, depending on the value of ρ improves the estimates considerably. In case of rounds 52 and 55, where the autocorrelation have been found significant, the reduction in the average SE of the estimates in comparison to the model without spatial autocorrelation, is considerable. Model 3A with spatial effect and without auxiliary variables is equally good. The spatial temporal model further improves the estimates taking into advantage of the state space considerations. It may be noted that for the round 52 (very high spatial autocorrelation), the estimates based on temporal models are worse than the estimates based on models without temporal considerations. Perhaps due to fixed regression and autocorrelation parameters, the estimates tend towards the average of the five rounds.

In order to judge the performances of the estimators under various models vis-a-vis under the most general model (spatial temporal model), data have been simulated under the spatial temporal model and true MSEs of the replicated estimates under each of the assumed models have been obtained. For this, we have conducted the simulation by taking the estimated parameters from the spatial temporal model, given in Table 2 and obtained the true replicated small area mean $\theta(b)$ for b^{th} replication $(b=1,2,\ldots,B)$ along with simulated observations $y(b)$ for a large number of replications. On this simulated dataset, for each replication, different models including spatial temporal model

Table 3
Estimates of Parameters for Small Area Estimation of MPCE Under Regression Temporal and Spatial Temporal Models

	ρ		σ_v^2		k	
Models	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Model 4	–	–	4,715.64	431.00	–	–
Model 5	0.79	0.04	2,163.50	245.50	0.53	0.07

Table 4
Average EBLUP for MPCE (Rs.), their Estimated SE and CV Under Regression, Spatial, Regression Temporal and Spatial Temporal Models

	NSSO Rounds					
Models	50	51	52	53	54	55
	Average Small Area Estimates					
Model 1	276.10	321.26	373.07	408.52	411.25	482.00
Model 2	272.87	312.53	354.45	397.52	400.87	471.99
Model 3	272.98	313.14	351.51	398.21	400.78	471.09
Model 3A	273.56	314.19	352.01	396.40	399.91	471.91
Model 4	274.13	305.62	345.54	383.53	399.56	463.32
Model 5	273.75	312.21	351.79	391.61	399.50	473.57
	Average Standard Errors (SE)					
Model 1	25.09	66.06	64.18	74.19	53.87	45.45
Model 2	17.10	33.65	29.09	39.85	32.68	30.59
Model 3	16.88	32.84	21.51	39.98	30.87	24.84
Model 3A	16.56	31.29	20.79	40.03	30.23	24.37
Model 4	19.51	34.91	35.19	37.79	35.14	33.15
Model 5	17.18	28.99	28.33	30.02	28.76	28.10
	Average Coefficient of Variation (CV) (%)					
Model 1	9.09	20.56	17.20	18.16	13.10	9.43
Model 2	6.27	10.79	8.21	10.01	8.15	6.48
Model 3	6.18	10.49	6.12	10.04	7.70	5.27
Model 3A	6.05	9.96	5.91	10.10	7.56	5.17
Model 4	7.12	11.42	10.18	9.85	8.79	7.15
Model 5	6.28	9.29	8.05	7.67	7.20	5.93

Table 5
Percentage Relative Efficiency [RMSE] of the Temporal Models in Comparison to other Models for MPCE

	NSSO Rounds					
	50	51	52	53	54	55
	Spatial Temporal Model [Model 5]					
Model 2	123.63	170.54	193.68	203.55	204.72	169.76
Model 3	100.24	133.82	149.70	165.46	165.85	154.23
Model 4	125.81	141.50	141.93	137.55	139.11	129.88
	Regression Temporal Model [Model 4]					
Model 2	100.71	134.50	156.35	165.30	163.13	152.56

have been applied and the small area mean estimators under each of them are obtained. While fitting the regression and spatial temporal models on the simulated datasets, the iterative maximisation process have the constrained value of $k <= 1$. Here we have taken $B = 5{,}000$ replications. The true MSEs of the estimators for i^{th} small area under a particular model $(k = 2-4)$ may be defined as

$$\text{MSE}(\theta_i^k) = \frac{1}{B}\sum_{k=1}^{B}[\hat{\theta}_i^k(b) - \theta_i(b)]^2, \quad i = 1, 2, \ldots, m.$$

The relative efficiency of the estimators under spatial temporal model (Model 5) against the estimators under models 2–4 have been judged by the ratio of their mean squared errors (RMSE) as

$$\text{RMSE}(k, \text{Temp}) = 100\frac{\sum_{i=1}^{m}\text{MSE}(\hat{\theta}_i^k)}{\sum_{i=1}^{m}\text{MSE}(\hat{\theta}_i^{\text{Temp}})}$$

where 'Temp' denotes the spatial temporal model and k denotes models 2, 3 and 4. Likewise the relative efficiency of the regression temporal model (Model 4) against the simple regression model (Model 2) has been found by simulating data with the estimated parameters given in Table 3, under the regression temporal model. The results have been shown in Table 5.

The results confirm the superiority of the spatial temporal model in comparison to other models for these parameters. The regression temporal model has also been found better than the simple regression model.

5. Conclusions

The Direct survey estimates based on the small sample can be considerably improved by using the area specific small area models. The spatial autocorrelation amongst the neighboring areas may be exploited for improving the direct survey estimates. However, the model must be applied after studying the significant correlation amongst the small areas by virtue of their neighborhood effects. In case of poor relation between the dependent and exogenous variables, the simple spatial model with intercept only, may equally improve the estimates. This model uses only the spatial autocorrelation to strengthen the small area estimates and do not require the use of exogenous variables. The spatial models, by using the appropriate weight matrix *W*, or a combination of *W* matrices, can considerably improve the estimates. Weight matrix should be based on logical considerations and it may be used effectively for the cases, where due to some reasons, reliable exogenous variables are not available. This aspect can be further exploited to find out the small area estimates for the areas which have been recently created/demarcated.

One has to be careful about the increase in the MSE due to the variability caused by replacing the parameters by their estimates. This gets reflected through the second order approximation to the MSE dealt in the paper. That is why many times the simple spatial model (with intercept) performs better than the spatial model involving more parameters. Use of time series data with fixed regression parameters across the time, further improves the small area estimates especially for the time points where the direct survey estimates have larger MSE. Spatial temporal models have advantage over temporal models without spatial consideration due to the inclusion of fixed spatial auto-correlation across the small areas. However, for some time points for which ρ may be very different than the rest, this may not hold due to estimates tending towards the average of five rounds. Here the temporal consideration can be started from a suitable initial time point. Finally the exogenous variables *X* and the weight matrix *W* supplement each other through the regression parameter β and the autocorrelation parameter ρ and a judicious use of them may result in considerable improvement in the small area estimates.

Acknowledgements

The unit level data for the research have been made available by the National Sample Survey Organisation (NSSO), Ministry of Statistics and Programme Implementation under a research collaboration between IIT Kanpur and the NSSO. The weight matrix containing the length of

the boundary between different small areas (districts) have been provided by the National Informatics Centre (NIC) of the Ministry of Information Technology, Government of India. We would like to thank the referees for their helpful comments which has considerably improved the paper.

Appendix

Theorem A.1: Under Regularity Conditions 1

$$\text{MSE}[\hat{\theta}(\hat{\psi})] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (5.1)$$

For proof of the Theorem, we use the following well known results (Srivastawa and Tiwari 1976). Let $U \sim N(0, \Sigma)$ then for the symmetric matrices A, B and C

$$E[U(U^T AU)U^T] = \text{Trace}(A\Sigma)\Sigma + 2\Sigma A\Sigma$$
$$E[U(U^T AU)(U^T BU)U^T] = \text{Trace}(A\Sigma)\text{Trace}(B\Sigma)\Sigma$$
$$+ 2[\text{Trace}(A\Sigma)\Sigma B\Sigma + \text{Trace}(B\Sigma)\Sigma A\Sigma + \text{Trace}(A\Sigma B\Sigma)\Sigma]$$
$$+ 4[\Sigma A\Sigma B\Sigma + \Sigma B\Sigma A\Sigma].$$

Proof of Theorem A.1

Kackar and Harville (1984) showed that $\text{MSE}[\hat{\theta}(\hat{\psi})] = \text{MSE}[\hat{\theta}(\psi)] + E[(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))^T]$. It is straight forward to show that $\text{MSE}[\hat{\theta}(\psi)] = g_1(\psi) + g_2(\psi)$. We need to prove that $g_3(\psi) = E[(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))^T] + o(m^{-1})$. Taylor Series expansion of $\hat{\theta}(\hat{\psi})$ around ψ and using $(\hat{\psi} - \psi) = O_p(m^{-1/2})$ and $(\partial^2\hat{\theta}(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}^*} = O_p(1)$ when $\| \hat{\psi}^* - \hat{\psi} \| \leqq \| \hat{\psi} - \psi \|$ we get

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi) \otimes I_m]^T \nabla\hat{\theta}(\psi) + O_p(m^{-1}). \quad (5.2)$$

Here $\nabla\hat{\theta}(\psi) = (\partial\hat{\theta}(\psi))/(\partial\psi) = [(\partial\hat{\theta}(\psi))/(\partial\rho), (\partial\hat{\theta}(\psi))/(\partial\sigma_v^2)]^T$. Using

$$\frac{\partial\hat{\theta}(\psi)}{\partial\psi_d} = \sum_{\alpha=1}^{p} \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\beta_\alpha}\Big|_{\beta=\hat{\beta}(\psi)} \frac{\partial\hat{\beta}(\psi)}{\partial\psi_d} + \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\psi_d}\Big|_{\beta=\hat{\beta}(\psi)}$$
$$d = 1, 2$$

where $\hat{\theta}^*(\beta,\psi) = X\beta(\psi) + \Lambda(\psi)[y - X\beta(\psi)]$, and the fact that $(\partial\hat{\beta}_\alpha(\psi))/(\partial\psi_d) = O_p(m^{-1/2})$ (Cox and Reid (1987)), we get from the above

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi)^T \otimes I_m]\nabla\hat{\theta}^*(\psi) + O_p(m^{-1}) \quad (5.3)$$

$$\text{where } \nabla\hat{\theta}^*(\psi) = \left[\frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\rho}, \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\sigma_v^2}\right]^T\Big|_{\beta=\hat{\beta}(\psi)}$$
$$= \text{L}(\psi)[y - X\hat{\beta}(\psi)].$$

Using the Regularity Conditions 1 and the fact that $\hat{\beta}(\psi) - \beta = O_p(m^{-1/2})$ we have

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]$$
$$= [(\hat{\psi} - \psi) \otimes I_m]^T L(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1})$$
$$= \sum_{d=1}^{2} (\hat{\psi}_d - \psi_d) L_d(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1}).$$

Further using the Taylor Series expansion of the Likelihood $S(\hat{\eta}) = 0$ around ψ where

$$S(\eta) = [S_\beta^T(\eta), S_\psi^T(\eta)]^T, \; S_\beta^T(\eta) = \underset{1\leq\alpha\leq p}{\text{Col}}\left[\frac{\partial\ell}{\partial\beta_\alpha}\right]$$

and the orthogonality of β and ψ, it follow that

$$(\hat{\psi} - \psi) = I_\psi^{-1}(\psi)S_\psi(\eta) + O_p(m^{-1}).$$

Writing

$$S_\psi(\psi) = \underset{1\leq d\leq 2}{\text{Col}}[S_d(\psi)] = [S_\rho(\psi), S_{\sigma_v^2}(\psi)]^T,$$
$$S_d(\psi) = \frac{\partial\ell}{\partial\psi_d} = -\frac{1}{2}\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right] + \frac{1}{2}[u^T B_d(\psi)u],$$
$$B_d(\psi) = \Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}, \; u = y - X\beta(\psi) \text{ and}$$
$$I_{de}(\psi) = \frac{1}{2}\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]$$

we get

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m][\text{S}_\psi(\psi) \otimes u]$$

and thus the expression

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)][\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]^T \text{ upto order } o(m^{-1})$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m] \underset{1\leq d\leq 2}{\text{Col}}[uS_d(\psi)] \underset{1\leq e\leq 2}{\text{Concat}}[S_e(\psi)u^T]$$
$$[I_\psi^{-1}(\psi) \otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m] \underset{1\leq d\leq 2}{\text{Col}} \underset{1\leq e\leq 2}{\text{Concat}}[uS_d(\psi)S_e(\psi)u^T]$$
$$[I_\psi^{-1}(\psi) \otimes I_m]L(\psi). \quad (5.4)$$

Now we can write the likelihood and its derivative as

$$\ell = \log L = \text{const.} - \frac{1}{2}\log[|\Sigma|] - \frac{1}{2}u^T\Sigma^{-1}u$$
$$\frac{\partial\ell}{\partial\psi_d} = -\frac{1}{2}\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right] + \frac{1}{2}u^T B_d(\psi)u,$$
$$B_d(\psi) = \Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}$$
$$E\left[-\frac{\partial^2\ell}{\partial\psi_d\partial\psi_e}\right] = \frac{1}{2}\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right] = I_{de}(\psi)$$

where information matrix $I_\psi(\psi) \equiv I_{de}(\psi)$.

The expectation of a typical element of the inner most terms in the expression (5.4) becomes

$$E[uS_d(\psi)S_e(\psi)u^T] =$$

$$E\begin{bmatrix} u[u^T B_d(\psi)u][u^T B_e(\psi)u]u^T \\ -u\,\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right][u^T B_e(\psi)u]u^T \\ -u[u^T B_d(\psi)u]\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]u^T \\ +u\,\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right]\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]u^T \end{bmatrix}$$

and by applying the results of Srivastawa and Tiwari (1976), it becomes

$$E[uS_d(\psi)S_e(\psi)u^T] =$$
$$\frac{1}{2}\text{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]\Sigma + 2\left[\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right].$$

Substituting these in the expression (5.4) and also the second expression being of order $O(m^{-1})$, we can get the following upto order $o(m^{-1})$

$$[\hat{\theta}(\hat{\psi})-\hat{\theta}(\psi)][\hat{\theta}(\hat{\psi})-\hat{\theta}(\psi)]^T$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m]\underset{1\le d\le 2}{\text{Col}}\,\underset{1\le e\le 2}{\text{Concat}}[I_{de}(\psi)\Sigma]$$
$$[I_\psi^{-1}(\psi)\otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m][I_\psi(\psi)\otimes\Sigma][I_\psi^{-1}(\psi)\otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes\Sigma]L(\psi).$$

Theorem A.2: Under Regularity Conditions 1

$$E[g_1(\hat{\psi})+g_3(\hat{\psi})-g_4(\hat{\psi})-g_5(\hat{\psi})] = g_1(\psi)+o(m^{-1}), \quad (5.5)$$

$$E[g_2(\hat{\psi})] = g_2(\psi)+o(m^{-1}),$$
$$E[g_3(\hat{\psi})] = g_3(\psi)+o(m^{-1}) \quad (5.6)$$

$$\text{and}\quad E[g_5(\hat{\psi})] = g_5(\psi)+o(m^{-1}). \quad (5.7)$$

Proof of Theorem A.2

Taylor Series expansion of $g_1(\hat{\psi})$ around ψ and using $\hat{\psi}-\psi = O_p(m^{-1/2})$ when $\|\hat{\psi}^*-\psi\| <= \|\hat{\psi}-\psi\|$, we get

$$g_1(\hat{\psi}) = g_1(\psi)+[(\hat{\psi})-(\psi)^T\otimes I_m]\nabla g_1(\psi)$$
$$+\frac{1}{2}[(\hat{\psi}-\psi)^T\otimes I_m]\nabla^2 g_1(\psi)[(\hat{\psi}-\psi)\otimes I_m]$$
$$+o_p(m^{-1})$$
$$\nabla g_1(\psi) = \left[\frac{\partial g_1(\psi)}{\partial\rho}\frac{\partial g_1(\psi)}{\partial\sigma_v^2}\right]^T,$$
$$\nabla^2 g_1(\psi) = \underset{1\le d\le 2}{\text{Col}}\left[\underset{1\le e\le 2}{\text{Concat}}\frac{\partial^2 g_1(\psi)}{\partial\psi_d\partial\psi_e}\right]$$
$$\frac{\partial g_1(\psi)}{\partial\psi_d} = R\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}R$$
$$\frac{\partial^2 g_1(\psi)}{\partial\psi_d\partial\psi_e} = -2R\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\Sigma^{-1}R$$
$$+R\Sigma^{-1}\frac{\partial^2\Sigma}{\partial\psi_d\partial\psi_e}\Sigma^{-1}R.$$

Using the fact that $\Sigma(\psi)$ and its derivatives are symmetric, we have the second term of the expression as

$$[(\hat{\psi}-\psi)^T\otimes I_m]\nabla^2 g_1(\psi)[(\hat{\psi}-\psi)\otimes I_m]$$
$$= -L^T(\psi)[I_\psi^{-1}(\psi)\otimes\Sigma]L(\psi)$$
$$+\frac{1}{2}\underset{m}{\text{Trace}}\left[[I_2\otimes(R\Sigma^{-1})]\frac{\partial^2\Sigma}{\partial\psi\partial\psi^T}[I_\psi^{-1}(\psi)\otimes(\Sigma^{-1}R)]\right]$$
$$= -g_3(\psi)+g_5(\psi)$$

where $I_\psi^{-1}(\psi) = \text{Var}(\psi)$ is information matrix, the asymptotic variance of ψ. The first term in the expression $[(\hat{\psi}-\psi)^T\otimes I_m]\nabla g_1(\psi)$ reduces to $g_4(\psi)$ because of $E(\hat{\psi}-\psi) = b_{\hat{\psi}}(\psi)$ up to order $o(m^{-1})$ (Peers and Iqbal 1985).

The second part of the Theorem follows from the Taylor series expansion of $g_2(\hat{\psi})$, $g_3(\hat{\psi})$ and $g_5(\hat{\psi})$, each around ψ and using $\hat{\psi}-\psi = O_p(m^{-1/2})$ and $(\partial^2 g_2(\psi))/(\partial\psi_d\,\partial\psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, $(\partial^2 g_3(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$ and $(\partial^2 g_5(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, respectively where $\|\hat{\psi}^*-\psi\| \le \|\hat{\psi}-\psi\|$.

Theorem A.3: Under Regularity Conditions 2

$$\text{MSE}([\hat{\theta}_t(\hat{\psi})] = g_{12t}(\psi)+g_{3t}(\psi)+o(m^{-1}). \quad (5.8)$$

Proof of Theorem A.3

The proof is basically on the line of Theorem A.1 and with the use of the results of (Srivastawa and Tiwari (1976)) mentioned therein.

$$\text{MSE}([\hat{\theta}_t(\hat{\psi})]$$
$$= \text{MSE}([\hat{\theta}_t(\psi)]+E([\theta_t(\psi)-\theta_t)(\theta_t(\psi)-\theta_t)^T]$$
$$= g_{12t}(\psi)+E[(\theta_t(\psi)-\theta_t)(\theta_t(\psi)-\theta_t)^T]. \quad (5.9)$$

Taylor series expansion of $\theta_t(\psi)$ around ψ and using $(\hat{\psi}-\psi)=O_p(m^{-1/2})$ and $(\partial^2\hat{\theta}(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}^*}=O_p(1)$ when $\|\hat{\psi}^*-\psi\| \leqq \|\hat{\psi}-\psi\|$ we have

$$\begin{aligned}[\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi)] &= [(\hat{\psi}-\psi)\otimes I_m]^T\nabla\hat{\theta}_t(\psi)+O_p(m^{-1}) \\ &= \sum_{d=1}^{3}[(\hat{\psi}_d-\psi_d)L_{td}(\psi)e_t(\psi)]+O_p(m^{-1}). \qquad (5.10)\end{aligned}$$

Further using the Taylor series expansion of the Likelihood equation $S(\hat{\eta})=0$ and the orthogonality of β and ψ, it follows

$$(\hat{\psi}-\psi)=I_\psi^{-1}(\psi)S(\psi)+O_p(m^{-1}). \qquad (5.11)$$

Substituting the expression for $(\hat{\psi}-\psi)$ in equation (5.10), we have up to order $o(m^{-1})$

$$[\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi)]=L_t^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m][S_\psi(\psi)\otimes e_t] \qquad (5.12)$$

and

$$\begin{aligned}&[(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))^T] \\ &= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m]\underset{1\le d\le 3}{\text{Col}}\ \underset{1\le e\le 3}{\text{Concat}} \\ &[e_tS_d(\psi)S_e(\psi)e_t^T][I_\psi^{-1}(\psi)\otimes I_m]L(\psi) \qquad (5.13)\end{aligned}$$

where

$$S_\psi(\psi)=\underset{1\le d\le 3}{\text{Col}}\,[S_d(\psi)],\quad S_d(\psi)=\frac{\partial\ell}{\partial\psi_d}.$$

Using the expression for derivatives of likelihood, we have

$$S_d(\psi)=\frac{1}{2}\left[\begin{array}{l}\text{Trace}[C_{1d}(\psi)]-\sum_{t=1}^{T}\text{Trace}\left[H_t^{-1}\frac{\partial H_t}{\partial\psi_d}\right] \\ +\sum_{t=1}^{T}[e_t^TB_{td}(\psi)e_t]\end{array}\right]$$

$$-\left[e_t^TH_t^{-1}\frac{\partial e_t}{\partial\psi_d}\right]C_{1d}(\psi)=\left[(X_1^TH_1^{-1}X_1)^{-1}X_1^TH_1^{-1}\frac{\partial H_1}{\partial\psi_d}H_1^{-1}X_1\right],$$

$$B_{td}(\psi)=H_t^{-1}\frac{\partial H_t}{\partial\psi_d}H_t^{-1}.$$

By applying the considerations $e_t\sim N(0,H_t)$, $\text{Corr}(e_i,e_j)=0$ for $i\neq j$, $\text{Corr}(e_i,(\partial e_i)/(\partial\psi_d))=0$ and $\text{Corr}(e_i,(\partial^2e_i)/(\partial\psi_d\partial\psi_e))=0$ due to the fact that $(\partial e_i)/(\partial\psi_d)=(\partial(y_i-U_t\hat{\alpha}_{t|t-1}))/(\partial\psi_d)$ being linear function of $(y_1,y_2,\ldots,y_{t-1})$ is uncorrelated with e_t, we get the expectation of the inner most terms of the expression (5.13) as

$$\begin{aligned}E[e_tS_d(\psi)S_e(\psi)E_t^T] &= K_{de}(\psi)H_t+2\left[\frac{\partial H_t}{\partial\psi_d}H_t^{-1}\frac{\partial H_t}{\partial\psi_d}\right] \\ &+\frac{1}{2}\left[\text{Trace}[B_{td}(\psi)]\frac{\partial H_t}{\partial\psi_e}+\text{Trace}[B_{te}(\psi)]\frac{\partial H_t}{\partial\psi_d}\right] \\ &+\frac{1}{4}\text{Trace}[B_{td}(\psi)]\text{Trace}[B_{te}(\psi)]H_t\end{aligned}$$

where

$$K_{de}(\psi)=\frac{1}{2}\sum_{t=1}^{T}\text{Trace}\left[H_t^{-1}\frac{\partial H_t}{\partial\psi_d}H_t^{-1}\frac{\partial H_t}{\partial\psi_e}\right].$$

The middle three terms in the expression being of order $O(1)$ which along with $I_\psi^{-1}(\psi)$ in the expression given below makes them of order $o(m^{-1})$,

$$\begin{aligned}E[(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))^T] &= g_{3t}(\psi) \\ &= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m][K_\psi(\psi)\otimes H_t] \\ &\quad [I_\psi^{-1}(\psi)\otimes I_m]L(\psi)+o(m^{-1}) \\ &= L^T(\psi)[I_\psi^{-1}(\psi)K_\psi(\psi)I_\psi^{-1}(\psi)\otimes H_t]L(\psi)+o(m^{-1}).\end{aligned}$$

Theorem A.4: Under Regularity Conditions 2

$$\begin{aligned}E[g_{12t}(\hat{\psi})+g_{3t}(\hat{\psi})+g_{31t}(\hat{\psi})-g_{4t}(\hat{\psi})-g_{5t}(\hat{\psi})] &= g_{12t}(\psi)+o(m^{-1}) \\ E[g_{3t}(\hat{\psi})] &= g_{3t}(\psi)+o(m^{-1})\end{aligned}$$

and

$$E[g_{5t}(\hat{\psi})]=g_{5t}(\psi)+o(m^{-1}).$$

Proof of Theorem A.4

The proof is essentially based on the line suggested in proving Theorem A.2. Using Taylor series expansion of $g_{12t}(\hat{\psi})$ around ψ, we get

$$\begin{aligned}g_{12t}(\hat{\psi}) &= g_{12t}(\psi)+[(\hat{\psi}-\psi)\otimes I_m]^T\nabla g_{12t}(\psi) \\ &+\frac{1}{2}[(\hat{\psi}-\psi)^T\otimes I_m]\nabla^2g_{12t}(\psi)[(\hat{\psi}-\psi)\otimes I_m] \\ &+o_p(m^{-1})\end{aligned}$$

$$\nabla g_{12t}(\psi)=\underset{1\le d\le 3}{\text{Col}}\,[\nabla g_{12td}(\psi)],\ \nabla g_{12td}(\psi)=\frac{\partial g_{12t}(\psi)}{\partial\psi_d}$$

$$\nabla^2g_{12t}(\psi)=\underset{1\le d\le 3}{\text{Col}}\left[\underset{1\le e\le 3}{\text{Concat}}\frac{\partial^2g_{12t}(\psi)}{\partial\psi_d\partial\psi_e}\right]$$

$$\frac{\partial g_{12t}(\psi)}{\partial\psi_d}=R\Sigma^{-1}\frac{\partial R}{\partial\psi_d}\Sigma^{-1}R$$

$$\begin{aligned}\frac{\partial^2g_{12t}(\psi)}{\partial\psi_d\partial\psi_e} &= -2R\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\Sigma^{-1}R \\ &+R\Sigma^{-1}\frac{\partial^2\Sigma}{\partial\psi_d\partial\psi_e}\Sigma^{-1}R.\end{aligned}$$

Using the fact that $\Sigma(\psi)$ and its derivatives are symmetric, we have the second term of the expression as

$$
\begin{aligned}
&[(\hat{\psi}-\psi)^T \otimes I_m]\nabla^2 g_{12t}(\psi)[(\hat{\psi}-\psi)\otimes I_m] \\
&= -L^T(\psi)[I_\psi^{-1}(\psi)\otimes\Sigma]L(\psi) \\
&\quad + \frac{1}{2}\underset{m}{\text{Trace}}\left[[I_3\otimes(R\Sigma^{-1})]\frac{\partial^2\Sigma}{\partial\psi\partial\psi^T}[I_\psi^{-1}(\psi)\otimes(\Sigma^{-1}R)]\right] \\
&= -g_{3t}(\psi) = g_{5t}(\psi)
\end{aligned}
$$

where $I_\psi^{-1}(\psi) = \text{Var}(\psi)$ is the asymptotic variance of ψ. The first term in the expression $[(\hat{\psi}-\psi)^T \otimes I_m]\nabla g_{12t}(\psi)$ reduces to $g_{4t}(\psi)$ because of $E(\hat{\psi}-\psi) = b_{\hat{\psi}}(\psi)$ up to order $o(m^{-1})$ (Peers and Iqbal 1985).

The second part of the Theorem follows from the Taylor series expansion of $g_{3t}(\hat{\psi})$ and $g_{5t}(\hat{\psi})$, each around ψ and using $\hat{\psi}-\psi = O_p(m^{-1/2})$ and $(\partial^2 g_{3t}(\psi))/(\partial\psi_d\,\partial\psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$ and $(\partial^2 g_{5t}(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, respectively where $\|\hat{\psi}^* - \hat{\psi}\| \leq \|\hat{\psi}-\psi\|$.

References

Cliff, A.D., and Ord, J.K. (1981). *Spatial Processes, Models and Applications*. Pion Limited, London.

Cox, D.R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, Series B, 49, 1-39 (with discussion).

Cressie, N. (1990). Small-Area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.

Cressie, N. (1992). REML estimation in empirical bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimates for best linear unbiased predictors in small area estimation problem. *Statistica Sinica*, 10, 613-627.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistics Association*, 74, 267-277.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistics Association*, 79, 853-862.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1997). Report No. 404(50/1.0/1), Consumption of some important commodities in India. *NSS* 50th *Round, July 1993-June 1994*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 436(51/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 51st *Round, July 1994-June 1995*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 440(52/1.0/1), Household consumption expenditure and employment situation in India, *NSS* 52nd *Round, July 1995-June 1996*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 442(53/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 53rd *Round, January-December 1997*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1999). Report No. 448(54/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 54th *Round, January-June 1998.*

National Sample Survey Organisation, Ministry of Statistics and Programme Implementaion, Govt. of India (2001). Report No. 472(55/1.0/1), Differences in level of consumption among socio-economic groups. *NSS* 55th *Round, July 1999-June 2000*.

Peers, H.W., and Iqbal, M. (1985). Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society*, Series B, 47, 547-554.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean square error of small area estimates. *Journal of the American Statistics Association*, 85, 163-171.

Rao, C.R., and Toutenbourg, Heldge (1999). *Linear Models*: *Least Squares and Alternatives*. Second Edition, New York: Springer.

Rao, J.N.K. (1999). Some recent advances in model based small area estimation. *Survey Methodology*, 25, 175-186.

Ripley, B.D. (1981). *Spatial Statistics*. New York: John Wiley & Sons, Inc.

Sallas W.M., and Harville D.A. (1994). Noninformative priors and restricted likelihood estimation in the Kalman filter. *Bayesian Analysis of Time Series and Dynamic Models*, (Ed. James C. Spall). New York: Marcel Dekker Inc., 477-508.

Singh, A.C., Mantel, H.J. and Thomas B.W. (1994). Time series EBLUPs for small area using survey data. *Survey Methodology*, 20, 33-43.

Singh, A.C., Stukel, D. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.

Srivastava, V.K., and Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 197-212
Statistics Canada, Catalogue No. 12-001-XPB

Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey

Liv Belsby, Jan Bjørnstad and Li-Chun Zhang [1]

Abstract

This paper considers the problem of estimating, in the presence of considerable nonignorable nonresponse, the number of private households of various sizes and the total number of households in Norway. The approach is model-based with a population model for household size given registered family size. We account for possible nonresponse biases by modeling the response mechanism conditional on household size. Various models are evaluated together with a maximum likelihood estimator and imputation-based poststratification. Comparisons are made with pure poststratification using registered family size as stratifier and estimation methods used in official statistics for The Norwegian Consumer Expenditure Survey. The study indicates that a modeling approach, including response modeling, poststratification and imputation are important ingredients for a satisfactory approach.

Key Words: Household size; Nonresponse; Imputation; Poststratification.

1. Introduction

This work is motivated by the considerable nonresponse rate in the Norwegian Consumer Expenditure Surveys (CES) for private households, for example 32% in the 1992 survey. Nonresponse involves both noncontact and refusal. We focus on the problem of nonignorable nonresponse that occurs when estimating the number of households of various sizes and the total number of households.

We shall consider a completely model-based approach; modeling and estimating the distribution of household size given registered family size and the response mechanism conditional on the household size. This model takes into account that the nonresponse mechanism may be nonignorable, in the sense that the probability of response is allowed to depend on the size of the household. The response model is used to correct for nonresponse. Model-based approaches with nonresponse included, sometimes called the prediction approach, have been considered by, among others, Little (1982), Greenlees, Reece and Zieschang (1982), Baker and Laird (1988), Bjørnstad and Walsøe (1991), Bjørnstad and Skjold (1992) and Forster and Smith (1998).

For various models of household size and response we consider mainly two model-based approaches, a maximum likelihood estimator and imputation-based poststratification after registered family size. These methods are compared to pure poststratification and the methods in current use in CES.

The main issue here is a comparison of models and methods with estimation bias as the basic problem. In addition, standard errors of the estimates and differences of the estimates, conditional on the sizes of post-strata determined by family size, are estimated using a bootstrap approach. In addition to assessing the statistical uncertainty of the estimators, this is done to help evaluate the extent to which differences between the proposed estimators are attributable to sampling error, nonresponse bias or both. However, in this evaluation we keep in mind the following quote from Little and Rubin (1987, page 67): "It is important to emphasize that in many applications the issue of nonresponse bias is often more crucial than that of variance. In fact, it has been argued that providing a valid estimate of sampling variance is worse than providing no estimate if the estimator has a large bias, which dominates the mean squared error."

Section 2 describes the data-structure and the sample design of CES, and Section 3 considers modeling issues. Section 3.1 presents the various models for household size and response to be considered for the 1992 CES, Section 3.2 describes the maximum likelihood method for parameter estimation, and in Section 3.3 the models are evaluated. A family size group model for household size and a logistic link for the response probability using household size as a categorical variable give the best fit of the models under consideration. Section 3.4 gives the estimated household size distributions for different family sizes and estimated response probabilities for different household sizes.

1. Liv Belsby, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lbe@ssb.no; Jan F. Bjørnstad, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: jab@ssb.no and Li-Chun Zhang, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lcz@ssb.no.

Section 4 considers model-based estimation, the imputation method, imputation-based estimators and the variance estimation method. It is shown that for the chosen model for household size from Section 3.3, the maximum likelihood estimator and the imputation-based poststratified estimator are identical.

Section 5 deals with the main goal of estimating the total number of household of various sizes based on the 1992 CES, using the estimators in Section 4. The model that gave the best fit seems to work well for our estimation problem. We conclude that poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

2. Norwegian Consumer Expenditure Survey

The population totals within household-size categories provide a more correct number of dwellings than the totals within family-size categories from the Norwegian Family Register. Furthermore, the authorities for evaluating eventual policy intervention aimed at housing construction use the estimated number of households. Estimating household-size totals is therefore an important issue in social planning. It is invariably affected by nonignorable nonresponse, no matter what kind of survey one uses. Hence, it is a good illustration for how to handle nonresponse bias. We shall base our estimation on the Norwegian Consumer Expenditure Surveys (CES), where it is important to gain information about the composition of households, since household size influences consumption.

The actual CES, the survey for expenditure variables, is a sample of private households from all private households in Norway. This is done by selecting a sample of persons and including the whole households these persons belong to. Persons older than 80 years old are excluded since they often live in institutions. For our purpose, the units of interest in the survey are *persons* between the ages of 16 and 80 living in private households, and the variable of interest is the size of the *household* the person belongs to, which is observed only in the response sample of the persons selected.

The sample design is a three-stage self-weighting sample of persons. That is, every person in the population has the same inclusion probability to the total sample. The first two stages select geographical areas in a stratified way, while at the third stage persons are selected randomly from the chosen geographical areas. The primary sampling units (PSU) at stage 1 consists of the municipalities in Norway. Municipalities with less than 3,000 inhabitants are grouped together such that each PSU consists of at least 3,000 persons. The PSUs are first grouped into 10 regions and within each region stratified according to size (number of inhabitants) and type of municipality (*i.e.*, industrial structure and centrality). Totally, we have 102 strata. Towns of more than 30,000 inhabitants are their own strata and therefore selected with certainty at stage 1. For the other strata, one PSU is selected with probability proportional to size. At the second stage, the selected PSUs are divided into three smaller areas (secondary sampling units, SSU) and one of these is selected at random. Finally, at the third stage, for each of the selected SSU, a random sample of persons is selected. The sample sizes for each selected SSU are determined such that the resulting total sample of persons is self-weighting.

Our application is based on the data from the 1992 CES. CES is a yearly survey and since 1992 a modified Horvitz-Thompson estimator, including a correction for nonresponse by estimating response probabilities given household size, has been employed (see Belsby 1995). The weights equal the inverse of the probability of being selected multiplied with the conditional probability of response given selected. Since 1993 the probability of response is estimated with a logistic model with auxiliary variables being place of residence (rural/urban), and household size. For most of the nonrespondents the family size is used as a substitute for the household size.

A household is defined as persons having a common dwelling and sharing at least one meal each day (having common board). For a complete description of CES we refer to Statistics Norway (1996). In CES, the auxiliary variables known for the total sample, including the nonrespondents, are the family size, the time of the survey (summer/not summer), and the place of residence (urban/rural). *Families* are registered in Norwegian Family Register, (*NFR*), and may differ from the household the persons in the family belong to, both by definition and because of changes not yet registered. Hence, the registered *family* size from *NFR* differs to some extent from the household size. Initially, based on experience from previous surveys, all the auxiliary variables and household size are assumed to affect the response rate.

Table 1 shows the data for the 1992 CES with a total sample of 1,698 persons. The households with size five and greater are collapsed due to the low frequency in the sample of households. We base our modeling and estimation on two corresponding tables, one for the persons in rural areas and one for the persons in urban areas. These data are given in table A1 in appendix A1.

For example, the number 48 in cell (1,2) means that of the 162 persons registered to live alone in the response sample, 48 are actually living in a two-persons household. This is explained mostly by young people's tendency to cohabitate without being married; see Keilman and Brunborg (1995).

Table 1
Family and household sizes for the 1992 Norwegian Consumer Expenditure Survey

	Household size							
Family size	1	2	3	4	≥ 5	Total	Nonresponse	Response rate
1	83	48	20	9	2	162	153	0.514
2	9	177	37	4	3	230	160	0.590
3	10	25	131	40	6	212	91	0.700
4	2	13	37	231	17	300	123	0.709
≥ 5	1	4	4	17	181	207	60	0.775
Total	105	267	229	301	209	1,111	587	0.654

3. Modeling of Household Size and Nonresponse

We shall assume a population model for the household size, given auxiliary variables, *i.e.*, we model the conditional probability. To take nonresponse into account in the statistical analysis, we must model the response mechanism, *i.e.*, the distribution of response conditional on the household size and auxiliary variables. The sampling mechanism for persons is ignorable for the survey we consider, *i.e.*, is independent of the population vector of household sizes. The statistical analysis is therefore done *conditional* on the total sample, following the likelihood principle (see Bjørnstad 1996). Hence, probability considerations based on the sampling design is irrelevant in the statistical analysis. This is the so-called prediction approach. However, when evaluating the estimation methods with regard to statistical uncertainty, we do this from a common randomization perspective as described in Section 4.3.

For CES, the auxiliary vector consists of the family size, place of residence divided into rural and urban areas, and time of the data collection.

3.1 The Models

Let us first consider a simple model for the household size, denoted by *Y*. Let $\mathbf{x}$ denote all auxiliary variables. The household size is assumed to depend only on the family size *x*, and as such is a model with a restricted parametric link function, but with no additional assumptions,

$$P(Y_i = y \mid \mathbf{x}_i) = P(Y_i = y \mid x_i) = p_{y, x_i}, \qquad (3.1)$$

where

$$\sum_y p_{y, x_i} = 1, \text{ for each possible value of } x_i.$$

The model (3.1) is flexible in the sense that it does not include any restrictions on the assumed model function of x_i. The drawback is the high number of parameters compared with a model using a logistic type model with a linear, in $\mathbf{x}$, link function (the function linking $P(Y = y)$ with $\mathbf{x}$). If nonresponse is ignored the estimates in this model would simply be the observed rates.

Household size defines ordered categories. Thus a natural choice for a model is the cumulative logit model, known as the proportional-odds model (see McCullagh and Nelder 1991), assuming (with θ_y increasing in y)

$$P(Y_i \leq y \mid \mathbf{x}) = \begin{cases} \dfrac{1}{1 + \exp(-\theta_y + \beta^t \mathbf{x})} & \text{for } y = 1, 2, 3, 4 \\ 1 & \text{for } y \geq 5. \end{cases}$$

However, a goodness of fit test, with $\mathbf{x}$ consisting of family size and place of residence, indicated that this model fits the data badly. Thus we choose to reject it.

It is assumed that the probability of nonresponse may depend on the household size. For example, one-person households are less likely to respond than households of larger size since larger households are easier to "find at home". Nonresponse is indicated by the variable R, where $R_i = 1$ if person i responds and 0 otherwise. Let R_s be the vector of these indicators in the total sample. From Bjørnstad (1996), the response mechanism (RM), *i.e.*, the conditional distribution of R_s given the $\mathbf{x}-$values in the population and the $y-$values in the total sample, is defined to be ignorable if it can be discarded in a likelihood-based analysis. This means that RM is ignorable if this conditional distribution of R_s does not depend on the unobserved $y-$values, coinciding with the definition used by Little and Rubin (1987, pages 90, 218). For our case it is assumed that all pairs (Y_i, R_i) are independent. Then RM is ignorable if Y_i and R_i are independent. Hence, nonignorable response mechanism is equivalent to

$$P(Y_i = y_i \mid \mathbf{x}_i, r_i = 0) \neq P(Y_i = y_i \mid \mathbf{x}_i, r_i = 1)$$

and then both are different from $P(Y_i = y_i \mid \mathbf{x}_i)$.

Thus estimating the parameters in the model for $P(Y = y \mid \mathbf{x})$ using only the response sample, ignoring that the probability of response depends on the household size, would most likely give biased estimates for the unknown parameters. Also the poststratification estimator would give

biased estimates because it assumes that the distribution of R only depends on the auxiliary $\mathbf{x}$. *E.g.*, the observed lower response rate among one-person families indicates that the same may hold for one-person households. If so, the estimated probability of household size 1, based on respondents only, would be too small. Poststratification with respect to family size will most likely correct only some of this bias.

The model for the probability of response, given auxiliary variables and household size y_i, is assumed to be logistic. It depends on the auxiliary variables $\mathbf{z}_i$, which includes part of $\mathbf{x}_i$, expressed by

RM1$(y, \mathbf{z})$:

$$P(R_i = 1 | y_i, \mathbf{z}_i) = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)}. \quad (3.2)$$

Here, α and γ are scalar parameters and ψ is a vector. The variable y_i has an order. Motivated by this fact, and to avoid introducing many parameters, y_i is used in (3.2) as an ordinal variable rather than a class variable. Thus the logit function,

$$\log\{P(R_i = 1 | y_i, \mathbf{z}_i) / P(R_i = 0 | y_i, \mathbf{z}_i)\} = \alpha + \gamma y_i + \psi^t \mathbf{z}_i,$$

is linear in y_i. To avoid the assumption of linear logit in y_i, we also consider a model with y_i as a categorical variable, *i.e.*,

$$\text{RM2}(y, \mathbf{z}): P(R_i = 1 | y_i, \mathbf{z}_i) = \frac{1}{1 + \exp\begin{pmatrix} -\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) \\ -\alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi^t \mathbf{z}_i \end{pmatrix}}, \quad (3.3)$$

where the indicator variable $I_y(y_i)$ equals 1 if $y_i = y$ and 0 otherwise. The drawback with this model is that it includes three parameters more than model (3.2).

3.2 Maximum Likelihood Parameter Estimation

All the selected persons in the sample are from different households (duplicates have been removed), The population model then assumes that the household sizes Y_i are statistically independent. For *this* variable, interviewer- or cluster- effect plays no role.

Let us consider the likelihood function for estimating the unknown parameters, assuming that all pairs (Y_i, R_i) are independent and response model RM1 given by (3.2). To simplify notation we relabel the observations such that observations 1 to n_r are the respondents and observations $n_r + 1$ to n are the nonrespondents. With response model RM2 the expression for the likelihood is of the same form with (3.3) replacing (3.2).

For the respondents let $L_i = P(Y_i = y_i \cap R_i = 1 | \mathbf{x}_i)$. Then, for model (3.1)

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)} \cdot p_{y_i, x_i}, \; i = 1, ..., n_r \quad (3.4)$$

For the nonrespondents let $L_i = P(R_i = 0 | \mathbf{x}_i)$. Then

$$L_i = \sum_{y=1}^{5} \frac{1}{1 + \exp(\alpha + \gamma y + \psi^t \mathbf{z}_i)} \cdot p_{y, x_i}, i = n_r + 1, ..., n. \quad (3.5)$$

The likelihood function for the entire sample of persons from different households is given by

$$L(\theta, \beta, \alpha, \gamma, \psi) = \prod_{i=1}^{n} L_i. \quad (3.6)$$

For $i = 1, ..., n_r$, L_i is according to (3.4) and for $i = n_r + 1, ..., n$, L_i is given by (3.5).

Estimates are found by maximizing the likelihood function (3.6). The maximization was done numerically using the software TSP (1991) see Hall, Cummins and Schnake (1991). The optimizing algorithm is a standard gradient method, using the analytical first and second derivatives. These are obtained by the program, saving us a substantial piece of programming. The model fitting is based on the chi-square statistic and on the $t-$values, provided by TSP, where the standard errors are derived from the analytical second derivatives. The $t-$values have to be interpreted with some care, since the unbiasedness of the estimated standard errors depends on how well the model is specified as well as the number of observations compared with the number of parameters.

3.3 Evaluation of the Models for Household Size and Response

We present the fit of the models with the Pearson goodness-of-fit statistics. The model study is based on the 1992 CES. The parameters are considered to be significant when the absolute $t-$values are greater than 2. However, we do not want a model that is too restrictive, and therefore some variables are kept even though their absolute $t-$values are less than 2.

In the response models RM1 and RM2 we use the variable $\mathbf{z} = z$, place of residence. We let $z = 0$ if rural area and $z = 1$ if urban area. It was observed in the CES $1986-88$ and CES $1992-94$, see Statistics Norway (1990, 1996), that there is more nonresponse during the summer. Therefore, the time of the survey was also included in the model, that is whether or not the data were collected in the period May $21-$ August 12. However, the time of the survey was found to be nonsignificant, with $t-$value clearly less than 2. Also the family size was found to be nonsignificant. But if the household size is omitted in the response model then the family size turns out to be significant.

Ideally, we want to take a look at the empirical logit function for response with respect to the household size. However, household size is unavailable for the non-respondents. As a replacement we plot the logit-function against the family size; see figure 1. From family size one to two the two functions for rural and urban families increase in a fairly parallel way . However, for family size three and four the logit functions depart from being linear and parallel. Thus we suspect that coding the household size as a categorial variable, as in model RM2, will give better fit than restricting the logit functions to be parallel for rural and urban and linear with respect to the household size, as in model RM1.

In order to test the goodness of fit of the models, we consider the Pearson chi-square statistic, conditional on the auxiliary variables x, z. Given rural or urban type of residence and registered family size, there are six possible outcomes; household sizes 1, ..., 5 and nonresponse. Altogether there are ten multinomial trials and sixty cells. For family sizes (1,2) and (4,5), the extreme household sizes (4,5) and (1,2), respectively, are combined because the expected sizes under the models are too small. This reduces the number of cells to 52. The degrees of freedom (d.f.) is calculated as: number of cells – number of trials – number of parameters. For model (3.1) & RM1(y, z), d.f. = $52 - 10 - (20 + 3) = 19$, and for (3.1) & RM2(y, z), d.f. = $52 - 10 - (20 + 6) = 16$. For model (3.1) & RM1(y, z) the Pearson statistic χ^2 is 26.35 and the $p-$value is 0.121. And for model (3.1) & RM2(y, z) χ^2 is 21.77 and the $p-$ value is 0.151.

By studying the standardized residuals, (*observed-expected*)$/\sqrt{\hat{\text{V}}\text{ar}(\text{observed})}$, we find that the main reason for the better fit is that model (3.1) & RM2(y, z) does a better job of predicting the observed counts for the urban area where the response rate is lowest (see appendix A1). Thus the data indicates that coding the household size as a categorial variable, as in RM2, improves the fit compared to using it as an ordinal variable. The model (3.1), with the restricted parametric link function, combined with RM2 is the best of the models we have considered so far.

3.4 Estimated Household Size Distribution and Response Probabilities

Table 2 displays the estimates for the population model (3.1) together with the logistic response model RM2 in (3.3).

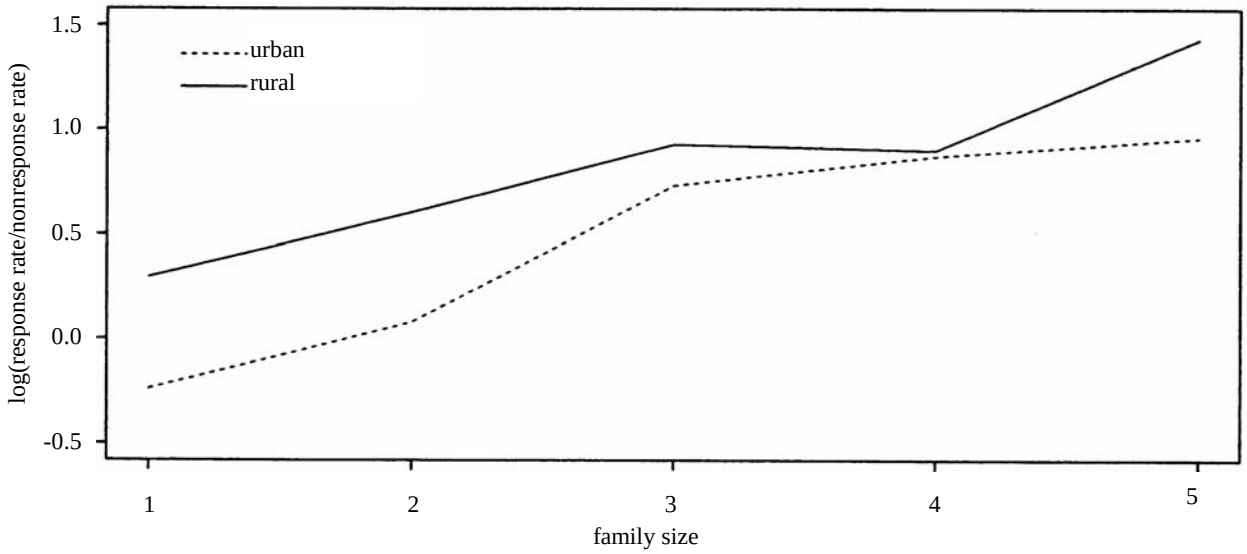


Figure 1. The logit function for the empirical response rate with respect to family size 1, ..., 5 in urban and rural areas, respectively. The computation is based on respondents and nonrespondents from Table 1 in Appendix A1.

Table 2
1992 CES. Parameter Estimates, in Percentages, for the Population Model with a Restricted Parametric Link Function, $p_{y,x}$, Combined with the Logistic Response Model RM2(y, z). In Parentheses are the Estimates for the Population Model, Ignoring the Response Mechanism

	Household size				
Family size, x	1	2	3	4	5 or more
1	60.01 (51.23)	26.75 (29.63)	8.35 (12.35)	4.09 (5.56)	0.80 (1.23)
2	5.27 (3.91)	79.80 (76.98)	12.48 (16.09)	1.47 (1.74)	0.98 (1.30)
3	7.53 (4.72)	14.45 (11.79)	56.67 (61.79)	18.85 (18.87)	2.50 (2.83)
4	1.06 (0.67)	5.31 (4.33)	11.38 (12.33)	77.20 (77.00)	5.05 (5.67)
5 or more	0.84 (0.48)	2.60 (1.93)	1.96 (1.93)	9.05 (8.21)	85.55 (87.44)

Let us interpret some of the values in the household model. Taking the response mechanism into account has largest effect on the estimated household distribution for one-person families. The probability that a household size equals one, given that the family size is one, is estimated as 60.01%. The estimate based on the traditional approach, ignoring the nonresponse, is 51.23%. The response model "adjusts" the observed rate among the respondents to a higher value. This seems reasonable since the rate of non-respondents is higher for small households. The estimated probability of household size five or more, given family size of five or more is 85.55%, which differs little from the observed rate among the respondents, 87.44%. This indicates that, given family size five or more, the household size distribution is about the same among respondents and nonrespondents.

Table 3 presents the estimated response probabilities based on RM2 in combination with the population model (3.1). Furthermore, we present estimated response probabilities based on a saturated model, with perfect fit, presented in Section 4.2. The model, defined by (4.9), assumes that the response probability for persons with the same household size within rural/urban area, respectively, is identical for different family sizes. Moreover, the model for household size depends on place of residence and family size, but with no restriction on the link function. We note that RM2(y, z) satisfies (4.9b), but is more restrictive. Model (4.9) allows for more freedom than model (3.1) with RM2(y, z).

Table 3
Estimated Probability of Response Based on the Logistic Model RM2 in Combination with (3.1), and the Saturated Model (4.9). The Estimates are Given in Percentages

	Household size				
Place of residence	1	2	3	4	5 or more
	Estimated response probabilities for model RM2				
Rural	47.77	60.90	79.16	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46
	Estimated response probabilities for the saturated model				
Rural	50.79	62.37	76.90	70.57	83.07
Urban	35.17	50.85	74.79	70.68	72.89

The estimated response probabilities reflect the lower response rate among one-person households, and the lower response rate in urban areas. Households of size five and higher have the highest response rate. The models estimate, surprisingly maybe, that the the probability of response is higher for households of size three than for households of size four. This may be explained by the fact that women often choose to have two children, and that three-person-households mostly consist of mother, father and a *small* child. Such a family will tend to stay at home and thus be more accessible than a typical four-persons-family with two older children.

The higher estimated response rate for households of size three compared to size four is equivalent to the ratio $P(Y=3|R=1)/P(Y=3|R=0)$ being greater than the ratio $P(Y=4|R=1)/P(Y=4|R=0)$. This is consistent with the household distribution in table 2, where we estimate that $P(Y=4)\approx P(Y=4|R=1)$, *i.e.*, $P(Y=4|R=0)\approx P(Y=4|R=1)$. On the other hand, the estimates in table 2 indicate that $P(Y=3|R=1)>P(Y=3)$ which means that $P(Y=3|R=1)>P(Y=3|R=0)$.

We see that the logistic model RM2 combined with the population model with the restricted parametric link $p_{y,x}$ acts as a smoother of the estimates based on the saturated model in (4.9), because of the added assumption of parallel logits of the response probabilities for urban and rural areas.

4. Estimators for Household Size Totals

In this section we present the estimators for household size totals and the method for variance estimation. We use a maximum likelihood estimator with the restricted parametric link function in (3.1) as population model. It is shown that this estimator is identical to an imputation-based poststratified estimator, which again turns out as a standard poststratification when the response mechanism is ignored. Furthermore, we present an imputed poststratified estimator, based on a saturated model for household size and response probability.

4.1 Estimators Based on a Restricted Parametric Link Function as Population Model

With N_y denoting the total number of persons living in households of size y, the number of households of size y equals $H_y = N_y/y$. The total number of households is denoted by H, $H=\sum_y H_y$.

The statistical problem is to estimate H_y for $y=1, ..., J$ and H. The largest size J is chosen such that there are few households of size greater than J. Strictly speaking, H_J is the number of households of size J or more, and likewise for N_J. In our application we choose $J=5$ due to the low frequency in the sample of households of size greater than five. We can write $N_y=\sum_{i=1}^{N} I(Y_i=y)$, where the indicator function $I(Y_i=y)=1$ if $Y_i=y$, and 0 otherwise. Hence, with $\mathbf{x}=(\mathbf{x}_1, ..., \mathbf{x}_N)$,

$$E(H_y|\mathbf{x})=\frac{1}{y}\sum_{i=1}^{N} P(Y_i=y|\mathbf{x}_i).$$

A maximum likelihood based estimator for H_y can be obtained by estimating $E(H_y|\mathbf{x})$, *i.e.*, replacing $P(Y_i=y|\mathbf{x}_i)$ by the maximum likelihood estimator

$\hat{P}(Y_i = y \mid \mathbf{x}_i)$. The data is stratified according to family sizes 1, ..., K, where the last category contains persons belonging to families of sizes $\geq K$. Using the model with the restricted parametric link function, defined in (3.1), Y is assumed to depend only on the family size x, and the estimator takes the form

$$\hat{H}_y = \frac{1}{y}\sum_{x=1}^{K} M_x \hat{P}(Y = y \mid x) \qquad (4.1)$$

where M_x (M_K) denotes the number of persons in the population with registered family size $x(\geq K)$. The M_x's are known auxiliary information from the Norwegian Family Register.

A common approach to correct for nonresponse is by imputation of the missing values in the sample. Based on the estimated distribution for Y for a given family size and place of residence for the nonrespondents, $\hat{P}(Y = y \mid x, z, r = 0)$, we assign the nonrespondents to the values 1, ..., 5 in proportions given by $\hat{P}(Y = y \mid x, z, r = 0)$ for $y = 1, ..., 5$. Let $n^*_{xy}(0)$ ($n^*_{xy}(1)$) be the number of imputed values with family size x and household size y, for rural (urban) areas and let $m_{xu}(0)$ ($m_{xu}(1)$) be the number of missing observations for persons in rural (urban) areas with family size x. Then

$$n^*_{xy}(z) = m_{xu}(z) \cdot \hat{P}(Y = y \mid x, z, r = 0), z = 0, 1. \quad (4.2)$$

and

$$n^*_{xy} = n^*_{xy}(0) + n^*_{xy}(1)$$

is the total number of imputed values with family size x and household size y, *i.e.*, n^*_{xy} is the estimated expected number of households of size y, *given* family size x and $r = 0$.

The following general result holds, showing that with population model (3.1), the maximum likelihood estimator (4.1) is identical to an imputation-based poststratified estimator.

Theorem. Assume model (3.1) for Y. That is, $P(Y = y \mid x, z) = p_{y,x}$ is independent of z, but otherwise the $p_{y,x}$'s are completely unknown with the only restriction $\sum_y p_{y,x} = 1$, for all values of x. The response mechanism is arbitrarily parametrized, *i.e.*, no assumption is made about $P(R = 1 \mid Y = y, x, z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by, for $x = 1, ..., K$,

$$\hat{p}_{y,x} = \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}},$$

where n_{xy} is the number of respondents belonging to a family of size x and household size y, m_x (m_K) is the number of respondents belonging to families of size $x(\geq K)$, and $m_{xu} = m_{xu}(0) + m_{xu}(1)$.

Proof. See Appendix A2.

The theorem implies that the estimator can be written as the imputation-based poststratified estimator, using family size as the stratifying variable ,

$$\hat{H}^{I}_{y,\text{post}} = \frac{1}{y}\sum_{x=1}^{K} M_x \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}. \qquad (4.3)$$

Assuming ignorable response mechanism and using the model (3.1), the likelihood function is given by $\prod_{i=1}^{n_r} P(Y_i = y_i \mid x_i)$. Then the maximum likelihood estimate $\hat{P}(Y = y \mid x)$ is simply the observed rate among the respondents with household size y, given family size x. Thus the maximum likelihood estimator turns out to be identical to the standard poststratified estimator, with family size as the stratifying variable,

$$\hat{H}_{y,\text{post}} = \frac{1}{y}\sum_{x=1}^{K} M_x \frac{n_{xy}}{m_x}. \qquad (4.4)$$

For a general study of poststratification see, for example Holt and Smith (1979) and Särndal, Swensson and Wretman (1992, chapter 7.6).

To illustrate the effects of nonresponse modeling and poststratification, we also present estimates based on the regular expansion estimator, given by

$$\hat{H}_{y,e} = \frac{1}{y} \cdot N \frac{n_y}{n_r} \qquad (4.5)$$

and the imputation-based expansion estimator given by

$$\hat{H}^{I}_{y,e} = \frac{1}{y} \cdot N \frac{n_y + n^*_y}{n}. \qquad (4.6)$$

Here, n_y is the number of respondents in households of size y, n_r is the total number of respondents, and $n^*_y = \sum_x n^*_{xy}$. The estimator (4.5) does not seek to correct for nonresponse nor use the family population distribution as a post-stratifying tool to improve the estimation, while estimator (4.6) tries to take the response mechanism into account, but cannot correct for nonrepresentative samples.

4.2 Imputation-based Poststratification with a Saturated Model

We now proceed to an intuitive method of imputation that was used to estimate response probabilities for a modified Horvitz-Thompson estimator in the official statistics from the 1992 CES (described in Belsby 1995). We will use this imputation method for the poststratified estimator (4.3).

The imputation method consists of distributing, within rural/urban area, the $m_{xu}(z)$ nonresponse units over the household sizes 1, ..., 5 in such a way that, given

household size, the rate of nonresponse is the same for all family sizes. It implicitly assumes that the response probability for persons with the same household size within rural/urban area is identical for different family sizes. Denote the number of nonresponse persons with family size x and household size y and place of residence z obtained in this manner by $h_{xy}(z)$. The corresponding number among the respondents is $n_{xy}(z)$. The values of $h_{xy}(z)$ are determined by the equations

$$\frac{h_{xy}(z)}{h_{xy}(z)+n_{xy}(z)} = \frac{h_{iy}(z)}{h_{iy}(z)+n_{iy}(z)}, \quad z=0,1. \qquad (4.7)$$

When $n_{xy}(z)=0$, we let $h_{xy}(z)=0$. The equation (4.7) is solved under the conditions

$$\sum_y h_{xy}(z) = m_{xu}(z);\ x=1,2,3,4,5 \text{ and } z=0,1. \qquad (4.8)$$

Solving (4.7) and (4.8) requires, for each value of z, one row $(n_{x1}(z), n_{x2}(z), ..., n_{x5}(z))$ of nonzeros, which holds for our case. The imputed values $h_{xy}(z)$ determined by (4.7) and (4.8) correspond to the imputation method described by (4.2) for the following model:

$$P(Y=y \mid x,z) = p_{y,x,z} \text{ with no restrictions} \qquad (4.9a)$$

$$P(R=1 \mid Y=y, x, z) = q_{y,z}, \text{ independent of } x. \qquad (4.9b)$$

This can be seen as follows:

For the ten multinomial trials determined by the different $(x,z)-$ values, we have 50 unknown cell probabilities $\pi_{yx,z} = P(Y=y,\ R=1 \mid x,\ z)$. With no restrictions on cell probabilities, the maximum likelihood estimates (mle) are given by observed relative frequencies,

$$\hat{\pi}_{yx,z} = \frac{n_{xy}(z)}{m_x(z)+m_{xu}(z)}.$$

This also holds when $n_{xy}(z)=0$. Now, it can be shown that there is a one-to-one correspondence between $\boldsymbol{\pi} = (\boldsymbol{\pi}_0,\ \boldsymbol{\pi}_1)$ and $(\boldsymbol{p}_0,\ \boldsymbol{q}_0,\ \boldsymbol{p}_1,\ \boldsymbol{q}_1)$, where $\boldsymbol{\pi}_z = (\pi_{yx,z} : y = 1, ..., 5\ ;\ x = 1, ..., 5)$, $\boldsymbol{p}_z = (p_{yx,z} : y = 1, ..., 5\ ;\ x = 1,\ ..., 5)$ and $\boldsymbol{q}_z = (q_{1,z}, ..., q_{5,z})$. Since $\pi_{yx,z} = p_{y,x,z} \cdot q_{yz}$, the mle of $p_{yx,z}$ and $q_{y,z}$ must satisfy

$$\hat{p}_{yx,z} \cdot \hat{q}_{y,z} = \frac{n_{xy}(z)}{m_x(z)+m_{xu}(z)} \qquad (4.10)$$

and are uniquely determined by $\hat{\pi}_{yx,z}$.

Consider $h_{xy}(z)$, given by (4.5) & (4.6). Let $h_y(z) = \sum_x h_{xy}(z)$ and $n_y(z) = \sum_x n_{xy}(z)$. From (4.7),

$$\frac{h_j(z)}{h_j(z)+n_j(z)} = \frac{h_{xj}(z)}{h_{xj}(z)+n_{xj}(z)}, \text{ if } n_{xj}(z)>0. \qquad (4.11)$$

From (4.10) and (4.11) we have that the following intuitive estimates also are mle.

$$\hat{q}_{y,z} = \frac{n_y(z)}{n_y(z)+h_y(z)} \qquad (4.12)$$

$$\hat{p}_{y,x,z} = \frac{n_{xy}(z)+h_{xy}(z)}{m_x(z)+m_{xu}(z)} \qquad (4.13)$$

(also when $n_{xy}(z) = h_{xy}(z) = 0$).

(We can also show (4.12) and (4.13) by maximizing the loglikelihood directly.) Next, we show that the imputed values (4.2) for the model (4.9) equal $h_{xy}(z)$. From (4.2), we have $n^*_{xy}(z) = m_{xu}(z) \cdot \hat{P}(Y=y \mid x, z, r=0)$. Under model (4.9) and estimates (4.12) and (4.13), we find that

$$\hat{P}(Y=y \mid x,z,R=0) = \frac{\hat{P}(Y=y \mid x,z) - \hat{P}(Y=y, R=1 \mid x,z)}{\hat{P}(R=0 \mid x,z)}$$

$$= \frac{\hat{p}_{yx,z} - \hat{\pi}_{yx,z}}{1-\sum_y \hat{\pi}_{yx,z}}$$

$$= \frac{n_{xy}(z)+h_{xy}(z)-n_{xy}(z)}{m_{xu}(z)} = \frac{h_{xy}(z)}{m_{xu}(z)},$$

and it follows that $n^*_{xy}(z) = h_{xy}(z)$. If $n_{xy}(z)=0$, then $\hat{p}_{y,x,z} = \hat{\pi}_{yx,z} = 0$, and $n^*_{xy}(z) = 0$. We note that model (4.9) is saturated and will, from (4.10), give perfect fit.

The imputation-based expansion estimates (4.6), with model (4.9), are identical to the modified Horvitz-Thompson estimates with $\hat{q}_{y,z} = n_y(z)/[n_y(z)+n^*_y(z)]$ (from (4.12)) as the estimated response probabilities, used in the official statistics from the 1992 CES. This follows from the fact that the modified Horvitz-Thompson estimator of N_y is given by

$$\hat{N}_{y,\mathrm{HT}} = \sum_{i \in s_r} \frac{I(Y_i=y)}{\pi_i},$$

where $\pi_i = P($ person i is selected to the sample and responds). Hence,

$$\pi_i = \frac{n}{N}\hat{P}(R_i=1 \mid x_i, z_i, Y_i=y) = \frac{n}{N}\hat{q}_{y,z_i}$$

and

$$\hat{N}_{y,\mathrm{HT}} = \frac{N}{n}\left(\frac{n_y(0)}{\hat{q}_{y,0}} + \frac{n_y(1)}{\hat{q}_{y,1}}\right). \qquad (4.14)$$

Here,

$$\hat{N}_{y,\mathrm{HT}} = \frac{N}{n}\left(\frac{n_y(0)}{n_y(0)/(n_y(0)+n^*_y(0))} + \frac{n_y(1)}{n_y(1)/(n_y(1)+n^*_y(1))}\right)$$

$$= N\frac{n_y+n^*_y}{n}.$$

So this modified Horvitz-Thompson estimator suffers from the same negative feature as the imputation-based expansion estimator (4.6); it cannot correct for the bias in an unrepresentative sample. For a general description of the modified Horvitz-Thompson method see, *e.g.,* Särndal *et al.* (1992, chapter 15).

4.3 Variance Estimation

Variance estimation of the various estimates are obtained by bootstrapping. It can be carried out under the modeling or quasi-randomization framework (Little and Rubin 1987). For instance, to estimate the variance under model (3.1) and RM1 (3.2), we may apply the parametric bootstrap with the estimated parameters (Efron and Tibshirani 1993). However, it is not clear how to compare the variances estimated under the alternative models. We have therefore chosen to estimate the variances of the different estimators under a common quasi-randomization framework. We assume simple random sampling conditional to the family size, which is the only assumption we make for variance estimation. Unconditionally we have a self-weighting, but not simple random, sample, and therefore this is a rather crude approximation to the actual conditional sampling design. However, for a comparative study of the estimators the approximation will serve this purpose well. The nonresponse indicator r_i is considered to be a constant associated with person i. We draw the bootstrap sample, resampling $(y_i, z_i, r_i = 1)$, $(z_i, r_i = 0)$ randomly with replacement, as described by Shao and Sitter (1996, Section 5), within each post-stratum of $\{i; x_i = x\}$. While the sizes of the sample post-strata are fixed, both the number of nonrespondents and the number of persons from urban or rural areas vary from one bootstrap sample to another. We calculate the bootstrap estimates in the same way as based on the observed data. In particular, the bootstrap data are imputed in the same way as the original data if the estimator is imputation-based. Finally, the estimated variances and standard errors are obtained by the usual Monte Carlo approximation based on 500 independent bootstrap samples.

5. Estimated Number of Households of Different Sizes Based on the 1992 Norwegian Consumer Expenditure Survey

In this section we present the estimated number of households of sizes one to five and more, and the total number of households for the population in Norway aged less than eighty years old. The estimation uses the data from CES 1992, and is based on the estimators considered in Section 4. To compute the estimates we need the number of families of different sizes in the population, *i.e.,* M_x, at the time of the 1992 survey. The actual number at the time of the survey is not recorded. As an approximation we use the numbers at January 1, 1993. These are given in table 4.

Table 4
Families and Persons with Age Less than 80 Years in Norway at January, 1993

Number of persons in family	Families	Persons
1 person	793,869	793,869
2 persons	408,440	816,880
3 persons	261,527	784,581
4 persons	266,504	1,066,016
5 or more persons	127,653	670,528
Total	1,857,993	4,131,874

Note that the average family size for families with 5 or more persons is 670,528/127,653 = 5.25. We use 5.25 as an estimate of the average household size for households of size 5 or more, and divide by 5.25 instead of 5 in all estimates of H_5.

5.1 Maximum Likelihood Estimation and Poststratification

The estimated household distributions are presented in table 5. The estimates are based on the maximum likelihood (m.l.) estimator (4.1) using the population model with the restricted parametric link function $p_{y,x}$ in combination with the response models RM1(y, z) and RM2(y, z). To illustrate the effect of nonresponse modeling versus poststratification we also present the standard poststratified estimator (4.4). We recall that this is the maximum likelyhood estimator when ignoring the response mechanism. Furthermore, we present the estimated household size distribution based on the imputation-based poststratification (4.3) with the saturated model (4.9). For assessing the sampling variability of the different estimators, the estimated standard errors are also included.

The three models that take the response mechanism into account give higher total number of households. They also give considerable higher numbers of one-person-households. This seems sensible since we expect the one-person households to have the highest nonresponse rate. And thus, these estimates are most influenced by taking the response mechanism into account. We note that the restricted parametric link model (3.1) together with the logistic response model RM2(y, z) gives practically the same poststratified estimates as model (4.9), with also approximately the same standard errors. Because of the freedom of model (4.9), with perfect fit, it seems that model (3.1) & RM2(y, z) works well for estimating the number of households of different sizes. Regarding the uncertainty of the estimates, we see as one might expect that the standard errors typically seem to increase with the number of unknown parameters in the underlying model. Also, the total number of households is rather accurately estimated, not counting possible bias, while it's clearly most difficult to estimate the number of one-person households.

In order to evaluate the extent to which the differences between the estimates are due to sampling error or non-response bias, we consider the estimated standard errors of the differences of the point estimates. Some of these are given in table 6, using mostly the imputation-based post-stratification with the saturated model as a reference. For short, we use the terms Est1 – Est4 for the estimates defined as they appear in table 5:

Est1: M.l. estimator based on population model $p_{y,x}$ and response model RM1

Est2: M.l. estimator based on population model $p_{y,x}$ and response model RM2

Est3: Imputation-based poststratification based on the saturated model (4.9)

Est4: Poststratified estimator without imputation.

Based on tables 5 and 6 we can conclude that Est4 and Est3 have different expected values in estimating H_1, H_3, H_5 and H. Regarding the other comparisons, we see that in estimating H_3 there is a significant difference between Est1 and Est2/Est3, and note from earlier discussions in Section 3.3 that RM2 gives a better fit to the data than RM1.

The estimates based on the expansion estimator $\hat{H}_{y,e}$, given by (4.5), in 100's, are 390,500, 496,500, 283,900, 279,900, 148,000 and 1,598,800 with estimated standard errors equal to 33,100, 21,700, 14,600, 11,600, 6,100 and 23,700 for H_1, ..., H_5 and H, respectively. The standard errors for the differences between these estimates and the Est3-estimates are 52,800, 30,900, 19,100, 10,800, 5,400 and 32,000 for H_1, ..., H_5 and H respectively. These expansion estimates indicate serious bias due to non-response, especially the estimates for H_1, H_5 and H, with poststratification correcting for some of the bias (probably about 50% for the estimates of H_1 and H). We also note that the standard errors for the poststratified estimator and this simple expansion estimator are about the same. So by reducing the bias with poststratification one reduces the total error as well.

Poststratification corrects for the bias caused by the discrepancy between the family size distributions in the response sample and the population. From table 1 and table 4 we see that these family size distributions are given by (in percentages), for $x = 1, ..., 5$:

Response sample: 14.6 – 20.7 – 19.1 – 27.0 – 18.6
Population: 19.2 – 19.8 – 19.0 – 25.8 – 16.2.

Since the number of one-person families is much too low in the response sample, so will the expansion estimate of H_1 be. With post strata determined by family size, post-stratification corrects for the family size bias in the response sample, but does implicitly assume that nonrespondents and respondents have the same household size distribution, for a fixed family size. Or, in other words, the respondents are treated as a random subsample of sampled units with the same family size, as mentioned by Little (1993). This is most likely not the case. We recall that the family size variable was not significant when the household variable was included in the response models. Thus it seems reasonable to assume, as in our response models, that response rates will vary with the actual household sizes rather than the registered family sizes. Typically, estimates of the number of one-person households will be biased when the nonrespondents are ignored.

Table 5
Estimated Household Totals for Persons Aged Less than 80 Years in Norway at January 1, 1993, in Units of 100.
In Parentheses, the Estimated Standard Error of the Estimates

	Maximum likelihood estimator with nonignorable response mechanism				**Imputation-based poststratification**		**Ignoring the response mechanism**	
Household size, y	Population model $p_{y,x}$ and response model RM1 (y, z)	%	Population model $p_{y,x}$ and response model RM2 (y, z)	%	Saturated population and response model	%	Poststratified estimator	%
1	558,800 (38,900)	32	595,400 (48,000)	34	596,600 (53,500)	34	486,000 (35,800)	29
2	520,200 (20,600)	30	525,800 (27,400)	30	523,600 (29,800)	30	507,800 (20,000)	30
3	278,900 (13,800)	16	249,100 (20,300)	14	250,000 (19,800)	14	286,200 (14,100)	17
4	258,900 (9,800)	15	269,000 (11,600)	15	268,900 (11,500)	15	270,600 (10,100)	16
≥ 5	125,800 (4,700)	7	126,000 (5,100)	7	126,200 (5,000)	7	131,300 (4,700)	8
Total	1,742,600 (25,600)	100	1,765,300 (29,700)	100	1,765,300 (31,900)	100	1,681,900 (23,300)	100

Table 6
Estimated Standard Errors of the Differences of the Point Estimates in Table 5

Household size	Est1−Est2	Est1−Est3	Est2−Est3	Est4−Est3
1	29,700	37,000	16,600	42,400
2	19,300	22,200	8,800	23,100
3	15,400	15,200	5,300	15,500
4	6,700	6,500	1,800	6,600
≥ 5	1,700	1,700	500	1,900
Total	15,300	18,800	8,900	23,300

After having corrected for nonresponse bias by completing the sample with imputed values, the sample itself may be skewed compared to the population. To illustrate the effect of poststratification to correct for this, we shall compare, using the saturated model (4.9), the imputation-based poststratified estimates Est3 with the imputation-based expansion estimates given by (4.6): 583,900, 567,700, 244,300, 259,300, 122,400 and 1,777,600 for H_1, ..., H_5 and H, respectively. As noted in Section 4.2, see (4.14), these estimates are identical to the modified Horvitz-Thompson estimates. The standard errors for these estimates are practically the same as for Est3. Hence, the alternative poststratified estimation methods based on nonignorable response models have standard errors at least no worse than the modified Horvitz-Thompson estimator. So if one reduces the bias with the alternative methods, one reduces the total error too. The standard errors of the differences between Est3 and this modified Horvitz-Thompson estimator in the estimates of H_1, ..., H_5 and H are 3,500, 2,200, 1,100, 600, 200 and 2,100 respectively. Clearly these two methods give significantly different estimates for all household size totals. In this comparison, one feature stands out. The expansion estimate of the number of two-persons households, 567,700, is clearly too high, as seen by comparing the family size distributions in the total sample and the population (in percentages), for $x = 1, \ldots, 5$:

Population: 19.2 – 19.8 – 19.0 – 25.8 – 16.2
Sample: 18.6 – 23.0 – 17.8 – 24.9 – 15.7.

The sample proportion of persons in two-persons families is much too high, and even though we have corrected for nonresponse bias, the expansion estimator, and then also the modified Horvitz-Thompson estimator cannot correct for a nonrepresentative sample. This will necessarily lead to biased estimates of H_2. We need poststratification to correct for a skewed sample. One can regard the difference in expected values for these estimators of H_2 as being close to the bias for the modified Horvitz-Thompson estimator, and note that an approximate 95% confidence interval for this difference is (39,800, 48,400).

For robustness considerations we also present the estimates from the cumulative logit model mentioned in Section 3.1 together with RM1 (y, z), which we know fits the data poorly. They are, in 100's: 591,800, 501,000, 265,200, 267,300, 128,200 and 1,753,500 for H_1, ..., H_5 and H, respectively. Compared to table 5, this seems to indicate that a reasonable model for response plays a more important role than a good population model. It is also evident that nonresponse modeling makes a difference, as seen when compared to poststratification and simple expansion.

5.2 Comparison with the Currently Used Estimates in CES, the Quality Survey for the 1990 Census and a Projection Study

Since 1993, an alternative, computationally simpler, modified Horvitz-Thompson estimator of type (4.14) has been in use in the production of official statistics from CES, see (Belsby 1995). We recall from Section 2 that the weights are the inverse sampling probabilities of the households, multiplied with the estimated probability of response. The response probabilities are estimated using a logistic model similar to RM2 (y, z) with place of residence and household size as explanatory variables. For the nonrespondents with unknown household size the registered family size is used instead, replacing (3.5). Thus, the weights may be regarded as an approximation to using (3.5). Of course, (3.5) is possible only when a population model is considered, which CES has not done. Table 7 presents estimated household distribution based on this CES-modified Horvitz-Thompson estimator.

The quality survey for the Census 1990, PES *1990*, contains 8,280 respondents and uses practically the same household definition as CES. The response rate was 95%. The H_y – estimates uses poststratification with respect to household size in the Census. However, no attempts were made to correct for possible nonresponse bias with respect to actual household size. PES deals with the whole population. Table 7 has the estimates for the $0 - 79$ age group with the same poststratification method as in *PES*.

Table 7 also presents estimates based on the Household Projections study by Keilman and Brunborg (1995). This study simulates household structure for the period 1990 to 2020. The data sources are 28,384 individuals from the 1990 Population and Housing Census and 1988 Family and Occupation Survey. Keilman and Brunborg project for the whole population in 1992. We adjust their estimates to the $0 - 79$ age group.

Table 7
Estimated Household Size Totals for Persons Less than 80 Years in Norway at January 1, 1993 with CES-modified Horvitz-Thompson, PES 1990 and Projections, in Units of 100

Household size	CES-Modified Horvitz-Thompson	%	PES 1990	%	Projections	%
1	622,900	35	626,000	35	668,300	37
2	518,500	29	494,200	28	549,000	30
3	259,900	15	291,500	16	211,900	12
4	258,500	15	250,000	14	221,500	12
≥ 5	124,600	7	115,300	6	97,500	5
Unknown					78,500	4
Total	1,784,400	1	1,777,000	99	1,826,700	100

Table 8
Estimated Probability of Response Based on the Method Used in CES Since 1993, in Percentages

	Household size				
Place of residence	1	2	3	4	5 or more
	CES-method				
Rural	44.53	66.24	74.55	73.54	80.07
Urban	36.01	57.90	67.25	66.09	73.80
	Model $p_{y,x}$ in (3.1) combined with RM2(y, z)				
Rural	47.77	60.90	79.05	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46

The estimates in table 7 support our impression that the estimates based on modeling the response mechanism leads to less biased estimates compared with ignoring the response mechanism as in mere poststratification or simple expansion. This is especially true for the one-person households and the total. The current "official estimator", the modified Horvitz-Thompson seems to give estimates of the right magnitude and in fact is closer to the results of PES 1990 than the modelbased estimates. However, this is more by accident. As a *method* it has some problems even in a representative sample. We can study this by estimating the response probabilities. Table 8 presents the results together with the estimates based on RM2(y, z) & (3.1) from table 3.

Compared to the estimated response probabilities based on model RM2(y, z) with (3.1), we see that replacing household size with family size in the nonresponse group is not a satisfactory approximation. Hence, if compared with the modified Horvitz-Thompson estimator in Section 5.1 based on the saturated model (4.9), the latter one would be preferred. For this particular survey, the CES approach overestimates the probability of response for household of size 2, which in a representative sample would lead to underestimating of H_2. The estimated response probabilities will most likely be biased when we are using family size in place of household size in the nonresponse group when estimating the parameters in the response model. This bias is an additional problem to the previously mentioned one, that the modified Horvitz-Thompson estimates will be similar to the imputation-based expansion estimates and cannot correct for nonrepresentative samples (as has been a problem in CES since 1993). In the 1992 CES, however, the sample is skewed with a too high proportion of families of size 2, and the H_2 – estimate will be of the right magnitude, by accident.

6. Conclusions

We have investigated modeling and methodological issues for estimating the total number of households of different sizes in Norway, based on the Norwegian Consumer Expenditure Survey (CES). The main issue is how to correct for bias due to nonignorable nonresponse. The existing estimation method in CES is a modified Horvitz-Thompson estimator that includes a correction for nonresponse by estimating response probabilities. We have considered basically two modelbased approaches, a maximum-likelihood estimator and imputation-based poststratification after registered family size. With a population model that corresponds to a group model after family size only, these two estimators are identical. This family group model for household size and a logistic link for the response probability using household size as a categorical variable seem to work well for our estimation problem.

In analyzing the 1992 CES, we find serious bias due to nonresponse, especially the estimates for H_1 and H, with pure poststratification (without imputation) correcting for

some of the bias (probably about 50% for the estimates of H_1 and H). Poststratification does not, however, take into account possible nonresponse bias dependent on household size. Our response models assume that the response rates will vary with the actual household sizes rather than the registered the family sizes, and it is quite evident that such nonresponse modeling makes a difference, leading to less biased estimates than mere poststratification or simple expansion, especially of H_1 and H.

The modified Horvitz-Thompson estimates used in the official statistics from CES correspond to imputation-based expansion estimates. Hence, they cannot correct for nonrepresentative samples. The study in this paper shows that, in addition to a nonignorable response model it is also necessary to poststratify according to family size, *i.e.*, using a population model given family size. Hence poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

In any estimation problem of totals in survey sampling, one must be aware of the fact that a Horvitz-Thompson estimator cannot correct for skewed samples, even when modified with good response estimates. Poststratification should always be considered as well as imputation based on a response model, nonignorable when needed.

Appendix A1

The data for rural and urban areas separately are given in table A1.

Appendix A2

Theorem. Assume model (3.1) for Y. *i.e.*, $P(Y=y\,|\,x,\ z)=p_{y,x}$ is independent of z, but otherwise the $p_{y,x}$'s are completely unknown with the only restriction being that $\sum_y p_{y,x}=1$, for all values of x, for all k. The response mechanism is arbitrarily parametrized, *i.e.*, no assumption is made about $P(R=1|Y=y,\ x,\ z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by

$$\hat{p}_{y,x}=\frac{n_{xy}+n^{*}_{xy}}{m_x+m_{xu}}.$$

Proof. Let $q_{yx,z}=P(R=1|Y=y,\ x,\ z)$. The log likelihood is given by

$$\begin{aligned}\ell &= \sum_x\sum_y n_{xy}\,p_{y,x}+\sum_{z=0}^{1}\sum_x\sum_y n_{xy}(z)q_{yx,z}\\ &\quad+\sum_{z=0}^{1}\sum_x m_{xu}(z)\log P(R=0\,|\,x,z)\\ &=\sum_x\sum_y n_{xy}\,p_{y,x}+\sum_{z=0}^{1}\sum_x\sum_y n_{xy}(z)q_{yx,z}\\ &\quad+\sum_{z=0}^{1}\sum_x m_{xu}(z)\log(1-\sum_{y=1}^{5}p_{y,x}q_{yx,z}).\end{aligned}$$

We use the Lagrange method and maximize $G=\ell+\sum_{x=1}^{5}\lambda_x(\sum_{y=1}^{5}p_{y,x}-1)$.

Let the solutions be $\hat{p}_{y,x}(\lambda_x)$, and determine the λ_x's such that $\sum_y\hat{p}_{y,x}(\lambda_x)=1$, for all x. No matter how the $q_{yx,z}$'s are parametrized, the mle $\hat{p}_{y,x}$ must satisfy, by solving the equations $\partial G/\partial p_{y,x}=0$,

$$\frac{n_{xy}}{\hat{p}_{y,x}}-\sum_{z=0}^{1}m_{xu}(z)\frac{\hat{q}_{yx,z}}{\hat{P}(R=0\,|\,x,z)}+\lambda_x=0 \qquad \text{(A1)}$$

which is equivalent to:

$$\begin{aligned}n_{xy}&=\hat{p}_{y,x}\sum_{z=0}^{1}\frac{m_{xu}(z)}{\hat{P}(R=0\,|\,x,z)}\\ &\quad-\sum_{z=0}^{1}m_{xu}(z)\frac{\hat{P}(R=0,Y=y\,|\,x,z)}{\hat{P}(R=0\,|\,x,z)}-\hat{p}_{y,x}\lambda_x.\end{aligned}$$

Table A1
Family and Household Sizes for the 1992 Norwegian Consumer Expenditure Survey, Split into Rural and Urban Areas. The Upper Entry is for the Urban Group

		Household size								
Family size		1	2	3	4	≥ 5	Total response	Non-response	Total	Response rate
1	urban	28	24	7	2	0	61	78	139	0.439
	rural	55	24	13	7	2	101	75	176	0.574
2	urban	6	70	12	3	0	91	84	175	0.520
	rural	3	107	25	1	3	139	76	215	0.647
3	urban	4	8	57	11	3	83	40	123	0.675
	rural	6	17	74	29	3	129	51	180	0.717
4	urban	0	3	15	80	5	103	43	146	0.705
	rural	2	10	22	151	12	197	80	277	0.711
≥ 5	urban	0	1	0	6	66	73	28	101	0.723
	rural	1	3	4	11	115	134	32	166	0.807
Total urban		38	106	91	102	74	411	273	684	0.601
Total rural		67	161	138	199	135	700	314	1014	0.690

We determine λ_x by summing over y:

$$m_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0 \mid x, z)} - \sum_{z=0}^{1} m_{xu}(z) \frac{\hat{P}(R=0 \mid x, z)}{\hat{P}(R=0 \mid x, z)} - \lambda_x,$$

hence

$$\lambda_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0 \mid x, z)} - (m_x + m_{xu}).$$

It follows from (A1) that $\hat{p}_{y,x}$ satisfies the following relation:

$$\hat{p}_{y,x} = \frac{n_{xy}}{\left(m_x + m_{xu} - \sum_{z=0}^{1} m_{xu}(z) \frac{\hat{P}(R=0 \mid Y=y, x, z)}{\hat{P}(R=0 \mid x, z)} \right)}. \quad \text{(A2)}$$

The imputed values are given by , from (4.2),

$$n^*_{xy}(z) = m_{xu}(z)\, \hat{p}_{y,x} \frac{\hat{P}(R=0 \mid Y=y, x, z)}{\hat{P}(R=0 \mid x, z)}$$

and, from (A2),

$$\hat{p}_{y,x} = n_{xy} \Big/ \left(m_x + m_{xu} - \sum_{z=0}^{1} \frac{n^*_{xy}(z)}{\hat{p}_{y,x}} \right) = n_{xy} \Big/ \left(m_x + m_{xu} - \frac{n^*_{xy}}{\hat{p}_{y,x}} \right)$$

or equivalently,

$$\hat{p}_{y,x}(m_x + m_{xu}) - n^*_{xy} = n_{xy},$$

$$\textit{i.e.,}\ \ \hat{p}_{y,x} = \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}. \quad \text{Q.E.D}$$

Appendix A3

Table A2

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM1(*y*, *z*)

		Household size					
Family size		1	2	3	4	≥ 5	Total
1	urban	77.8	44.1	12.9	3.9	0.3	139
	rural	103.6	43.1	18.4	8.7	2.3	176
2	urban	10.8	137.9	22.1	3.8	0.4	175
	rural	7.5	168.6	33.9	1.7	3.3	215
3	urban	7.5	14.3	81.3	16.4	3.6	123
	rural	10.7	25.3	104.8	35.6	3.7	180
4	urban	0.8	6.4	21.9	110.3	6.6	146
	rural	3.5	16.7	35.1	206.9	14.8	277
≥ 5	urban	0.5	2.4	1.0	9.0	88.2	101
	rural	1.6	4.7	5.2	14.4	140.1	166
Total /urban		97.4	205.1	139.2	143.4	99.1	684
rural		126.9	258.4	197.4	267.3	164.2	1,014

Table A3

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM2 (*y*, *z*)

		Household size					
Family size		1	2	3	4	≥ 5	Total
1	urban	81.6	42.7	10.4	4.0	0.3	139
	rural	107.5	41.5	15.9	8.8	2.3	176
2	urban	11.9	140.4	18.3	3.9	0.5	175
	rural	8.6	170.9	30.3	1.8	3.4	215
3	urban	9.4	16.1	75.2	18.6	3.7	123
	rural	13.4	27.7	96.5	38.5	3.9	180
4	urban	0.8	6.2	18.9	113.5	6.6	146
	rural	3.7	16.2	29.2	213.1	14.8	277
≥ 5	urban	0.5	2.3	0.6	9.3	88.3	101
	rural	1.7	4.6	4.6	14.9	140.2	166
Total /urban		104.2	207.7	123.4	149.3	99.4	684
rural		134.9	260.9	176.5	277.1	164.6	1,014

Appendix A4

Table A4
The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (4.9), *i.e*, Imputations Determined by (4.7) and (4.8)

	Household size					
Family size	1	2	3	4	≥ 5	Total
1 urban	79.6	47.2	9.4	2.8	0.0	139
rural	108.3	38.5	16.9	9.9	2.4	176
2 urban	17.1	137.7	16.0	4.2	0.0	175
rural	5.9	171.6	32.5	1.4	3.6	215
3 urban	11.4	15.7	76.2	15.6	4.1	123
rural	11.8	27.3	96.2	41.1	3.6	180
4 urban	0.0	5.9	20.0	113.2	6.9	146
rural	3.9	16.0	28.6	214.0	14.5	277
≥ 5 urban	0.0	2.0	0.0	8.5	90.5	101
rural	2.0	4.8	5.2	15.6	138.4	166
Total /urban	108.1	208.5	121.6	144.3	101.5	684
rural	131.9	258.2	179.4	282.0	162.5	1,014

Table A5
The Total Numbers of Family and Household Sizes for Imputed Complete Sample. Based on Model (4.9)

	Household size					
Family size	1	2	3	4	≥ 5	Total
1	187.9	85.7	26.3	12.7	2.4	315
2	23.0	309.2	48.6	5.7	3.6	390
3	23.2	43.0	172.4	56.7	7.7	303
4	3.9	21.9	48.7	327.2	21.3	423
≥ 5	2.0	6.8	5.2	24.1	229.0	267
Total	240.0	466.6	301.1	426.3	264.0	1,698

References

Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Belsby, L. (1995). Forbruksundersøkelsen. Vektmetoder, frafallskorrigering og intervjuer-effekt. (The consumer survey. Weight methods, nonresponse correction and interviewer effect), Notater 95/18 Statistics Norway.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

Bjørnstad, J.F., and Skjold, F. (1992). Interval estimation in the presence of nonresponse. *The American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*. 233-238.

Bjørnstad, J.F., and Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. *The American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*, 152-156.

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

Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Journal of the Royal Statistical Society B*, 60, 57-70.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

Hall, B.H., Cummins, C. and Schnake, R. (1991). *TSP Reference Manual, Version 4.2A*, Palo Alto California: TSP International.

Holt, D., and Smith, T.M.F. (1979). Post-stratification, *Journal of the Royal Statistical Society A*, 142, 33-46.

Keilman, N., and Brunborg, H. (1995). *Household Projections for Norway, 1990-2020, Part 1: Macrosimulation*, Reports 95/21, Statistics Norway.

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

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

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

McCullagh, P., and Nelder, J.A. (1991). *Generalized Linear Models, 2nd ed*. London: Chapman & Hall.

Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Statistics Norway (1990). *Survey of Consumer Expenditure 1986-88.* Official Statistics of Norway NOS B919.

Statistics Norway (1996). *Survey of Consumer Expenditure 1992-1994*. Official Statistics of Norway NOS C317.

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

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 213-225
Statistics Canada, Catalogue No. 12-001-XPB

Bayesian Analysis of Nonignorable Missing Categorical Data: An Application to Bone Mineral Density and Family Income

Balgobin Nandram, Lawrence H. Cox and Jai Won Choi [1]

Abstract

We consider a problem in which an analysis is needed for categorical data from a single two-way table with partial classification (*i.e.*, both item and unit nonresponses). We assume that this is the only information available. A Bayesian methodology permits modeling different patterns of missingness under ignorability and nonignorability assumptions. We construct a nonignorable nonresponse model which is obtained from the ignorable nonresponse model via a model expansion using a data-dependent prior; the nonignorable nonresponse model robustifies the ignorable nonresponse model. A multinomial-Dirichlet model, adjusted for the nonresponse, is used to estimate the cell probabilities, and a Bayes factor is used to test for association. We illustrate our methodology using data on bone mineral density and family income. A sensitivity analysis is used to assess the effects of the data-dependent prior. The ignorable and nonignorable nonresponse models are compared using a simulation study, and there are subtle differences between these models.

Key Words: Bayes factor; Chi-squared statistic; Importance function; Markov chain Monte Carlo; Multinomial-Dirichlet model; Robust; Two-way categorical table.

1. Introduction

It is a common practice to use two-way categorical tables to present survey data. For many surveys there are missing data, and this gives rise to partial classification of the sampled individuals. Thus, for the two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing); see Little and Rubin (2002, section 1.3) for definitions of the three missing data mechanisms (MCAR, MAR, MNAR). Thus, there are four tables (one table with the complete cases, and three possible supplemental tables: one table with row classification only, one table with column classification only, and one table with neither row nor column classification). One may not know how the data are missing. Thus, we use a model in which the likelihood function accounts for differences between the observed data and missing data (*i.e.*, nonignorable missing data); see Rubin (1976) and Little and Rubin (2002) for the relation between ignorability/nonignorability and these three missing data mechanisms. Because there are well-known advantages of the Bayesian method over the non-Bayesian method for these problems, we propose a Bayesian analysis of a general $r \times c$ categorical table, consisting of a table with complete cases and three supplemental tables. Specifically, we develop a Bayesian method to estimate the cell probabilities and test for association between the two categorical variables.

We assume that the only information available to the data analysts is the complete cases and the three supplemental tables. Specifically, we assume that there is no information (either from covariates or prior information) about nonignorability. In our Bayesian approach, the survey design features have been suppressed (*i.e.*, there are no survey weights and there are no clustering or stratification). Sometimes survey data are presented to the public with certain features of the data suppressed for reasons of convenience and confidentiality. We recognize that both the ignorable and the nonignorable nonresponse models may be incorrect when they do not take account of these features. However, the parameters in the ignorable nonresponse model are identifiable and estimable, and one can take advantage of this fact to construct a nonignorable nonresponse model which is related to the ignorable nonresponse model. Also, in the ignorable nonresponse model we assume that there is a MAR mechanism that drives the nonresponse, and there may be information in the incomplete cases (*i.e.*, the two tables with observed row and column margins). Without any information about the degree of nonignorability, it is sensible to generalize the ignorable nonresponse model. This is how we attempt to accomplish our objectives in this paper.

This paper has five sections. In section 1 we have further discussion of the problem, and we review related methodology. In section 2, we describe a 3×3 table of bone mineral density (BMD) and family income (FI) from the third National Health and Nutrition Examination Survey

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute road, Worcester MA 01609, E-mail: balnan@wpi.edu; Lawrence H. Cox and Jai Won Choi, Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. E-mail: lgc9@cdc.gov, jwc7@cdc.gov.

(NHANES III). This is used mainly for illustration. In section 3, we describe the methodology to obtain estimates of the cell probabilities, and we use the Bayes factor to test for association of the two attributes. We accomplish these objectives by first constructing an ignorable nonresponse model, and we show how to expand an ignorable nonresponse model into a nonignorable nonresponse model. In section 4, we analyze the NHANES III data to demonstrate our methods. Also, a simulation study gives further comparison of the ignorable and the nonignorable nonresponse models, and a sensitivity analysis shows that inference is not too sensitive to the choice of an important prior distribution. Finally, section 5 has concluding remarks.

1.1 Discussion of the Problem

We do not know whether an ignorable nonresponse model or a nonignorable nonresponse model is appropriate, but it is worthwhile noting that Cohen and Duffy (2002) point out that "Health surveys are a good example, where it seems plausible that propensity to respond may be related to health." Thus, nonignorable nonresponse models are important candidates for the analysis of data from health surveys. For a general $r \times c$ categorical table (two categorical variables, one with r categories and the other with c categories) with nonresponse, our objectives are to show how to (a) make inference about the cell probabilities, and (b) test for no association between the two categories using the Bayes factor. While (a) comes directly from the modeling, (b) needs one extra step.

Let $\boldsymbol{I}_i$ be the cell indicator for the i^{th} individual in a $r \times c$ table for $i = 1, \ldots, n$ individuals. Then, it is well known that if the $\boldsymbol{I}_i$ are *independent and identically* distributed, the Pearson's chi-squared statistic has $\chi^2_{(r-1)(c-1)}$. Otherwise the Pearson's chi-squared statistic does not have a $\chi^2_{(r-1)(c-1)}$, and this is true when there are missing data and the respondents and nonrespondents differ. When this is the case, adjustments must be made to the Pearson's chi-squared statistic. Within the non-Bayesian framework Chen and Fienberg (1974) and Wang (2001) have corrections for incomplete two-way tables. Although not directly relevant here, it is pertinent to mention that similar adjustments have been made for cluster sampling and stratified random sampling (Rao and Scott 1981, 1984). The works of Chen and Fienberg (1974) and Wang (2001) can essentially handle item nonresponse only; unit nonresponse is excluded because the modeling is motivated by the ignorable nonresponse models (*e.g.*, see discussion in Kalton and Kasprzyk 1986).

The Bayesian method permits us to use a procedure that does not rely on asymptotic theory, incorporate nonignorable missingness into the modeling and obtain an alternative to Pearson's chi-squared statistic for testing for no association; see Little (2003) for a discussion of the well-known advantages of the Bayesian approach in survey sampling. Our alternative to the Pearson chi-squared statistic is based on the Bayes factor (Kass and Raftery 1995). This is a statistic that compares a model with association and one with no association via the ratio of their marginal likelihoods under the ignorable and the nonignorable nonresponse models separately.

Little and Rubin (2002, chapter 15) discuss the nonignorable nonresponse problem. For example, Rubin, Stern and Vehovar (1995) (also discussed in Little and Rubin 2002, page 345) provide an interesting analysis of the November/December 1990 Slovenian Public Opinion survey in which there were data on 2,074 prospective voters in their plebiscite with three dichotomous variables; there is 12% nonresponse. They fit both ignorable and nonignorable nonresponse models (loglinear with all interactions) to the data, and they were satisfied with the ignorable nonresponse model. However, they stated "Of course, this does not mean that MAR should be automatically applied in all cases. Analyses assuming MAR are not likely to be adequate if a survey has large amounts of nonresponse, if covariate information is limited, or for cases where the missing-data mechanism is clearly nonignorable (*e.g.*, censored data)."

1.2 Related Methodology

Our methodology is different from Rubin, Stern and Vehovar (1995). We start with Nandram and Choi (2002 a, b) in which a parameter γ centers (can be viewed as an index) the nonignorable nonresponse model on the ignorable nonresponse model. When $\gamma = 1$, the nonignorable nonresponse model is the ignorable nonresponse model, and thus, the nonignorable nonresponse model "degenerates" into the ignorable nonresponse model when $\gamma = 1$; see also Forster and Smith (1998). This is useful because the nonignorable nonresponse model contains the ignorable nonresponse model as a special case; thereby expressing uncertainty about ignorability. Draper (1995) called this a *continuous model expansion*, and he has recommended the use of a continuous model expansion over a discrete model expansion (*i.e.*, finite mixtures) whenever it is possible. We simply call the continuous model expansion an *expansion* model. Nandram and Choi (2002 a, b) obtain the centering by taking $\gamma \mid \nu \sim \text{Gamma}(\nu, \nu)$ in which $E(\gamma \mid \nu) = 1, \quad \text{var}(\gamma \mid \nu) = 1/\nu$.

Nandram and Choi (2002 a) analyze binary data on household crimes in the National Crime Survey, and Nandram and Choi (2002 b) analyze binary data on doctor visits in the National Health Interview Survey. While Nandram and Choi (2002 a) has more comparisons, Nandram and Choi (2002 b) has more sensitivity analyses. Nandram, Han and Choi (2002) describe two hierarchical

Bayesian models, an ignorable and a nonignorable nonresponse model, for the analysis of count data from several areas, the counts in each area being described by a multinomial distribution. In all these works the issue of association is not relevant because there is a single categorical variable.

The approach in Nandram and Choi (2002 a, b) is attractive, but it does not apply immediately to the current application on $r \times c$ categorical table. Specifically, only one centering parameter was needed in Nandram and Choi (2002 a, b). To extend the method of Nandram and Choi (2002 a, b), one needs rc centering parameters. Each of these parameters has to have a distribution centered at one to allow degeneration to the ignorable nonresponse model. There are also inequality constraints that must be included in the nonignorable nonresponse model. Thus, while this idea is attractive, the methodology needed to apply the work of Nandram and Choi (2002 a, b) is much beyond the scope of our current paper.

Nandram, Liu, Choi and Cox (2005) extend the work of Nandram, Han and Choi (2002) in two important directions to (a) consider several two-way categorical tables instead of one-way tables and (b) develop a method to study the association between the two categorical variables. Nandram, Liu, Choi and Cox (2005) analyze data on the relation between bone mineral density (BMD) and age from thirty-five counties in the third National Health and Nutrition Examination Survey. In each county the data are categorized into two levels of age and three levels of BMD (*i.e.*, there are thirty-five 2×3 categorical tables). Note that the age of everyone is observed, but the BMD values for a large number of individuals are not observed. Thus, for each county there is a single table with complete cases, and one table with row totals (*i.e.*, the ages of these individuals are known, but their BMD values are missing). Here, our objective is to extend the work of Nandram, Liu, Choi and Cox (2005) to a *general* $r \times c$ categorical table. This is an important advance because now there are three supplemental tables (one table with row classification only, one table with column classification only, and one table with neither row nor column classification) instead of just one with row totals as in Nandram, Liu, Choi and Cox (2005).

2. Data on Bone Mineral Density and Family Income

We briefly describe the 3×3 categorical table of bone mineral density (BMD) and family income (FI). FI is a discrete variable, and there are three levels: low, medium and high. While BMD is a continuous variable, the World Health Organization has classified BMD into three levels: normal, osteopenia and osteoporosis; see Looker, Orwoll, Johnston, Lindsay, Wahner, Dunn, Calvo and Harris (1997, 1998). BMD is used to diagnose osteoporosis, a disease of elderly females, and in NHANES III it is measured for individuals at least twenty years old (*i.e.*, we use the data on white females only with chronic conditions older than twenty years).

Among those participated in the examination stage, about 62% of the individuals have both FI and BMD observed, 8% with only BMD observed, 29% with only income observed, 1% with neither income nor BMD. The dataset, used in our study, is presented in Table 1 as a 3×3 categorical table of BMD and FI. Our problem is to estimate the proportion of individuals at each BMD-FI level and to test for association between BMD and FI. In NHANES III the response rate increases up to age twenty years, and stabilizes after that age; race, sex and the sampling weights play a minor role (see Nandram and Choi 2005). Thus, for this application we assume that the only data available are the four tables of BMD and FI, and we develop a methodology for this situation.

Table 1
Classification of Bone Mineral Density (BMD) and Family Income (FI) for 2,998 White Females, at Least 20 years Old (20+)

	FI				
BMD	0	1	2	Missing	Sum
0	621	290	284	135	1,330
1	260	131	117	69	577
2	93	30	18	27	168
Missing	456	156	266	45	923
Sum	1,430	607	685	276	2,998

Note: BMD: 0(> 0.82g/cm^2; normal), 1(> 0.64, ≤ 0.82g/cm^2; osteopenia), 2(≤ 0.64g/cm^2; osteoporosis); FI: 0(< \$20,000), 1(≥ \$20,000, < \$45,000), 2(≥ \$45,000); BMD is only measured for age 20+.

It is difficult to assess an association between BMD and FI when there are many individuals not completely classified (*i.e.*, missing data). As discussed in the literature, not necessarily on NHANES III, there are several potentially important confounding variables such as age, smoking, dietary calcium intake, estrogen replacement therapy, physical activity, educational attainment, health status and alcohol consumption (see Ganry, Baudoin and Fardellone 2000). Farahmand, Persson, Michaelsson, Baron, Parker and Ljunghall (2000) stated that for postmenopausal women, aged 50–81 years, from six counties in Sweden, higher household income is associated with decreased hip fracture risk. Using complete data from NHANES III, Lauderdale and Rathouz (2003) studied the regression of bone mineral content on economic indicators (*e.g.*, education and poverty income ratio). An adjustment was made for other factors such as age, height and weight. They conclude that "Bone density does not reflect economic conditions as

strongly or consistently as physical stature." Unfortunately, these works do not address the nonignorability of the missing data; missing data are not discussed. Also, the response rate to income items is usually low.

We have looked at the data for the complete cases more closely. We fit a multinomial-Dirichlet model with association and one with no association . The model with association is $\boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomial}(n, \boldsymbol{p})$ and $\boldsymbol{p} \sim \text{Dirichlet}(1, \ldots, 1)$. Note that by no association we mean that $p_{jk} = p_j^{(1)}\, p_k^{(2)}$, $j = 1, \ldots, r$, $k = 1, \ldots, c$, where $\sum_{j=1}^{r} p_j^{(1)} = 1$ and $\sum_{k=1}^{c} p_k^{(2)} = 1$. Thus, for the model with no association, $\boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomial}(n, \boldsymbol{p})$, $\boldsymbol{p}^{(1)} \sim \text{Dirichlet}(1, \ldots, 1)$, and independently $\boldsymbol{p}^{(2)} \sim \text{Dirichlet}(1, \ldots, 1)$, where $\boldsymbol{p}^{(1)}$ and $\boldsymbol{p}^{(2)}$ have r and c components respectively. It is easy to show that the marginal likelihood with association (as) is $p_{\text{as}}(\boldsymbol{n}) = (rc-1)!\,n!/(n+rc-1)!$ and with no association (nas) is

$$p_{\text{nas}}(\boldsymbol{n}) = p_{\text{as}}(\boldsymbol{n}) \frac{(r-1)!(c-1)!}{(rc-1)!} \frac{(n+rc-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_{j=1}^{r} n_{j\cdot}! \prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!}.$$

Consider our data in Table 1 again. Under independence (*i.e.*, no association) the observed chi-squared statistic is 12.7 on 4 degrees of freedom with a p–value of 0.013 and the hypothesis of no association is rejected. On the logarithmic scale, the marginal likelihoods are $p_{\text{nas}}(\boldsymbol{n}) = -46.2$ and $p_{\text{as}}(\boldsymbol{n}) = -49.6$ resulting in a log Bayes factor of 3.40 for evidence of no association relative to association. Therefore, while the chi-squared test provides strong evidence against no association, the log Bayes factor provides strong evidence for no association. Thus, there is a contradictory evidence for no association. See Mirkin (2001) for a review of interpretations of the chi-squared statistic as a measure of association or independence.

How sensitive is the Bayes factor to the choice of the prior distributions? First, note that the prior density that any reasonable person might use in this problem is the Dirichlet distribution. For the model with association we have selected the prior distributions to be $\boldsymbol{p} \sim \text{Dirichlet}(\boldsymbol{\gamma})$, and for the model with no association $\boldsymbol{p}^{(1)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ and independently $\boldsymbol{p}^{(2)} \sim \text{Dirichlet}(\boldsymbol{\beta})$. Let $n_{j\cdot}^{(1)} = \sum_{k=1}^{c} n_{jk}$, $j = 1, \ldots, r$ and $n_{\cdot k}^{(2)} = \sum_{j=1}^{r} n_{jk}$, $k = 1, \ldots, c$. Then, it is easy to show that the Bayes factor for a test of association versus no association is

$$\text{BF} = \frac{D_{rc}(\boldsymbol{n}+\boldsymbol{\gamma}) / D_r(\boldsymbol{n}_{\cdot}^{(1)}+\boldsymbol{\alpha}) D_c(\boldsymbol{n}_{\cdot}^{(2)}+\boldsymbol{\beta})}{D_{rc}(\boldsymbol{\gamma}) / D_r(\boldsymbol{\alpha}) D_c(\boldsymbol{\beta})},$$

where $D_r(.)$ refers to the Dirichlet function with r components, *etc.*; see section 3.1 for notations. Then, we choose each of the components of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to be κ (*e.g.*, in $p_{\text{as}}(\boldsymbol{n})$ and $p_{\text{nas}}(\boldsymbol{n})$, $\kappa = 1$). Sensitivity to the choice of prior distributions can be studied in terms of κ. Here $\kappa = 1$ corresponds to the prior distributions that are usually used in the multinomial-Dirichlet model, and $\kappa = 0.50$, Jeffreys' prior. Thus, we have chosen $\kappa = 0.25, 0.5, 1.0, 1.5, 2, 3$, and the corresponding Bayes factors (log scale) are 4.7, 3.6, 3.4, 3.9, 4.7, 6.6. Thus, while the Bayes factor is sensitive to the choice of the prior distributions, it is not too sensitive. Of course, if there is informative prior information, in which κ is substantially large, it is a different issue.

The Pearson chi-squared statistic is dominated by cells (3, 1) and (3, 3) with squares of the Pearson residuals being 4.61 and 6.15 respectively (the observed chi-squared statistic is 12.7). It is interesting that the Bayes factor tends to smooth this effect out. We have collapsed the two categories, osteopenia and osteoporosis, into a single category. For this 2×3 categorical table, the chi-squared test statistic is 1.7 on 2 degrees of freedom with a p–value of 0.42. The marginal likelihoods are $p_{\text{nas}}(\boldsymbol{n}) = -28.2$ and $p_{\text{as}}(\boldsymbol{n}) = -32.0$ resulting in a log Bayes factor of –3.81. Therefore, both tests suggest no association for this 2×3 table. Thus, based on these data it is hard to believe that there is an association between BMD and FI. The question that now arises is "Can this conclusion change if we take into account the incomplete data?"

3. Methodology and Nonresponse Models

First, we describe the notation. Second, we describe the ignorable nonresponse model. Third, we construct a nonignorable nonresponse model by expanding the ignorable nonresponse model. Fourth, we discuss the Bayes factor. Finally, we describe how to specify an important prior distribution.

3.1 Notation

For a $r \times c$ categorical table, let $I_{jk\ell} = 1$ if ℓ^{th} individual falls in the j^{th} row and k^{th} column and 0 otherwise. Also, let $J_{s\ell} = 1$ if the ℓ^{th} individual falls in table s ($s = 1$: complete cases; $s = 2$: table with row totals; $s = 3$: table with column totals; $s = 4$: table with individuals unclassified), and $J_{s\ell} = 0$ otherwise, $s = 1, 2, 3, 4$ with $\sum_{s=1}^{4} J_{\ell s} = 1$. The vector $\boldsymbol{J}_{\ell} = (J_{1\ell}, J_{2\ell}, J_{3\ell}, J_{4\ell})'$ has its components corresponding to the four tables.

Let p_{jk} be the probability that an individual belongs to cell (j, k) of the $r \times c$ table, and let π_{sjk} be the probability that an individual belongs to the s^{th} table, given that cell status (j, k). For the ignorable nonresponse model $\pi_{sjk} = \pi_s$, but for a nonignorable nonresponse model π_{sjk} depends on at least one of j and k as well. We will also let

$\boldsymbol{p}$ be the vector p_{jk}, $j=1, \ldots, r$, $k=1, \ldots, c$, and $\boldsymbol{\pi}_{jk}$ be a vector with components $\{\pi_{sjk}, s=1, \ldots, 4\}$, $j=1, \ldots, r$, $k=1, \ldots, c$.

Then, we take

$$\boldsymbol{I}_\ell \mid \boldsymbol{p} \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{p}\}, \quad (1)$$

where $\sum_{j=1}^r \sum_{k=1}^c p_{jk}=1$, $p_{jk} \geq 0$, $j=1, \ldots, r$, $k=1, \ldots, c$. For the parameters $\boldsymbol{p}$ we take

$$\boldsymbol{p} \sim \text{Dirichlet}(1, \ldots, 1),\ p_{jk} \geq 0,\ \sum_{j=1}^r \sum_{k=1}^c p_{jk}=1. \quad (2)$$

Henceforth, we will use the notation that a k-dimensional vector, $\boldsymbol{x} \sim \text{Dirichlet}(\boldsymbol{ct})$ to mean $f(\boldsymbol{x})=\{\prod_j^k x_j^{c_j t-1}\}/D_k(\boldsymbol{ct})$, $x_j \geq 0$, $\sum_{j=1}^k x_j=1$, where $D_k(\boldsymbol{ct})=\{\prod_{j=1}^k \Gamma(c_j t)\}/\Gamma(t)$ is the Dirichlet function with $c_j>0$, $\sum_{j=1}^k c_j=1$.

Assumptions (1) and (2) are the same for both the ignorable and nonignorable nonresponse models, and they are standard when there are no missing data.

Let the cell counts be $y_{sjk}=\sum_{\ell=1}^n I_{jk\ell} J_{s\ell}$, $s=1, 2, 3, 4$ for the four cases. Here y_{1jk} are observed and y_{sjk}, $s=2, 3, 4$ are missing (*i.e.*, latent variables). For y_{1jk} we know that $\sum_{j=1}^r \sum_{k=1}^c y_{1jk}=n_0$, the number of individuals with complete data; for y_{2jk} we know that $\sum_{k=1}^c y_{2jk}=u_j$, where the row margins u_j, $j=1, \ldots, r$ are observed; for y_{3jk} we know that $\sum_{j=1}^r y_{3jk}=v_k$, where the column margins v_k, $k=1, \ldots, c$ are observed; and for y_{4jk} we know that $\sum_{j=1}^r \sum_{k=1}^c y_{4jk}=w$. Throughout we assume that all inference is conditional on $n_0, \boldsymbol{u}, \boldsymbol{v}$ and w, and we will suppress this notation whenever it is understood. Whenever it is convenient, we will use notations such as $\sum_{s,j,k} y_{sjk} \equiv \sum_{s=1}^4 \sum_{j=1}^r \sum_{k=1}^c y_{sjk}$, $\prod_{s,j,k} \pi_{sjk} \equiv \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}$ and $\boldsymbol{y}_{(1)}=(\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4)$, $\boldsymbol{y}_{(2)}=(\boldsymbol{y}_1, \boldsymbol{y}_3, \boldsymbol{y}_4)$ *etc.*, where $\boldsymbol{y}_s=(y_{sjk}, j=1, \ldots, r, k=1, \ldots, c)$, $s=1, 2, 3, 4$. Also, $\sum_{s,j,k}^{4,r,c} y_{sjk}=n$. We will also use $y_{s\cdot\cdot}=\sum_{j,k} y_{sjk}$, $y_{\cdot jk}=\sum_s y_{sjk}$ and $\boldsymbol{y}=(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4)$.

3.2 Ignorable Nonresponse Model

For the ignorable nonresponse model we take

$$\boldsymbol{J}_\ell \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}\}. \quad (3)$$

That is, there is no dependence on the cell status of an individual.

Then, the augmented likelihood function for $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w$ is

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w) \propto \left[\prod_{s=1}^4 \pi_s^{y_{s\cdot\cdot}}\right] \left[\prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk!}}\right], \quad (4)$$

subject to $\sum_{j=1}^r \sum_{k=1}^c y_{1jk}=n_0$, $\sum_{k=1}^c y_{2jk}=u_j$, $j=1, \ldots, r$, $\sum_{j=1}^r y_{3jk}=v_k$, $k=1, \ldots, c$, and $\sum_{j=1}^r \sum_{k=1}^c y_{4jk}=w$. There are three interesting features in (4). First, under ignorability the likelihood function separates into two pieces, one that contains the π_s only and the other the p_{jk}, and inference about these two parameters are unrelated. Second, inference about π_s is based only on the observed $y_{s\cdot\cdot}$ (*i.e.*, the sufficient statistics for π_1, π_2, π_3 and π_4 are essentially the proportions of cases in the first, second, third and fourth tables respectively). Third, under the ignorable nonresponse model, the u_j and the v_k contain information about the p_{jk}; w does not contain any information about the p_{jk}. This is easy to show; letting T denote the set $\{(\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4): \sum_{k=1}^c y_{2jk}=u_j, j=1, \ldots, r, \sum_{j=1}^r y_{3jk}=v_k, k=1, \ldots, c, \sum_{j=1}^r \sum_{k=1}^c y_{4jk}=w\}$, by (4)

$$\sum_{(\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4) \in T} \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!} = w! \prod_{j=1}^r \frac{u_j!}{\left\{\sum_{k=1}^c p_{jk}\right\}^{u_j}} \prod_{k=1}^c \frac{v_k!}{\left\{\sum_{j=1}^r p_{jk}\right\}^{v_k}} \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{1jk}}}{y_{1jk}!}.$$

Finally, for the parameters $\boldsymbol{\pi}$ we take

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1, \ldots, 1),\ \pi_s \geq 0,\ \sum_{s=1}^4 \pi_s=1. \quad (5)$$

Note that this is a uniform probability density in four-dimensional space, and there are no hyperparameters in this model. Thus, for the ignorable nonresponse model, combining (2) and (5), the joint prior density is

$$g_1(\boldsymbol{p}, \boldsymbol{\pi}) \propto 1,\ p_{jk} \geq 0, \sum_{j=1}^r \sum_{k=1}^c p_{jk}=1, \pi_s \geq 0, \sum_{s=1}^4 \pi_s=1, \quad (6)$$

which is proper.

Finally, combining the likelihood function in (4) with the joint prior density in (6) via Bayes' theorem, the joint posterior density of the parameters $\boldsymbol{\pi}$, $\boldsymbol{p}$ and $\boldsymbol{y}_{(1)}$ is

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \left[\prod_{s=1}^4 \pi_s^{y_{s\cdot\cdot}}\right] \left[\prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!}\right]. \quad (7)$$

A posteriori $\boldsymbol{p}$ and $\boldsymbol{\pi}$ are independent. Inference about $\boldsymbol{\pi}$ is easy because $\boldsymbol{\pi} \mid \boldsymbol{y}_1, \boldsymbol{y}_{(1)} \sim \text{Dirichlet}(y_{1\cdot\cdot}+1, \ldots, y_{4\cdot\cdot}+1)$, which is independent of $\boldsymbol{y}_{(1)}$. Inference about $\boldsymbol{p}$ can be obtained using a simple Gibbs sampler because, letting $q_{jk}^{(1)}=p_{jk}/\sum_{k'=1}^c p_{jk'}$ and $q_{jk}^{(2)}=p_{jk}/\sum_{j'=1}^r p_{j'k}$, the conditional probabilities are

$$\boldsymbol{p} \mid \boldsymbol{y} \sim \text{Dirichlet}(y_{\cdot 11}+1, \ldots, y_{\cdot rc}+1),$$

$$\boldsymbol{y}_{2j} \mid \boldsymbol{p}, u_j, \boldsymbol{y}_{(2)} \overset{\text{ind}}{\sim} \text{Multinomial}(u_j, \boldsymbol{q}_j^{(1)}),\ j=1, \ldots, r,$$

$$\boldsymbol{y}_{3k} \mid \boldsymbol{p}, v_k, \boldsymbol{y}_{(3)} \overset{\text{ind}}{\sim} \text{Multinomial}(v_k, \boldsymbol{q}_k^{(2)}),\ k=1, \ldots, c,$$

$$\boldsymbol{y}_4 \mid \boldsymbol{p}, w, \boldsymbol{y}_{(4)} \sim \text{Multinomial}(w, \boldsymbol{p}). \quad (8)$$

Clearly, the parameters $\boldsymbol{p}$ and $\boldsymbol{\pi}$ are identifiable and estimable. Also, note that $\boldsymbol{y}_4$ in (8) is a latent variable and that it does not contribute to inference about $\boldsymbol{p}$. Rather it assists in the computation by providing a simple Gibbs sampler. However, we note that information in $\boldsymbol{y}_4$, via *w*, is important under a nonignorable nonresponse model.

3.3 Nonignorable Nonresponse Model

For nonignorable missing data we take

$$\boldsymbol{J}_\ell \mid \{I_{jk\ell} = 1, I_{j'k'\ell} = 0,\ j \neq j', k \neq k', \boldsymbol{\pi}_{jk}\} \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}_{jk}\}. \quad (9)$$

Assumption (9) specifies that the probabilities an individual belongs to one of the four tables depend on the two characteristics (*i.e.*, row and column classifications) of the individual. In this manner we incorporate the assumption that the missing data is nonignorable. This is an extension of the model in Nandram, Han and Choi (2002). One can also have $\boldsymbol{\pi}_j$ or $\boldsymbol{\pi}_k$ instead of $\boldsymbol{\pi}_{jk}$; the methodology is similar.

Next, we need the likelihood function. Here the augmented likelihood function for $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1$ is

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w) \propto \left[\prod_{s,j,k}^{4,r,c} \frac{\pi_{sjk}^{y_{sjk}}}{y_{sjk}!}\right]\left[\prod_{j,k}^{r,c} p_{jk}^{y_{\cdot jk}}\right], \quad (10)$$

subject to $\sum_{j=1}^r \sum_{k=1}^c y_{1jk} = n_0$, $\sum_{k=1}^c y_{2jk} = u_j$, $j = 1, \ldots, r$, $\sum_{j=1}^r y_{3jk} = v_k$, $k = 1, \ldots, c$, and $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w$.

Observe that in (10) the parameters p_{jk} and π_{sjk} are not identifiable. Clearly, to estimate p_{jk} one needs to know $y_{\cdot jk}$, but only the y_{1jk} are known. Also, to estimate π_{sjk} one needs to know y_{sjk}, $s = 2, 3, 4$. Thus, y_{sjk}, $s = 2, 3, 4$ are also not identifiable. Putting very informative proper priors on the π_{sjk} will help, but this is not a practical solution. If an ignorable model (*i.e.*, $\pi_{sjk} = \pi_s$) is used, then all the parameters can be identified. Therefore, a sensible solution is to attempt to link the $\boldsymbol{\pi}_{jk}$ over (j, k) using a common feature. If the $\boldsymbol{\pi}_{jk}$ come from a common distribution with "known" parameters, we would be able to estimate them. That is, we must attempt to "borrow strength" as in small area estimation. This permits estimation of $\boldsymbol{y}_{(1)}$ which, in turn, will facilitate estimation of the p_{jk} and π_{sjk}.

For the $\boldsymbol{\pi}_{jk}$ we "center" the nonignorable nonresponse model on the ignorable nonresponse model. Specifically, we assume that

$$\boldsymbol{\pi}_{jk} \mid \boldsymbol{\mu}, \tau \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\mu_1\tau, \mu_2\tau, \mu_3\tau, \mu_4\tau),$$

$$\pi_{sjk} \geq 0, \sum_{s=1}^4 \pi_{sjk} = 1, \quad (11)$$

$j = 1, \ldots, r$, $k = 1, \ldots, c$. In (11) the parameter τ tells us about the closeness of the nonignorable nonresponse model to the ignorable nonresponse model. For example, if τ is small, the $\boldsymbol{\pi}_{jk}$ will be very different, and if τ is large, the $\boldsymbol{\pi}_{jk}$ will be very similar. Thus, inference may be sensitive to the choice of τ, and one has to be careful in choosing τ. In the absence of any information about nonignorability, it is natural to choose a prior density for τ so that the nonignorable nonresponse model generalizes the ignorable nonresponse model. This generalization is attained because as τ goes to infinity, the $\boldsymbol{\pi}_{jk}$ converge to the same value over (j, k) (not component-wise), the ignorable nonresponse model. The parameters $\boldsymbol{\mu}$ and τ are not identifiable because the $\boldsymbol{\pi}_{jk}$ are not. Thus, it is impossible to estimate $\boldsymbol{\mu}$ and τ without any information; a natural way to proceed is to attempt to use some of the data already observed.

Specifically, a priori we take μ and τ to be independent with

$$p(\boldsymbol{\mu}) = 1, \mu_s \geq 0, s = 1, 2, 3, 4, \quad \sum_{s=1}^4 \mu_s = 1,\ \tau \sim \text{Gamma}\,(\alpha_0, \beta_0), \tau \geq 0, \quad (12)$$

where α_0 and β_0 are to be specified; without any information about α_0 and β_0 one needs to use the data again. To help specify α_0 and β_0 for the nonignorable nonresponse model, we have used the ignorable nonresponse model. The prior on τ adds extra variation, thereby permitting some degree of nonignorability (see section 3.5). Note again that if τ is very large (*i.e.*, $\alpha_0 >> \beta_0$), this nonignorable nonresponse model degenerates into the ignorable nonresponse model. Thus, an issue of how sensitive inference is to this specification arises. Of course, one can choose other distributions for τ in (12) (*e.g.*, lognormal distribution), but this is really not the key issue.

Combining (2), (11) and (12), the joint prior density of $\boldsymbol{\pi}, \boldsymbol{p}, \boldsymbol{\mu}$ and τ is

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau) \propto \left\{\prod_{j=1}^r \prod_{k=1}^c \frac{\prod_{s=1}^4 \pi_{sjk}^{\mu_s\tau-1}}{D(\boldsymbol{\mu}\tau)}\right\}\tau^{\alpha_0-1}e^{-\beta_0\tau}. \quad (13)$$

Note again that (13) is a proper prior density. Finally, combining the likelihood function in (10) with the joint prior density in (13) via Bayes' theorem, the joint posterior density of the parameters $\boldsymbol{\pi}, \boldsymbol{p}, \boldsymbol{\mu}, \tau$ and the latent variables $\boldsymbol{y}_{(1)}$ is

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \left[\prod_{s,j,k}^{4,r,c} \left\{\frac{(\pi_{sjk}\, p_{jk})^{y_{sjk}}}{y_{sjk}!}\right\}\right] \left\{\prod_{j,k}^{r,c} \frac{\prod_{s=1}^4 \pi_{sjk}^{\mu_s\tau-1}}{D(\boldsymbol{\mu}\tau)}\right\}\tau^{\alpha_0-1}e^{-\beta_0\tau}. \quad (14)$$

In Appendix A we show how to fit the nonignorable nonresponse model to obtain the appropriate inference using the Gibbs sampler.

3.4 Bayes Factor: Tests of Association and Nonignorability

We construct a test for the association between BMD and FI. This test is an assessment of the assumption that $p_{jk} = q_{1j}q_{2k}$, $j = 1, \ldots, r$, $k = 1, \ldots, c$, and $\sum_{j=1}^{r} q_{1j} = 1$ and $\sum_{k=1}^{c} q_{2k} = 1$. We use the Bayes factor, the ratio of the marginal likelihoods under two scenarios (*e.g.*, association versus no association). Note that we observe $\boldsymbol{y}_1$, but $\boldsymbol{y}_{(1)}$ is a set of latent variables. So each marginal likelihood is simply the probability that $\boldsymbol{y}_1$ is the observed value of $\boldsymbol{Y}_1$, which we denote by $p(\boldsymbol{y}_1)$.

We set

$$C = \left\{ \boldsymbol{y}_{(1)} : \sum_{k=1}^{c} y_{2jk} = u_j,\ j = 1, \ldots, r;\ \sum_{j=1}^{r} y_{3jk} = v_k,\ k = 1, \ldots, c; \sum_{j=1}^{r}\sum_{k=1}^{c} y_{4jk} = w \right\}.$$

Then, letting $d = 3!n!(rc-1)!$ and $e = 3!n!(r-1)!(c-1)!$, the marginal likelihood for the ignorable (IG) nonresponse model is

$$p_{\mathrm{IG}}(\boldsymbol{y}_1) = \begin{cases} d \sum\limits_{\boldsymbol{y}_{(1)} \in C} \int\int \prod\limits_{s,j,k}^{4,r,c} \{(\pi_s p_{jk})^{y_{sjk}} / y_{sjk}!\} d\boldsymbol{\pi}\, d\boldsymbol{p}, \\ \qquad \text{association} \\ e \sum\limits_{\boldsymbol{y}_{(1)} \in C} \int\int\int \prod\limits_{s,j,k}^{4,r,c} \{(\pi_s q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} d\boldsymbol{\pi}\, d\boldsymbol{q}_1\, d\boldsymbol{q}_2, \\ \qquad \text{no association,} \end{cases} \tag{15}$$

and letting $\Omega_a = (\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau)$ and $\Omega_{na} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau)$, the marginal likelihood for the nonignorable (NIG) nonresponse model is

$$p_{\mathrm{NIG}}(\boldsymbol{y}_1) = \begin{cases} d \sum\limits_{\boldsymbol{y}_{(1)} \in C} \int_{\Omega_a} \prod\limits_{s,j,k}^{4,r,c} \{(\pi_{sjk} p_{jk})^{y_{sjk}} / y_{sjk}!\} g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau)\, d\Omega_a, \\ \qquad \text{association} \\ e \sum\limits_{\boldsymbol{y}_{(1)} \in C} \int_{\Omega_{na}} \prod\limits_{s,j,k}^{4,r,c} \{(\pi_{sjk} q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau)\, d\Omega_{na}, \\ \qquad \text{no association,} \end{cases} \tag{16}$$

where

$$g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau) = \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} \prod_{j=1}^{r} \prod_{k=1}^{c} \left\{ \frac{\prod_{s=1}^{4} \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\}. \tag{17}$$

The summation in the set C is computationally intensive because there are numerous points $\boldsymbol{y}_{(1)} \in C$ (*i.e.*, we need to sum over all of them). We avoid this problem by first summing over C analytically and the rest is obtained using Monte Carlo integration.

For the ignorable model it is easy to show that

$$p_{\mathrm{IG}}(\boldsymbol{y}_1) = \begin{cases} a = \dfrac{3!n!}{n+1} \dfrac{(rc-1)!}{(n+rc-1)!}, \\ \qquad \text{association} \\ b = \dfrac{3!n!}{n+1} \dfrac{(r-1)!(c-1)!}{(n+r-1)!(n+c-1)!} \dfrac{\prod_j y_{1j\cdot}! \prod_k y_{1\cdot k}!}{\prod_j \prod_k y_{1jk}!}, \\ \qquad \text{no association,} \end{cases} \tag{18}$$

where n is the total number of individuals in the entire table. We describe how to estimate $p_{\mathrm{NIG}}(\boldsymbol{y}_1)$ in Appendix B.

However, we note that a test for ignorability or nonignorability is tenuous because we assume that there is no information about ignorability or nonignorability. Yet, our nonignorable nonresponse model is a generalization of our ignorable nonresponse model. We believe that the test about association under the ignorable nonresponse model or nonignorable nonresponse model is reliable.

Finally, we note that the Bayes factor may be sensitive to prior specifications, especially when there are not enough data to estimate the parameters under test; see Sinharay and Stern (2002) for an interesting discussion on nested models. We have studied sensitivity of the Bayes factor with respect to the specification of α_0 and β_0 in (17); see section 3.5 and Table 6. This is useful because it is an important prior in our nonignorable nonresponse model. However, the main comparison is a test for no association under the ignorable nonresponse model and the nonignorable nonresponse model separately. The parameter τ only enters the nonignorable nonresponse model, and τ has the same prior under association and no association.

3.5 Specification of α_0 and β_0

The specification of the hyperparameters α_0 and β_0 in $\tau \sim \mathrm{Gamma}(\alpha_0, \beta_0)$ is a key issue in our method; see (12). This is important because we use this technique to robustify the ignorable nonresponse model; a sensitivity analysis is performed later. Note that $E(\tau) = \alpha_0 / \beta_0$; thus if $\alpha_0 >> \beta_0$, the nonignorable nonresponse model will be similar to the ignorable nonresponse model. Suppose we can observe a random sample $\tau^{(1)}, \ldots, \tau^{(M)}$ from $\mathrm{Gamma}(\alpha_0, \beta_0)$. Then, we can use a simple method (*e.g.*, the method of moments) to estimate α_0 and β_0.

How can we obtain a sample to fit $\mathrm{Gamma}(\alpha_0, \beta_0)$? The Gibbs sampler in (8) for the ignorable nonresponse model gives imputed values for the missing cell counts. We have imputed the missing cell counts M times, $M = 1{,}000$;

let $n_{1jk}^{(h)} \equiv y_{1jk}$ and $n_{sjk}^{(h)}, s = 2, 3, 4, \quad h = 1, \ldots, M$ denote the missing cell counts. Then, for each h we fit the nonignorable nonresponse model without the prior specification in (12),

$$(n_{111}^{(h)}, \ldots, n_{1rc}^{(h)}, \ldots, n_{411}^{(h)}, \ldots, n_{4rc}^{(h)}) \mid \boldsymbol{\pi}, \boldsymbol{p} \sim \text{Multinomial}\{n, (\pi_{111}p_{11}, \ldots, \pi_{4rc}p_{rc})\},$$

$$\boldsymbol{p} \sim \text{Dirichlet}(\mathbf{1}), \text{ and } \boldsymbol{\pi}_{jk} \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}) \text{ where } \alpha_s = \mu_s \tau, \; s = 1, 2, 3, 4.$$

After integrating out $\boldsymbol{p}$ and $\boldsymbol{\pi}_{jk}$, we get the likelihood function,

$$\prod_{j=1}^{r}\prod_{k=1}^{c}\left[\frac{\Gamma\left(\sum_{s=1}^{4}\alpha_s\right)}{\Gamma\left(\sum_{s=1}^{4}(\alpha_s + n_{sjk}^{(h)})\right)}\prod_{s=1}^{4}\frac{\Gamma(\alpha_s + n_{sjk}^{(h)})}{\Gamma(\alpha_s)}\right], \quad \alpha_s > 0, s = 1, 2, 3, 4. \tag{19}$$

Using the Nelder-Mead algorithm to maximize the likelihood function in (19) over $\alpha_s > 0, s = 1, 2, 3, 4,$ at the h^{th} iterate, we obtain the maximum likelihood estimators $\hat{\boldsymbol{\alpha}}^{(h)}, h = 1, \ldots, M$. Now letting $\tau^{(h)} = \sum_{s=1}^{4}\hat{\alpha}_s^{(h)}$, we view $\tau^{(h)}, h = 1, \ldots, M$ as a random sample from Gamma (α_0, β_0).

Finally, using the method of moments, we fit Gamma(α_0, β_0) to the "data," $\tau^{(h)}, h = 1, \ldots, M$, to get $\alpha_0 = a^2/b$ and $\beta_0 = a/b$, where $a = M^{-1}\sum_{h=1}^{M}\tau^{(h)}$ and $b = (M-1)^{-1}\sum_{h=1}^{M}(\tau^{(h)} - a)^2$. Thus, we have constructed a data-dependent prior distribution for τ. Our procedure gives $\alpha_0 = 125, \; \beta_0 = 0.35$ (*i.e.*, τ has mean 357 and standard deviation 31.9). In section 4 we discuss sensitivity to this choice.

4. Data and Empirical Analysis

We apply our methodology to the data in the 3×3 categorical table in Table 1. After we present results associated with the observed data and a sensitivity analysis, we describe a simulation study to assess the difference between the ignorable and the nonignorable nonresponse models.

4.1 Data Analysis

See Table 2 for a comparison of the ignorable nonresponse model and the nonignorable nonresponse model. We have also included the numerical standard error (NSE) which is a measure of how well the numerical results can be reproduced; we have used the batch-means method to compute it. Thus, one would be comfortable with small NSE's relative to the Monte Carlo estimates or the posterior means. For both models the NSE's are small with relatively larger values for the nonignorable nonresponse model (both near zero any way), indicating that the computations are repeatable. The posterior means (PM) are very similar for the two models. The posterior standard deviations (PSD) are larger for the nonignorable model, making the 95% credible intervals wider. Virtually all the 95% credible intervals under the ignorable nonresponse model are contained by those of the nonignorable nonresponse model.

Table 2
Comparison of the Posterior Means (PM), Posterior Standard Deviations (PSD), Numerical Standard Errors (NSE), and 95% Credible Intervals (CI) for p from the Ignorable and Nonignorable Nonresponse Models

Cell	$\hat{p}$	PM	PSD	NSE	CI
(a) Ignorable Model					
(1, 1)	0.337	0.330	0.005	0.001	(0.321, 0.339)
(1, 2)	0.157	0.142	0.003	0.001	(0.136, 0.147)
(1, 3)	0.154	0.168	0.004	0.001	(0.162, 0.175)
(2, 1)	0.141	0.142	0.004	0.001	(0.134, 0.148)
(2, 2)	0.071	0.066	0.002	0.001	(0.061, 0.070)
(2, 3)	0.063	0.071	0.003	0.001	(0.066, 0.078)
(3, 1)	0.050	0.053	0.003	0.001	(0.048, 0.059)
(3, 2)	0.016	0.016	0.001	0.000	(0.013, 0.019)
(3, 3)	0.010	0.012	0.002	0.000	(0.009, 0.015)
(b) Nonignorable Model					
(1, 1)	0.337	0.321	0.020	0.009	(0.278, 0.355)
(1, 2)	0.157	0.143	0.008	0.003	(0.126, 0.158)
(1, 3)	0.154	0.173	0.014	0.007	(0.140, 0.196)
(2, 1)	0.141	0.139	0.019	0.009	(0.109, 0.182)
(2, 2)	0.071	0.069	0.007	0.003	(0.056, 0.085)
(2, 3)	0.063	0.071	0.013	0.006	(0.053, 0.102)
(3, 1)	0.050	0.052	0.008	0.002	(0.040, 0.070)
(3, 2)	0.016	0.019	0.003	0.001	(0.014, 0.026)
(3, 3)	0.010	0.013	0.003	0.001	(0.009, 0.020)

Note: The ignorable nonresponse model has $\pi_{sjk} = \pi_s$, $s = 1, 2, 3, 4, \; j = 1, 2, 3, \; k = 1, 2, 3$. The observed value of p based on the complete data is $\hat{p}$.

In Table 3 we have also compared the estimation of π_s in the ignorable nonresponse model with π_{sjk} in the nonignorable nonresponse model. For the nonignorable nonresponse model we present the range of the posterior means (PM) for the nine cells of each $s, \; s = 1, 2, 3, 4$. This indicates the extent of the nonignorability. The PM's of π_s are within the range of the π_{sjk}, and as expected, the PSD's are larger for the nonignorable model. For example, over the nine cells the π_{1jk} vary from 0.388 to 0.656, and these two numbers differ significantly from 0.615, showing some degree of nonignorability. Thus, there is some difference between the ignorable and the nonignorable nonresponse models.

In Table 4 we have presented the logarithms of the Bayes factors for testing the goodness of fit of the ignorable nonresponse model and the nonignorable nonresponse model. There is "strong" evidence that the ignorable nonresponse model fits better than the nonignorable nonresponse model for these data (Kass and Raftery 1995). While the ignorable nonresponse model provides "strong" evidence for no association, the evidence from the nonignorable nonresponse model is "positive" as stated by Kass and Raftery (1995).

Thus, again there is a difference between the ignorable and the nonignorable nonresponse models. However, the NSE of 1.80 tends to nullify such differences. Our conclusion is that there is strong evidence to suggest no association between BMD and FI.

Table 3
Comparison of the Posterior Means (PM) and Posterior Standard Deviations (PSD) for π_{sjk} from the Ignorable and Nonignorable Nonresponse Models

	Ignorable	Nonignorable
π_1	0.615 (0.009)	0.388 (0.078) – 0.656 (0.044)
π_2	0.077 (0.005)	0.057 (0.017) – 0.195 (0.068)
π_3	0.292 (0.008)	0.217 (0.041) – 0.349 (0.053)
π_4	0.015 (0.002)	0.013 (0.005) – 0.152 (0.055)

Note: PSD's are in parentheses. For the ignorable nonresponse model the parameters are π_1, π_2, π_3 and π_4 and for the nonignorable nonresponse model the parameters are π_{sjk}, $s = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3$. Among the nine cells for each s we selected the smallest PM and the largest PM to form the range.

Table 4
Marginal Likelihoods and Bayes Factors for Testing Association Between BMD and FI Under the Ignorable and the Nonignorable Nonresponse Models

	Association	No association	Difference
Ignorable	−49.571	−46.173	−3.398
Nonignorable	−53.129	−50.132	−2.996
NSE	1.800	1.790	

Note: All entries (marginal likelihoods and their differences) are on the logarithmic scale. The Monte Carlo integration uses 50,000 iteractions. The NSEs, numerical standard errors, are small relative to the marginal likelihoods.

We have considered the relation between BMD and FI when the osteopenia and osteoporosis levels are collapsed into one level. Under the ignorable nonresponse model the log Bayes factor is –2.77 (log marginal likelihoods: –32.82 and – 29.05), and under the nonignorable nonresponse model the log Bayes factor is – 4.52 (log marginal likelihoods: –34.25 and – 4.52). Thus, the same conclusion is reached about no association between BMD and FI.

We have also separated out the data into two age groups: premenopausal (age at most 49 years old; young) and postmenopausal (age at least 50 years old; old). For the young group there were only 4 females with osteoporosis, and so we collapsed the females with osteopenia and osteoporosis. We fit both the ignorable and nonignorable nonresponse models to these data and got similar results. For the old group using the ignorable nonresponse model the log marginal likelihoods corresponding to no association and association are – 43.01 and – 38.91 giving a log Bayes factor of 4.10 for no association. Thus, there is strong evidence for no association between BMD and FI. For the young group using the ignorable nonresponse model the log marginal likelihoods corresponding to no association and association are – 29.93 and – 28.80 giving a log Bayes factor of 1.13 for no association. Thus, there is positive evidence for no association between BMD and FI for both age groups. Therefore, age is unlikely to play a role in the association of BMD and FI.

4.2 Sensitivity Analysis

We have studied the sensitivity of inference about the p_{jk} with respect to the prior distribution of τ. That is, we have taken $\tau \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is a sensitivity parameter that we have taken to be 1 in our analysis (note that $E(\tau) = \kappa\alpha_0 / \beta_0$).

Our procedure for the specification of α_0 and β_0 gives values of $\alpha_0 = 125$ and $\beta_0 = 0.35$; see section 3.5. Making κ bigger than 1 induces less changes in the posterior mean (PM) and posterior standard deviation (PSD) of the p_{jk} than for κ smaller than 1 because larger values of κ induces much smaller changes in the prior distribution of τ. In Table 5 we present PM's and PSD's of the p_{jk} for $\kappa = 0.25, 0.50, 1.00, 2.00, 4.00$. The PM's increase with κ and the PSD's decrease as κ increases from 0.25 to 4.00. Thus, there is some sensitivity to the specification of α_0 and β_0, but the changes are small. For example, the PM's of p_{11} are 0.31, 0.32, 0.33 at $\kappa = 0.25, 1.00, 4.00$ and the PSD's at these values of κ are 0.04, 0.02, 0.01.

Table 5
Sensitivity of the Posterior Means (PM) and Posterior Standard Deviations (PSD) of the p_{jk} to Choices of κ in the Nonignorable Nonresponse Model

κ	0.25		0.50		1.00		2.00		4.00	
Cell	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
(1, 1)	306.93	36.09	315.01	25.81	321.81	19.95	325.37	14.55	326.16	10.46
(1, 2)	141.12	15.52	139.86	11.91	142.66	8.44	142.63	6.68	143.42	5.01
(1, 3)	161.68	25.80	167.83	18.77	173.40	13.77	176.20	8.44	175.78	6.71
(2, 1)	143.18	34.20	142.62	24.92	138.57	18.82	137.23	13.59	137.26	9.70
(2, 2)	68.46	13.12	71.06	10.09	68.44	7.48	68.79	5.72	68.11	4.45
(2, 3)	79.78	22.83	75.97	17.86	71.11	12.56	68.09	7.84	68.34	6.38
(3, 1)	59.97	21.60	53.50	12.12	52.14	7.76	50.97	5.29	51.41	4.35
(3, 2)	21.43	7.76	20.02	4.89	18.96	23.28	18.67	2.78	17.84	2.23
(3, 3)	17.45	10.38	14.12	4.28	12.93	2.99	12.05	2.34	11.69	1.99

Note: All entries must be multiplied by 10^{-3}. In the nonignorable nonresponse model $\pi_{sjk} \overset{\text{iid}}{\sim} \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is the sensitivity parameter and $\alpha_0 = 125$ and $\beta_0 = 0.35$.

We have also studied the sensitivity of the Bayes factors to choices of κ (see Table 6). First, the NSE's decrease with κ, but the change is small. Note that we have used 50,000 iterations in the Monte Carlo integration; this sample size is needed for the Monte Carlo estimates to stabilize. The log marginal likelihoods do not change too much with κ. Because the log Bayes factors are small, some changes are reflected in inference: At $\kappa = 0.25, 0.50, 4.00$ there is "strong" evidence for no association, but at $\kappa = 1.00, 2.00$ there is "positive" (borderline) evidence for no association. Overall, there is some degree of evidence for no association. Thus, it is interesting that one does not need to worry too much about the choice for (α_0, β_0).

Table 6
Sensitivity of the Marginal Likelihoods and the Bayes Factor to Choices of κ in the Nonignorable Nonresponse Model

	Association		No Association		Bayes Factor
κ	ML	NSE	ML	NSE	
0.25	−53.37	1.90	−49.16	1.89	−4.21
0.50	−52.58	1.83	−49.49	1.82	−3.08
1.00	−52.58	1.80	−49.76	1.79	−2.82
2.00	−52.81	1.79	−49.83	1.78	−2.98
4.00	−52.95	1.78	−49.91	1.77	−3.04

Note: All entries are on the logarithm scale. In the nonignorable nonresponse model $\pi_{sjk} \overset{\text{iid}}{\sim} \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is the sensitivity parameter and $\alpha_0 = 125$ and $\beta_0 = 0.35$.

4.3 Simulation Study

We have performed a simulation study to further compare the ignorable and nonignorable nonresponse models. Our objective is to confirm differences that exist between the two models. In our situation a test based on the Bayes factor can confirm one or the other. With limited information about nonignorability (our current situation), it is sensible to fit an ignorable nonresponse model because all the parameters are identifiable in the ignorable nonresponse model. Thus, we proceed by comparing the ignorable and nonignorable nonresponse models when data are generated from (a) the ignorable nonresponse model and (b) the nonignorable nonresponse model. This is a typical Bayesian analysis.

We obtained the posterior means of the p_{jk} and the π_{sjk}, denoted by $\tilde{p}_{jk}$ and $\tilde{\pi}_{sjk}$ respectively, after the nonignorable nonresponse model is fit to the observed data. For the ignorable model we took $\tilde{\pi}_s = \sum_{j=1}^{r}\sum_{k=1}^{c}\tilde{\pi}_{sjk}/rc$, $s = 1, 2, 3, 4$. We obtained the cell counts for the ignorable model by drawing from

$$(y_{111}, \ldots, y_{1rc}, \ldots, y_{411}, \ldots, y_{4rc}) \mid \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}} \sim \text{Multinomial}\{n, (\tilde{\pi}_1\ \tilde{p}_{11}, \ldots, \tilde{\pi}_4\ \tilde{p}_{rc})\}$$

and for the nonignorable model by drawing from

$$(y_{111}, \ldots, y_{1rc}, \ldots, y_{411}, \ldots, y_{4rc}) \mid \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}} \sim \text{Multinomial}\{n, (\tilde{\pi}_{111}\tilde{p}_{11}, \ldots, \tilde{\pi}_{4rc}\tilde{p}_{rc})\},$$

where $n = 2{,}998$, the total number of individuals in the original data set (see Table 1). We have generated 1,000 datasets from each of the ignorable and nonignorable nonresponse models. Then, we fit the ignorable and nonignorable nonresponse models to each dataset in exactly the same manner for the observed data in Table 1, and we computed the posterior means (PM) and the posterior standard deviations (PSD) for the p_{jk}. In Table 7 we present the averages of the PM's and PSD's over the 1,000 datasets. The second column (labeled $\hat{p}$) has the posterior mean of p_{jk} for the observed data under the nonignorable nonresponse model (see Table 2b).

For (a) in Table 7 the PM's are very close to the $\hat{p}_{jk}$ for the ignorable nonresponse model, but not so close when the nonignorable nonresponse model is fit. It is noticeable that

Table 7
Comparison of the Ignorable and Nonignorable Nonresponse Models Via the Simulated Data and the Posterior Means (PM) and Posterior Standard Deviations (PSD) of the p_{jk}

	Simulated	Ignorable (a)				Nonignorable (b)			
	Fitted	Ignorable		Nonignorable		Ignorable		Nonignorable	
Cell	$\hat{p}$	PM	PSD	PM	PSD	PM	PSD	PM	PSD
(1, 1)	321.81	320.73	5.72	307.42	11.30	332.02	5.10	324.44	10.60
(1, 2)	142.66	142.96	4.24	146.44	7.34	141.81	3.30	143.44	5.43
(1, 3)	173.40	172.59	4.42	173.49	7.62	168.66	4.14	174.10	7.04
(2, 1)	138.57	138.82	4.81	135.32	9.82	143.63	4.52	139.20	9.74
(2, 2)	68.44	68.44	3.55	72.01	6.02	64.51	2.91	68.20	4.76
(2, 3)	71.11	71.41	3.65	75.00	6.30	70.85	3.76	69.63	6.58
(3, 1)	52.14	52.17	3.11	53.03	4.95	53.08	3.04	52.44	4.70
(3, 2)	18.96	19.35	2.08	21.65	2.98	15.08	1.72	17.32	2.48
(3, 3)	12.93	13.54	1.78	15.64	2.55	10.95	1.85	11.20	2.18

Note: Data are simulated from the ignorable nonresponse model in (a) or the nonignorable nonresponse model in (b), and both the ignorable and nonignorable nonresponse models are fit. We have generated 1,000 datasets, and we fit both the ignorable and nonignorable nonresponse models to each simulated dataset. The PM's and PSD's are averages over the 1,000 datasets and $\hat{p}$ is the posterior mean for the observed data which we used to generate the data sets. All entries must be multiplied by 10^{-3}.

the PSD's under the nonignorable nonresponse model are about twice as large as those under the ignorable nonresponse model. For (b) in Table 7 the PM's for the nonignorable nonresponse model are closer to the $\hat{p}_{jk}$ than those from the ignorable nonresponse model. However, in both cases the PSD's for the nonignorable nonresponse model are about twice those from the ignorable nonresponse model. For example, in Table 7 for the (1, 1) cell as compared with 0.322 for $\hat{p}$, in (a) the ignorable (nonignorable) model gives a PM of 0.321 (0.307), but in (b) the ignorable (nonignorable) model gives a PM of 0.332 (0.324) for other examples. Thus, the two models are indeed different for estimating $\boldsymbol{p}$.

We have also considered estimating the proportion P of simulated datasets in which the ignorable nonresponse model performs better than the nonignorable nonresponse model. It is expensive to compute the marginal likelihood under the nonignorable nonresponse model. We note again that it takes 50,000 iterations for the Monte Carlo estimate to stabilize; this is an enormous task for the simulation study because we need to calculate the marginal likelihoods for 1,000 datasets. Thus, we use a simple procedure to compare the two models, and we expect that this procedure would give a conclusion similar to a power calculation.

Specifically, we compute $\Delta^{(h)} = n\sum_{j=1}^{r}\sum_{k=1}^{c}(\hat{p}_{jk} - PM_{jk}^{(h)})^2 / PM_{jk}^{(h)}$, where $PM_{jk}^{(h)}$ is the posterior mean of p_{jk} corresponding to the h^{th} dataset. We denote $\Delta^{(h)}$ by $\Delta_{\text{IG}}^{(h)}$ for the ignorable nonresponse model and $\Delta_{\text{NIG}}^{(h)}$ for the nonignorable nonresponse model. An estimator of P, $\hat{P}$, is obtained by counting the number of the 1,000 experiments in which $\Delta_{\text{IG}}^{(h)} > \Delta_{\text{NIG}}^{(h)}$. For the data generated from the ignorable nonresponse model, $\hat{P}$ is 0.236 with a standard error of 0.013. For the data generated from the nonignorable nonresponse model, $\hat{P}$ is 0.920 with a standard error of 0.009. Thus, if the ignorable nonresponse model is expected to hold, about 24% of the time the nonignorable nonresponse model will beat it, and if the nonignorable nonresponse model is expected to hold, only about (1 – 0.920)100% ≈ 8% of the time the ignorable nonresponse model will beat it. Thus, there are latent differences between these two models. The nonignorable nonresponse model does capture some degree of nonignorability, and it robustifies the ignorable nonresponse model. We believe that this is a reasonable comparison between the ignorable and the nonignorable nonresponse models.

5. Concluding Remarks

There are two key methodological developments in this paper. Specifically, we have shown that (a) it is possible to analyze multinomial data from $r \times c$ categorical tables when there are both item and unit nonresponses, and the nonresponse mechanism may be nonignorable; and (b) by using the Bayes factor (ratio of the marginal likelihoods of two models), we can test for association between the two categories. Essentially, we have assumed that there is no information about nonignorability, all design features are suppressed and we have taken a conservative ground.

For the 3×3 categorical data of BMD and FI, we have shown how to estimate the cell probabilities accurately. For the complete cases, the Bayes factor shows "strong" evidence for no association between BMD and FI. For all the data, our Bayes factor shows that the evidence for no association is "strong" under the ignorable nonresponse model, and is "positive" under the nonignorable nonresponse model. Thus, there is virtually no difference between the two scenarios: data from only the complete cases are used and all the data are used. Also, based on the Bayes factor and our simulation study, while there are differences between the ignorable nonresponse model and the nonignorable nonresponse models, such differences are small. There are differences for inference about the proportions of individuals in various BMD-FI levels; the posterior means are similar but the posterior standard deviations under the nonignorable nonresponse model are larger than those under the ignorable nonresponse model.

Our simulation study supports two properties (subtle differences) of our models. First, the estimates of the cell probabilities from the ignorable (nonignorable) nonresponse model are closer to the true values when the ignorable (nonignorable) nonresponse model is expected to hold, but in either case the estimates from the nonignorable nonresponse model have about twice the standard deviations from the ignorable nonresponse model. Second, if the ignorable (nonignorable) nonresponse model is expected to hold, it can be beaten by the nonignorable (ignorable) nonresponse model. This happens a significantly larger proportion of time when the ignorable nonresponse model is expected to hold. Thus, there are differences between these models. We suggest fitting both models, and compute the Bayes factor to decide which one to use. We do not recommend using these models when there are appropriate covariates and/or prior information to explain nonignorability.

In future research one can attempt to reduce the number of parameters in the nonignorable nonresponse model to further reduce the effects of nonignorability. For example, it may be possible to consider representing the data in two categorical tables as follows. The three supplemental tables are collapsed into a single supplemental table with its j^{th} row having *at least* u_j individuals, and its k^{th} column having *at least* v_k individuals; the total number of individuals in this supplemental table is $w + \sum_{j=1}^{r} u_j + \sum_{k=1}^{c} v_k$;

see section 3.1 for notations. Finally, we note that a full analysis of data from a complex survey requires an input of information (covariates and prior information) about non-ignorability, sampling weights and clustering effects as well.

Appendix A
Fitting the Nonignorable Nonresponse Model

We show how to use the Gibbs sampler to make inference about the parameters in (14). The conditional posterior density of p is

$$\boldsymbol{p} \mid \boldsymbol{y} \sim \text{Dirichlet}(y_{\cdot 11}+1, \ldots, y_{\cdot rc}+1) \tag{A.1}$$

and the conditional posterior density of $\boldsymbol{\pi}_{jk}$ is

$$\boldsymbol{\pi}_{jk} \mid \{\boldsymbol{\mu}, \tau, \boldsymbol{y}\} \overset{\text{ind}}{\sim} \text{Dirichlet}\begin{pmatrix} y_{1jk}+\mu_1\tau, y_{2jk} \\ +\mu_2\tau, y_{3jk}+\mu_3\tau, y_{4jk}+\mu_4\tau \end{pmatrix} \tag{A.2}$$

with independence over $j=1, \ldots, r, k=1, \ldots, c$.

We need the conditional posterior probability mass functions of $\boldsymbol{y}_s, s=2, 3, 4$ given $\boldsymbol{y}_{(s)}, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c$. From (14) it is clear that the $\boldsymbol{y}_s, s=2, 3, 4$ are conditionally independent multinomial random vectors. Specifically,

$$\begin{aligned} \boldsymbol{y}_{2j} \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \\ \overset{\text{ind}}{\sim} \text{Multinomial}(u_j, \boldsymbol{q}_j^{(2)}),\ j=1, \ldots, r, \\ \boldsymbol{y}_{3k} \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \\ \overset{\text{ind}}{\sim} \text{Multinomial}(v_k, \boldsymbol{q}_k^{(3)}),\ k=1, \ldots, c, \\ \boldsymbol{y}_4 \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \\ \sim \text{Multinomial}(w, \boldsymbol{q}^{(4)}), \end{aligned} \tag{A.3}$$

where $q_{jk}^{(2)} = \pi_{2jk} p_{jk} / \sum_{k'=1}^{c} \pi_{2jk'} p_{jk'}, k=1, \ldots, c, q_{jk}^{(3)} = \pi_{3jk} p_{jk} / \sum_{j'=1}^{r} \pi_{3j'k} p_{j'k}, j=1, \ldots, r,$ and $q_{jk}^{(4)} = \pi_{4jk} p_{jk} / \sum_{j'=1}^{r} \sum_{k'=1}^{c} \pi_{4j'k'} p_{j'k'}, j=1, \ldots, r, k=1, \ldots, c.$

Next, we consider the hyper-parameters. Letting $\delta_s = \prod_{j=1}^{r} \prod_{k=1}^{c} \pi_{sjk}$, the joint conditional posterior density of $\boldsymbol{\mu}, \tau$ is

$$\begin{aligned} p(\boldsymbol{\mu}, \tau \mid \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \\ \propto \left[\left\{\prod_{s=1}^{4} \delta_s^{\mu_s \tau}\right\} / \{D(\boldsymbol{\mu}\tau)\}^{rc}\right] \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}, \end{aligned} \tag{A.4}$$

where $\sum_{s=1}^{4} \mu_s = 1, \mu_s \geq 0, s=1, 2, 3, 4, \tau > 0.$

We use the grid method to get samples from the conditional posterior density of $p(\boldsymbol{\mu} \mid \tau, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c)$ and $p(\tau \mid \boldsymbol{\mu}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, \quad k=1, \ldots, c)$. After transforming τ to $\varphi/(1-\varphi)$, the parameters now live on $(0, 1)$ with appropriate constraints, making the grid procedure convenient. We use 50 intervals of equal widths (obtained by experimentation) to draw $\boldsymbol{\mu}$ and φ, and a random deviate for τ is $\varphi/(1-\varphi)$.

The Gibbs sampler is executed by drawing a random deviate from each of the conditional posterior "densities", (A.1), (A.2), (A.3), and (A.4) in turn, iterating the entire procedure until convergence. This is an example of the griddy Gibbs sampler (Ritter and Tanner 1992).

Appendix B
Estimation of $p_{\text{NIG}}(y_1)$ in (16)

Letting n_m denote the number of incomplete cases (*i.e.*, $n = n_0 + n_m$), one can also show that for the model with association $p_{\text{NIG}}(\boldsymbol{y}_1) = a((n+1)!)/(n_0! n_m!)A$ and for the model with no association $p_{\text{NIG}}(\boldsymbol{y}_1) = b((n+1)!/(n_0! n_m!))B$, where a and b are given in (18),

$$\begin{aligned} A = \\ \int_{\Omega_a} \left\{\prod_{j,k} \pi_{1jk}^{y_{1jk}}\right\} \left\{\sum_{s=2}^{4} \sum_{j,k} \pi_{sjk} p_{jk}\right\}^{n_m} \left\{\frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111}+1, \ldots, y_{1rc}+1)}\right\} \\ \times \prod_{j,k} \left\{\frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)}\right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_a, \end{aligned}$$

$$\begin{aligned} B = \\ \int_{\Omega_{na}} \left\{\prod_{j,k} \pi_{1jk}^{y_{1jk}}\right\} \left\{\sum_{s=2}^{4} \sum_{j,k} \pi_{sjk} q_{1j} q_{2k}\right\}^{n_m} \\ \frac{\prod_j q_{1j}^{y_{1j}}}{D(y_{11\cdot}+1, \ldots, y_{1r\cdot}+1)} \times \frac{\prod_k q_{2k}^{y_{1jk}}}{D(y_{1\cdot 1}+1, \ldots, y_{1\cdot c}+1)} \\ \prod_{j,k} \left\{\frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)}\right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_{na}. \end{aligned} \tag{B.1}$$

Note that $0 < A, B < 1$ gives a useful diagnostic check on the computation.

We show how to compute A in (B.1) using Monte Carlo integration; the procedure to compute B is similar. We prefer the simpler procedure based on Monte Carlo integration with an importance function (Nandram and Kim 2002) rather than the method based on a continuation of the Gibbs sampler (Chib and Jeliazkov 2001).

For A we choose the importance function

$$\begin{aligned} \pi_{im}(\Omega_a) = \frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111}+1, \ldots, y_{1rc}+1)} \prod_{j,k} \left[\frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)}\right] \\ \frac{\prod_{s=1}^{4} \mu_s^{\tilde{\mu}_1 \tilde{\tau} - 1}}{D(\tilde{\boldsymbol{\mu}}\tilde{\tau})} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} \end{aligned}$$

where $\tilde{\mu}_s$ and $\tilde{\tau}$ are estimates obtained using a Gibbs output. We obtain a sample from $\pi_{im}(\Omega_a)$ by drawing $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$, $\boldsymbol{\mu} \sim \text{Dirichlet}(\tilde{\boldsymbol{\mu}}, \tilde{\tau})$, $\boldsymbol{\pi}_{jk} \mid \boldsymbol{\mu}, \tau \sim \text{Dirichlet}(\boldsymbol{\mu}, \tau)$ and $\boldsymbol{p} \mid \boldsymbol{y}_1 \overset{\text{ind}}{\sim} \text{Dirichlet}(y_{111}+1, \ldots, y_{1rc}+1)$.

Then, letting $w_h = \sum_j^r \sum_k^c y_{1jk} \log \pi_{1jk}^{(h)} + n_m \log[\sum_{s=2}^4 \sum_j^r \sum_k^c \pi_{sjk}^{(h)} p_{jk}^{(h)}] - \sum_{s=1}^4 (\tilde{\mu}_s \tilde{\tau} - 1) + \log \mu_s^{(h)} + \log(D(\tilde{\boldsymbol{\mu}}\tilde{\tau}))$, $h = 1, \ldots, M$, an estimator of A is $\hat{A} = M^{-1} \sum_{h=1}^M e^{\omega_h}$. The numerical standard error (NSE) of $\log(\hat{A})$ can be approximated. For letting $\bar{\omega} = M^{-1} \sum_{h=1}^M \omega_h$ and $S^2 = (M-1)^{-1} \sum_{k=1}^M (\omega_h - \bar{\omega})^2$, we have $\mathrm{Var}(\hat{A}) \approx e^{2\bar{\omega}} S^2/M$, $\mathrm{Var}(\log(\hat{A})) \approx (\mathrm{Var}(\hat{A})/e^{2\bar{\omega}}) \approx S^2/M$, and the NSE is $S/\sqrt{M}$ approximately. We start with $M = 10{,}000$ independent samples from the importance function, and increasing M until convergence which occurs about $M = 50{,}000$.

Acknowlegements

This was a part of the work done during the academic year 2003/2004 when Balgobin Nandram was on sabbatical as a Research Scientist at the National Center for Health Statistics, Hyattsville, Maryland. We are grateful to the Associate Editor and the two referees for their informative comments, and the three opportunities to revise the manuscript.

References

Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Chib, S., and Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96, 270-281.

Cohen, G., and Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents. *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, Series B, 57, 45-97.

Farahmand, B.Y., Persson, P.G., Michaelsson, K., Baron, J.A., Parker, M.G. and Ljunghall, S. (2000). Socioeconomic status, marital status and hip fracture risk: a population-based case control study. *Osteoporosis International*, 11, 803-808.

Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society*, Series B, 60, 57-70.

Ganry, O., Baudoin, C. and Fardellone, P. (2000). Effect of alcohol intake on bone mineral density in elderly women: The EPIDOS Study. *American Journal of Epidemiology*, 151, 8, 773-780.

Kass, R., and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Lauderdale, D.S., and Rathouz, P.J. (2003). Does bone mineralization reflect economic conditions? An examination using a national US sample. *Economics and Human Biology*, 1, 91-104.

Little, R.J. (2003). Bayesian Approach to Sample Survey Inference. In *Analysis of Survey Data*, (Eds. R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

Little, R.J.A., and Rubin D.B. (2002). *Statistical Analysis with Missing Data*. Edition, New York: John Wiley & Sons, Inc.

Looker, A.C., Orwoll, E.S., Johnston, C.C., Lindsay, R.L., Wahner, H.W., Dunn, W., Calvo, M.S. and Harris, T.B. (1997). Prevalence of low femoral bone density in older U.S. adults from NHANES III. *Journal of Bone and Mineral Research*, 12, 1761-1768.

Looker, A.C., Wahner, H.W., Dunn, W.L., Calvo, M.S., Harris, R.R., Heyse, S.P., Johnston, C.C. and Lindsay, R. (1998). Updated data on proximal femur bone mineral levels of us adults. *Osteoporosis International*, 8, 468-489.

Mirkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55, 111-120.

Nandram, B., and Choi, J.W. (2002 a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., and Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., and Choi, J.W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the nhanes data. *Survey Methodology*, 31, 73-84.

Nandram, B., Han, G. and Choi, J.W. (2002). A hierarchical Bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, 28, 145-156.

Nandram, B., and Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72, 319-340.

Nandram, B., Liu, N., Choi, J.W. and Cox, L.H. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Ritter, C., and Tanner, M.A. (1992). The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

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

Rubin, D.B., Stern, H.S. and Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

Sinharay, S., and Stern, H.S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56, 196-201.

Wang, H. (2001). *Two-way Contingency Tables with Marginally and Conditionally Imputed Nonrespondents*, Ph.D. Dissertation, *Department of Statistics*, University of Wisconsin-Madison.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 227-231
Statistics Canada, Catalogue No. 12-001-XPB

On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment

Jean-François Beaumont [1]

Abstract

Nonresponse weight adjustment is commonly used to compensate for unit nonresponse in surveys. Often, a nonresponse model is postulated and design weights are adjusted by the inverse of estimated response probabilities. Typical nonresponse models are conditional on a vector of fixed auxiliary variables that are observed for every sample unit, such as variables used to construct the sampling design. In this note, we consider using data collection process variables as potential auxiliary variables. An example is the number of attempts to contact a sample unit. In our treatment, these auxiliary variables are taken to be random, even after conditioning on the selected sample, since they could change if the data collection process were repeated for a given sample. We show that this randomness introduces no bias and no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, when nonresponse depends on the variables of interest, we argue that the use of data collection process variables is likely to reduce the nonresponse bias if they provide information about the variables of interest not already included in the nonresponse model and if they are associated with nonresponse. As a result, data collection process variables may well be beneficial to handle unit nonresponse. This is briefly illustrated using the Canadian Labour Force Survey.

Key Words: Nonresponse bias; Nonresponse model; Nonresponse variance; Number of attempts; Paradata; Response probability.

1. Introduction

Unit nonresponse is often handled in surveys by using a nonresponse weight adjustment method. The basic principle that is often chosen is to adjust the design weights by the inverse of estimated response probabilities (see, for example, Ekholm and Laaksonen 1991). These estimated response probabilities are obtained by postulating a model for the unknown nonresponse mechanism, which we call the nonresponse model. Key to reducing the nonresponse bias and variance as much as possible is to condition on a vector of auxiliary variables that are observed for every sample unit and that are good predictors of both nonresponse and the variables of interest (Little and Vartivarian 2005). Usually, the auxiliary variables are treated as being fixed both unconditionally and conditionally on the selected sample.

In this note, we consider using Data Collection Process (DCP) varimpbleps as potential auxiliary vmpriables to be included in the nonresponse model. An example is the number of attempts to contact a sample unit. Such type of data is sometimes called paradata (see Couper and Lyberg 2005 for a recent reference on paradata) and has been used to deal with unit nonresponse by Holt and Elliott (1991), among others. In our treatment, contrary to Holt and Elliott (1991), DCP variables are taken to be random, even after conditioning on the selected sample, since they could change if the data collection process were repeated for a given sample.

DCP variables may be particularly useful in cross-sectional surveys where the auxiliary variables available to handle unit nonresponse are often limited to variables used to construct the sampling design. Although such design variables are not useless, they are often neither very good predictors of nonresponse nor the variables of interest. The additional information from data collection process may be welcome in these cases. In longitudinal surveys, there is a wealth of potential auxiliary variables to deal with wave nonresponse. DCP information may thus not have the same importance to compensate for wave nonresponse than the importance it has to compensate for unit nonresponse in cross-sectional surveys. However, we have yet to study this in any depth. It may turn out that, at change points, DCP variables may matter greatly.

In section 2, we introduce notation and theory concerning the effect of using random auxiliary variables in the nonresponse model when estimating population totals. This issue of the randomness of DCP auxiliary variables was raised and debated at Statistics Canada's Advisory Committee on Statistical Methods after the paper by Alavi and Beaumont (2004) was presented. The goal of section 2 is thus to shed some light on this issue. The use of DCP variables to adjust design weights for nonresponse is briefly illustrated in section 3, using the Canadian Labour Force

1. Jean-François Beaumont, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.ca.

Survey (CLFS). The last section, section 4, contains a brief summary of the paper.

2. Theory

Let us assume that we are interested in estimating the population total $t_y = \sum_{k \in U} y_k$ of a variable of interest y for a certain fixed population U of size N. From this population, a random sample s of size n is selected according to a probability sampling design $p(s|\mathbf{D})$, where $\mathbf{D}$ is a $N-$row matrix containing $\mathbf{d}_k'$ in its k^{th} row and $\mathbf{d}$ is the vector of design variables. Let also assume that, in the absence of nonresponse, we would use the Horvitz-Thompson estimator $\hat{t}_y = \sum_{k \in s} w_k y_k$, where $w_k = 1/\pi_k$ is the design weight of unit k and $\pi_k = P(k \in s)$ is its selection probability.

Usually, due to a number of reasons, unit nonresponse occurs so that the variable y is only observed for a subset s_r of s, the respondents. Along with s_r, a random vector $\mathbf{z}$ of DCP variables is also observed for every sample unit according to a joint mechanism $\#q(\mathbf{Z}_s, s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$. As mentioned in the introduction, the number of attempts to contact a sample unit is an example of a DCP variable. The vector $\mathbf{z}$ of DCP variables and the set of respondents s_r are random after conditioning on the selected sample since these quantities would likely take different values if the data collection process were repeated for a given sample. The quantity $\mathbf{Z}_s$ is a $n-$row matrix containing $\mathbf{z}_k'$ in its k^{th} row, $\mathbf{Y}$ is a $N-$element vector containing y_k in its k^{th} element and $\mathbf{X}$ is a $N-$row matrix containing $\mathbf{x}_k'$ in its k^{th} row. The vector $\mathbf{x}$ is a vector of additional fixed auxiliary variables. For instance, these auxiliary variables could come from an administrative file or, in a longitudinal survey, they could be the variables of interest observed at the previous wave. As a result, the vector $\mathbf{x}$ may not be available for nonsample units. Table 1 summarizes the availability of the different types of variables for the respondents, nonrespondents and nonsample units.

Table 1
Availability of Variables

	y	z	x	d
Respondents: s_r	YES	YES	YES	YES
Nonrespondents: $s - s_r$	NO	YES	YES	YES
Nonsample units: $U - s$	NO	NO *	YES **	YES

* The vector $\mathbf{z}$ is not even defined for nonsample units.
** The vector $\mathbf{x}$ may not always be available for nonsample units.

The joint mechanism $\#q(\mathbf{Z}_s, s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ can be factorized into two distinct random mechanisms: i) $\#(\mathbf{Z}_s | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ and ii) $q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$. The former is called the DCP mechanism while the latter is called the nonresponse mechanism. This factorization will be useful later to obtain properties of our nonresponse-weight-adjusted estimator defined in equation (2.2) below. We assume that

$$q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s) = q(s_r | s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s), \quad (2.1)$$

where $\mathbf{D}_s$ and $\mathbf{X}_s$ are the sample portions of $\mathbf{D}$ and $\mathbf{X}$ respectively. This assumption implies that the nonresponse mechanism is independent of (or unconfounded with) $\mathbf{Y}$, after conditioning on s, $\mathbf{D}_s$, $\mathbf{X}_s$ and $\mathbf{Z}_s$, and that the data are missing at random. However, we make no explicit simplifying assumption about the DCP mechanism so that it may well depend on $\mathbf{Y}$, even after conditioning on $s, \mathbf{D}$ and $\mathbf{X}$.

To compensate for unit nonresponse, we consider the nonresponse-weight-adjusted estimator

$$\hat{t}_y^{\text{NWA}} = \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k, \quad (2.2)$$

where $p_k(\boldsymbol{\alpha}) = P(k \in s_r | s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s; \boldsymbol{\alpha})$ is the conditional response probability for a unit $k \in s$ and $\hat{\boldsymbol{\alpha}}$ is an estimator of the vector of unknown nonresponse model parameters $\boldsymbol{\alpha}$. Note that a nonresponse model is a set of assumptions about the unknown nonresponse mechanism $q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$; one of them being assumption (2.1). We assume that $\hat{\boldsymbol{\alpha}}$ is implicitly defined by the equation $\mathbf{U}_1(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, where $\mathbf{U}_1(.)$ is a vector of $q-$unbiased estimating functions for $\boldsymbol{\alpha}$; that is, $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = \mathbf{0}$. Therefore, $\mathbf{U}_1(.)$ is also $p\#q-$unbiased for $\boldsymbol{\alpha}$. In the remaining of the paper, we remove everywhere the conditioning on $\mathbf{Y}$, $\mathbf{D}$ and $\mathbf{X}$ when taking expectations and variances since these vectors are always treated as being fixed. For instance, we will write $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) | s, \mathbf{Z}_s\} = \mathbf{0}$ instead of $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = 0$. This will simplify considerably the notation.

Note that the nonresponse-weight-adjusted estimator (2.2) is implicitly defined by the equation

$$U_2(\hat{\boldsymbol{\alpha}}, \hat{t}_y^{\text{NWA}}) = \hat{t}_y^{\text{NWA}} - \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k = 0. \quad (2.3)$$

If the nonresponse model is correctly specified and, in particular, if assumption (2.1) is satisfied, then the estimating function $U_2(.,.)$ is $p\#q-$unbiased for t_y; that is, $\mathrm{E}_{p\#q}\{U_2(\boldsymbol{\alpha}, t_y)\} = 0$. To make assumption (2.1) as plausible as possible, it is important that the nonresponse model be conditional on design, auxiliary and DCP variables that are well correlated with y, provided that these variables are also associated with nonresponse. This recommendation should be useful to control the magnitude of the nonresponse bias, which may be unavoidable in real surveys.

This is also in line with the recommendation given in Little and Vartivarian (2005). Therefore, if DCP variables contain information about y above the information already contained in $\mathbf{d}$ and $\mathbf{x}$, then the use of DCP variables may be useful to reduce the nonresponse bias if they are associated with nonresponse.

Now, let $\boldsymbol{\theta}=(\boldsymbol{\alpha}', t_y)'$, $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\alpha}}', \hat{t}_y^{\mathrm{NWA}})'$ and $\mathbf{U}(\tilde{\boldsymbol{\theta}})=\{\mathbf{U}_1'(\tilde{\boldsymbol{\alpha}}), U_2(\tilde{\boldsymbol{\alpha}}, \tilde{t}_y)\}'$, for some vector $\tilde{\boldsymbol{\theta}}=(\tilde{\boldsymbol{\alpha}}', \tilde{t}_y)'$. As noted above, $\hat{\boldsymbol{\theta}}$ is implicitly defined by the equation $\mathbf{U}(\hat{\boldsymbol{\theta}})=\mathbf{0}$ and the estimating function $\mathbf{U}(.)$ is $p\#q-$ unbiased for $\boldsymbol{\theta}$ since $\mathbf{E}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\}=\mathbf{0}$. Using a first-order Taylor approximation (see Binder 1983), we have $\hat{\boldsymbol{\theta}}\approx \boldsymbol{\theta}-\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{U}(\boldsymbol{\theta})$, where $\mathbf{H}(\tilde{\boldsymbol{\theta}})=\mathbf{E}_{p\#q}\{\partial\mathbf{U}(\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}'\}$. The matrix $\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}$ is thus given by

$$\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}=\begin{pmatrix}\{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & \mathbf{0}\\ -\mathbf{H}_{21}(\boldsymbol{\theta})\{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & 1\end{pmatrix}, \qquad (2.4)$$

where $\mathbf{H}_{i1}(\tilde{\boldsymbol{\theta}})=\mathbf{E}_{p\#q}(\partial\mathbf{U}_i(\tilde{\boldsymbol{\theta}}))/(\partial\tilde{\boldsymbol{\alpha}}')$, for $i=1, 2$. Using conditions similar to those of Binder (1983), $\hat{\boldsymbol{\theta}}$ is asymptotically normal and asymptotically $p\#q-$ unbiased for $\boldsymbol{\theta}$. As a result, $\hat{t}_y^{\mathrm{NWA}}$ is asymptotically normal and asymptotically $p\#q-$ unbiased for t_y. Therefore, using DCP variables in the nonresponse model does not introduce any bias in the nonresponse-weight-adjusted estimator $\hat{t}_y^{\mathrm{NWA}}$ provided that the nonresponse model (specification of $q(s_r \mid s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s)$ and assumption 2.1) holds. Also, if the true unknown nonresponse mechanism depends on the sample portion of $\mathbf{Y}$, $\mathbf{Y}_s$, after conditioning on s, $\mathbf{D}_s$ and $\mathbf{X}_s$, then conditioning on a vector $\mathbf{z}$ of DCP variables is likely to reduce the nonresponse bias if the DCP mechanism depends on $\mathbf{Y}_s$, after conditioning on s, $\mathbf{D}_s$ and $\mathbf{X}_s$, which means that the DCP variables contain information about y not already contained in $\mathbf{d}$ and $\mathbf{x}$.

Continuing our Taylor linearization, and using the fact that

$$\begin{aligned}\mathbf{V}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\}=&\mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta})\mid s\}\\ &+\mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s,\mathbf{Z}_s\}\\ &+\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s,\mathbf{Z}_s\},\end{aligned}$$

the $p\#q-$ variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{p\#q}(\hat{\boldsymbol{\theta}})$, is approximated by

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}})=&\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta})\mid s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}\\ &+\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s,\mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}\\ &+\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s,\mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \qquad (2.5)\end{aligned}$$

The first term on the right-hand side of equation (2.5) is called the sampling variance of $\hat{\boldsymbol{\theta}}$, the second term is called the DCP variance of $\hat{\boldsymbol{\theta}}$ and the third term is called the nonresponse variance of $\hat{\boldsymbol{\theta}}$. The variance $\mathrm{V}_{p\#q}(\hat{t}_y^{\mathrm{NWA}})$ is approximated by the value in the last row and in the last column of equation (2.5). Using expression (2.4) and the fact that $\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s, \mathbf{Z}_s\}=(\mathbf{0}', t_y-\hat{t}_y)'$, the approximate variance (2.5) reduces to

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}})=&\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathrm{V}_p(\hat{t}_y)\end{pmatrix}+\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & 0\end{pmatrix}\\ &+\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta})\mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \qquad (2.6)\end{aligned}$$

The second matrix on the right-hand side of equation (2.6) corresponds to the DCP variance of $\hat{\boldsymbol{\theta}}$ and contains 0 for all its elements. Therefore, using random auxiliary (DCP) variables in the nonresponse model does not introduce any additional term of variance, as opposed to using only fixed auxiliary variables, when the nonresponse model is properly specified. Since DCP variables are likely to reduce the nonresponse bias if they are associated with y, then it seems beneficial to take advantage of them when handling unit nonresponse through a weight adjustment. Also, as pointed out by Little and Vartivarian (2005), adding auxiliary variables in the nonresponse model that are associated with y tends to reduce the nonresponse variance. The mean squared error can therefore be reduced on both counts.

A more detailed expression for the nonresponse variance term in equation (2.6) as well as a sampling and a nonresponse variance estimator can be obtained similarly as in Beaumont (2005). Beaumont (2005) also discusses the effect of estimating the nonresponse model parameters on the variance of an estimator of a population total.

3. The Example of the Canadian Labour Force Survey

The goal of this example is not to provide every detail of the analysis that was conducted on the Canadian Labour Force Survey (CLFS) data but simply to describe some issues related to the choice of the nonresponse model and to the estimation of response probabilities. With these points in mind, we then go on to discuss the main conclusions that were reached. Greater detail about the results of the investigations in the CLFS, implementation of the new method and a comparison with the previous method can be found in Alavi and Beaumont (2004).

The CLFS is a monthly survey with a stratified multistage sampling design (Gambino, Singh, Dufour, Kennedy and Lindeyer 1998). The information used to construct the sampling design and to draw a sample of dwellings is essentially geographic. The sample is divided into six representative rotation groups and each sampled dwelling stays in the sample for six consecutive months. One rotation group contains dwellings for which the members are interviewed

for the first time; another rotation group contains dwellings for which the members are interviewed for the second time and so on. Thus, for five rotation groups out of six, the sampled dwellings are common from one month to the next. Computer-assisted interviews are used to collect the survey information for every person in the selected households. With computer-assisted interviews, a large amount of DCP information is obtained for both responding and nonresponding households.

A logistic nonresponse model has been considered to model the unknown nonresponse mechanism $q(s_r \mid s, \mathbf{D}_s, \mathbf{Z}_s)$. With this model, the unknown response probability for household k is expressed as $p_k(\boldsymbol{\alpha}) = \{1 + \exp(-\boldsymbol{\alpha}'(\mathbf{zd})_k)\}^{-1}$ and sampled households are assumed to respond independently of one another. The vector $\mathbf{zd}$ is a vector that contains DCP variables $\mathbf{z}$, fixed design variables $\mathbf{d}$ as well as interactions between these two types of variables. No additional vector $\mathbf{x}$ of auxiliary variables was available. Two DCP variables were used: the number of attempts to contact a sampled household, which was divided into five categories, and the time of the last attempt, which was also divided into five categories. The design variables used were mainly geographic and also included the rotation group indicator. Due to potential interviewer and clustering effects, the above model may not be entirely realistic. It was used for its simplicity and because it appeared reasonable and an improvement over the previous method. Also, the estimated response probabilities resulting from this model were used only to provide a score and were not used directly to adjust design weights, as described below in this section.

The unknown vector $\boldsymbol{\alpha}$ was estimated by the maximum likelihood method using the $q-$unbiased estimating function

$$\mathbf{U}_1(\boldsymbol{\alpha}) = \sum_{k \in s} \{r_k - p_k(\boldsymbol{\alpha})\}(\mathbf{zd})_k, \qquad (3.1)$$

where $r_k = 1$, if $k \in s_r$, and $r_k = 0$, otherwise. Note that a design-weighted estimating function was not considered. This follows the practice recommended in Little and Vartivarian (2003) and can be justified by noting that the interest is in modelling the nonresponse mechanism only for sampled households $k \in s$ (not for the whole population) and that this mechanism is conditional on *s*. Also, the DCP variables are not even defined outside the sample. The use of design weights does thus not make sense in this context and increases the variance of $\hat{\boldsymbol{\alpha}}$ if the nonresponse model is correctly specified. Also, it is not clear that using a design-weighted estimating function would systematically bring robustness in this case. However, note that we do not ignore design information since it is included in the nonresponse model. This can be paralleled to the recommendation of including design information in imputation models (see, for example, Rubin 1996).

Stepwise logistic regression was performed for several months in order to determine appropriate design and DCP variables to be included in the final nonresponse model. In all months considered, the variable 'number of attempts' was the first to enter in the model and thus the most useful for explaining nonresponse. This variable was also correlated with the main variables of interest 'employment' and 'unemployment'. For instance, people belonging to respondent households with a large number of attempts, i.e. those that are difficult to reach, had a tendency to be more often employed (see Alavi and Beaumont 2004). Households with a large number of attempts had also a tendency to be nonrespondents. Therefore, it seems appropriate to give a larger weight adjustment to the responding households for which the number of attempts is large since their propensity to respond is lower and they are more likely to have characteristics similar to the nonrespondents.

The final nonresponse model chosen fit reasonably well the CLFS data in most months considered, according to the Hosmer-Lemeshow goodness-of-fit test. Nevertheless, the score method of Little (1986) was used to obtain some robustness against undetected model failures. The above logistic nonresponse model was first used to obtain an estimated response probability for every sampled household and then the sample was divided into about 50 homogeneous classes with respect to this estimated response probability using the clustering algorithm implemented in the procedure FASTCLUS of SAS. This large number of classes was possible given the large CLFS sample size. It was chosen so as to reduce the nonresponse bias not only at the population level but also for smaller domains. The nonresponse weight adjustment for a responding household *k* within a given class *c* was simply computed as the inverse of the unweighted response rate within class *c*. A threshold on the nonresponse weight adjustment was set to 2.5 to control the nonresponse variance of the nonresponse-weight-adjusted estimator. When needed, the application of this threshold was necessary only for a very small number of classes. These were the classes with the smallest estimated response probabilities. Without this threshold, nonresponse weight adjustments around 4 could occasionally be observed.

Another nonresponse model was considered in which the response probability for a household *k* is modelled as the product of the probability that household *k* be contacted, times the probability that this household respond, given it is contacted. The latter two probabilities were modelled separately. Although this model seems to be a better approximation of reality and gave slightly better results in the sense that it better explained nonresponse, the gains were not deemed sufficient to add this complexity in the nonresponse adjustment method. It may deserve further study.

4. Conclusion

An important contribution of this paper is that DCP information must be treated as being random when used in a nonresponse model. We then have shown that the use of such information to handle unit nonresponse through a weight adjustment does not introduce any bias and that there is no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, we have argued that if DCP information is associated with the variables of interest and with nonresponse, then its use is likely to reduce the nonresponse bias when the nonresponse mechanism depends directly on the variables of interest. We have also illustrated through the CLFS example that such information can be useful for dealing with unit nonresponse in a major survey.

The full response estimator that we have considered is the Horvitz-Thompson estimator. Our conclusions would have remained the same had we used instead a generalized regression estimator. We have used the Horvitz-Thompson estimator for its simplicity and because it was sufficient to show our main point.

Acknowledgements

I would like to thank the members of Statistics Canada's Advisory Committee on Statistical Methods for raising issues with the application of the proposed method in the Labour Force Survey and, particularly, J.N.K. Rao and Chris Skinner for useful discussions following the presentation to the Committee. I would also like to sincerely thank the Associate Editor for his comments and suggestions. They were very useful and improved the clarity of the paper. Finally, I am much indebted to Asma Alavi and Cynthia Bocci of Statistics Canada for preparing computer programs used to analyse Labour Force Survey data.

References

Alavi, A., and Beaumont, J.-F. (2004). Nonresponse adjustment plans for the Labour Force Survey. Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, May 2-3, 2004.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, Series B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Couper, M., and Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute* (to appear).

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 7, 325-337.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.

Holt, D., and Elliott, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 40, 333-342.

Little, R.J. (1986). Survey nonresponse adjustment for estimate of means. *International Statistical Review*, 54, 139-157.

Little, R.J., and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology* , 31, 161-168.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 233-238
Statistics Canada, Catalogue No. 12-001-XPB

On the Correlation Structure of Sample Units

Alfredo Bustos [1]

Abstract

In this paper we make explicit some distributional properties of sample units, not usually found in the literature; in particular, their correlation structure and the fact that it does not depend on arbitrarily assigned population indices. Such properties are relevant to a number of estimation procedures, whose efficiency would benefit from making explicit reference to them.

Key Words: Census; Survey; Sampling; Sample units; Probability function; Mean; Covariance.

1. Introduction

In recent times, population and household censuses, as we know them, have become more difficult to perform for a number of reasons. Alternative ways of securing more frequent information for the production of local, state and national statistical results have been proposed. Continuous large national surveys, among them those known as rolling censuses, with large sample sizes and complex designs, are being considered.

However, in order to produce results at the local authority level the way a census does, different techniques for estimation as well as for validation and, in some cases, for imputation have to be developed and their efficiency improved. One way of achieving greater efficiency consists of taking into account all relevant information available. Of course, this includes the stochastic properties of sample units.

In what follows, beginning from basic principles, we derive a general explicit form for the probability function of an ordered sample. We also show how that function, as well as the inclusion probabilities, can be computed. Finally, we give a general form for the correlation matrix of sample units, which depends solely on inclusion probabilities, so that linear and maximum-likelihood estimation procedures can benefit from it.

2. The Basic Model

The basic model we start from represents the sequential random drawing of n units from a population U formed by N such units, and may be stated as follows. Let N and n be two positive constants such that $n \le N$, and let V represent an $N \times n$ matrix, whose components are each distributed as Bernoulli random variables with, possibly, different parameters. Then,

$$V_{N\times n} = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \vartheta_{13} & \cdots & \vartheta_{1n} \\ \vartheta_{21} & \vartheta_{22} & \vartheta_{23} & \cdots & \vartheta_{2n} \\ \vartheta_{31} & \vartheta_{32} & \vartheta_{33} & \cdots & \vartheta_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{N1} & \vartheta_{N2} & \vartheta_{N3} & \cdots & \vartheta_{Nn} \end{bmatrix}. \qquad (1.1)$$

Also part of the model is the restriction imposed on each column of V to add to one. In other words, we require that

$$\sum_{I=1}^{N} \vartheta_{Ik} = 1, \text{ for } k = 1, ..., n \qquad (1.2)$$

be satisfied.

This is required because if the j^{th} draw results in population unit I being selected, then entry (I, j) takes the value of one while all other entries of column j are equal to zero. Note that this is equivalent to imposing a non-stochastic constraint on the behavior of all components of the i^{th} column of V, regardless of the sampling scheme. Therefore, entries belonging to the same column do not behave independently.

When sampling takes place with replacement (WR), the sum of the elements of the I^{th} row of the above matrix is distributed as a Binomial (n, p_I) since each column is distributed independently of other columns. On the other hand, when sampling takes place without replacement (WOR), the total of row I can take only two values: one, if the I^{th} unit is drawn at some stage, or zero, otherwise, bringing us back to the Bernoulli case.

Disjoint subsets of rows may be formed according to different criteria. For instance, when rows are grouped with regard to their spatial vicinity, one could speak about clusters or primary sampling units. When one or more statistical indicators form the basis for the groupings, the term strata is usually used.

1. Victor Alfredo Bustos y de la Tijera, Instituto Nacional de Estadística, Geografia e Informática, H. de Nacozari 2301, 20270, Aguascalientes, Ags., México. E-mail: alfredo.bustos@inegi.gob.mx.

Let us now define the inclusion probabilities as

$$\pi_I^{(k)} = P(\text{population unit } I \text{ in sample of size } k) \\ = 0 \text{ if } k = 0. \tag{2}$$

Note that $\pi_I^{(n)} = \pi_I$, commonly refered to as the inclusion probability for unit *I*.

Now let $\underline{\vartheta}_{\circ j}$ represent the j^{th} column and $\underline{\vartheta}_{I\circ}$ the I^{th} row of matrix *V*. Therefore, based on the following expression,

$$f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) = f(\underline{\vartheta}_{\circ 1}) f(\underline{\vartheta}_{\circ 2} \mid \underline{\vartheta}_{\circ 1}) \\ f(\underline{\vartheta}_{\circ 3} \mid \underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}) ... f(\underline{\vartheta}_{\circ n} \mid \underline{\vartheta}_{\circ 1}, ..., \underline{\vartheta}_{\circ n-1}) \tag{3}$$

we can write the joint probability function of the elements of *V* as:

$$f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) = \prod_{k=1}^{n} \left[\prod_{I=1}^{N} (\pi_I^{(k)} - \pi_I^{(k-1)})^{\vartheta_{Ik}} \right] \\ = \prod_{k=1}^{n} \left[\prod_{I=1}^{N} (p_I^{(k)})^{\vartheta_{Ik}} \right] \tag{4}$$

subject to

$$\sum_{I=1}^{N} \vartheta_{Ik} = 1, k = 1, \ldots, n \text{ and}$$

$$\sum_{k=1}^{N} \vartheta_{Ik} \le \begin{cases} 1, \text{ WOR} \\ n, \text{ WR} \end{cases} I = 1, \ldots, N;$$

and here $p_I^{(k)}$, defined as $p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)})$, stands for the probability that population unit *I* is included in the sample at the k^{th} draw. The above function is useful for calculating the probability of any ordered sample of size *n*. Clearly, when the order of inclusion can be ignored, the probability of a given sample would be obtained by adding the n! values obtained through (4).

3. The Implications of Sampling on the Stochastic Properties of Population Units

Consequently,

$$E(\vartheta_{Ik}) = p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)}) \tag{5}$$

and therefore, we can write

$$E[V] = \begin{bmatrix} p_1^{(1)} & p_1^{(2)} & p_1^{(3)} & \cdots & p_1^{(n)} \\ p_2^{(1)} & p_2^{(2)} & p_2^{(3)} & \cdots & p_2^{(n)} \\ p_3^{(1)} & p_3^{(2)} & p_3^{(3)} & \cdots & p_3^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_N^{(1)} & p_N^{(2)} & p_N^{(3)} & \cdots & p_N^{(n)} \end{bmatrix}. \tag{6}$$

From here, step-by-step inclusion probabilities, in WOR sampling situations, may be recursively computed, as is shown in (7), below.

$$p_I^{(k)} = \begin{cases} p_I & \text{if } k = 1 \\ p_I^{(k-1)} \sum_{J \neq I}^{N} \dfrac{p_J^{(k-1)}}{1 - p_J^{(k-1)}} & \text{if } k > 1. \end{cases} \tag{7}$$

Note that (7) enables us to compute the desired probabilities at two different moments: first, when no draw has actually occurred, which explains why we average over the whole population, and secondly, when the result of the previous draw is known, at which time the probability of the J^{th} population unit, say, entering the sample equals one and all other probabilities for that draw are equal to zero. Hence, at least in theory, we can compute the inverse of the so called expansion factors or weights for one stage sampling, or stage by stage in multistage sampling. Clearly,

$$\pi_I^{(n)} = \sum_{k=1}^{n} p_I^{(k)}. \tag{8}$$

If we define the joint inclusion probabilities as

$$\pi_{IJ}^{(k)} = P\left(\begin{matrix} \text{population units } I \text{ and} \\ J \text{ in sample of size } k \end{matrix} \right), \tag{9}$$

then we have that they can also be computed as follows:

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right). \tag{10}$$

For example, in simple random sampling WR (SRS/WR), expressions (7), (8) and (10) result in (7.1), (8.1) and (10.1),

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \ge 1 \tag{7.1}$$

$$\pi_I^{(n)} = \frac{n}{N} \tag{8.1}$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \\ = \sum_{j=1}^{n-1} \left(\frac{n-j}{N^2} + \frac{n-j}{N^2} \right) = \frac{n(n-1)}{N^2}. \tag{10.1}$$

While in SRS/WOR we get expressions (7.2), (8.2) and (10.2), instead.

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \ge 1 \tag{7.2}$$

$$\pi_I^{(n)} = \frac{n}{N} \tag{8.2}$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \text{ where } J \neq I$$

$$= \sum_{j=1}^{n-1}\left(\frac{n-j}{N(N-1)} + \frac{n-j}{N(N-1)} \right) = \frac{n(n-1)}{N(N-1)}. \tag{10.2}$$

Let us now consider the row vectors $\underline{\vartheta}_{I\circ}$. Then, for the covariance matrix between different rows, we get

$$\text{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ}) = \begin{bmatrix} -p_I^{(1)} p_J^{(1)} & 0 & \cdots & 0 \\ 0 & -p_I^{(2)} p_J^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -p_I^{(n)} p_J^{(n)} \end{bmatrix}_{n\times n} \tag{11}$$

whenever I is different from J.

When sampling takes place WR, and therefore, $p_I^{(j)} = p_I \ \forall\ j = 1, ..., n,$ the covariance matrix for the I^{th} row vector is given by

$$\text{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I q_I & 0 & 0 & \cdots & 0 \\ 0 & p_I q_I & 0 & \cdots & 0 \\ 0 & 0 & p_I q_I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_I q_I \end{bmatrix}_{n\times n}. \tag{12.1}$$

In a WOR setting the above covariance matrix becomes

$$\text{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I^{(1)}(1-p_I^{(1)}) & -p_I^{(1)} p_I^{(2)} & \cdots & -p_I^{(1)} p_I^{(n)} \\ -p_I^{(1)} p_I^{(2)} & p_I^{(2)}(1-p_I^{(2)}) & \cdots & -p_I^{(2)} p_I^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_I^{(1)} p_I^{(n)} & -p_I^{(2)} p_I^{(n)} & \cdots & p_I^{(n)}(1-p_I^{(n)}) \end{bmatrix}_{n\times n}. \tag{12.2}$$

Let $\underline{\vartheta}$ represent the N–dimensional vector which results from adding the columns of V. Clearly, the components of this vector may be expressed as the product of $\underline{\vartheta}_{I\circ}$ by a vector whose components are all equal to one. In other words,

$$\underline{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vdots \\ \vartheta_N \end{pmatrix} = \begin{pmatrix} \underline{\vartheta}_{1\circ}^T \underline{1} \\ \underline{\vartheta}_{2\circ}^T \underline{1} \\ \underline{\vartheta}_{3\circ}^T \underline{1} \\ \vdots \\ \underline{\vartheta}_{N\circ}^T \underline{1} \end{pmatrix}. \tag{13}$$

Some distributional properties of these sums may be then obtained directly from those of the rows or the columns of matrix V.

For instance, their expected values are given as

$$E(\vartheta_I) = E(\underline{\vartheta}_{I\circ}^T \underline{1}) = E\left(\sum_{k=1}^{n} \vartheta_{Ik} \right)$$

$$= \sum_{k=1}^{n} p_I^{(k)} = \pi_I^{(1)} + \sum_{k=2}^{n} (\pi_I^{(k)} - \pi_I^{(k-1)}) = \pi_I^{(n)}. \tag{14}$$

From (1.2), we get the non-stochastic restriction:

$$\underline{1}'\,\underline{\vartheta} = \vartheta_1 + \vartheta_2 + \vartheta_3 + ... + \vartheta_N = n. \tag{15}$$

From (14) and (15), well known propositions (16) and (17) follow immediately,

$$E[\underline{\vartheta}'] = (\pi_1^{(n)}, \pi_2^{(n)}, \pi_3^{(n)}, \ldots, \pi_N^{(n)}) \tag{16}$$

$$\pi_1^{(n)} + \pi_2^{(n)} + \pi_3^{(n)} + \ldots + \pi_N^{(n)} = n. \tag{17}$$

For the second order moments, we get

$$\text{Cov}(\vartheta_I, \vartheta_J) = \text{Cov}(\underline{1}'\,\underline{\vartheta}_{I\circ}, \underline{1}'\,\underline{\vartheta}_{J\circ})$$

$$= \underline{1}'\,\text{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ})\underline{1} = -\sum_{k=1}^{n} p_I^{(k)} p_J^{(k)}$$

$$= \begin{cases} -np_I p_J & \text{WR} \\ (\pi_{IJ}^{(n)} - \pi_I^{(n)} \pi_J^{(n)}) & \text{WOR}, \end{cases} \tag{18}$$

which clearly indicates that the covariance is never positive. In turn, the variances are given by

$$\text{Var}(\vartheta_I) = \text{Var}(\underline{1}'\,\underline{\vartheta}_{I\circ}) = \underline{1}'\,\text{Cov}(\underline{\vartheta}_{I\circ})\underline{1}$$

$$= \begin{cases} np_I q_I & \text{WR} \\ \pi_I^{(n)}(1-\pi_I^{(n)}) & \text{WOR}. \end{cases} \tag{19}$$

Another important consequence of (15) has to do with the second order moments of the stochastic vector $\underline{\vartheta}$.

$$0 = \text{Var}(n) = \text{Var}(\underline{1}'\,\underline{\vartheta}) = \underline{1}'\text{Cov}(\underline{\vartheta})\underline{1} = \underline{1}'C\underline{1}. \tag{20}$$

Clearly, the diagonal elements of matrix C, the covariance matrix of $\underline{\vartheta}$, are not all equal to zero. Therefore, randomly drawing a fixed-size simple introduces a dependency in the population units which results in non-null covariances implying that matrix C is singular. Otherwise, it is impossible for (20) to be satisfied.

As a matter of fact, it is possible to prove that the sum of any row (or column) of C must be equal to zero, which is a stronger statement. Given that the covariance between a random variable and a constant equals zero, we get

$$0 = \mathrm{Cov}(\vartheta_I, n) = \mathrm{Cov}(\vartheta_I, \vartheta_1 + \vartheta_2 + \vartheta_3 + \cdots + \vartheta_N)$$
$$= C_{I1} + C_{I2} + \cdots + C_{IN}$$
$$= \mathrm{Var}(\vartheta_I) + \sum_{J \neq I} \mathrm{Cov}(\vartheta_I, \vartheta_J). \quad (21)$$

We have thus proven that in WOR sampling (22.1) holds.

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}). \quad (22.1)$$

The same statement can be proven algebraically by noting that

$$\sum_{J \neq I} \pi_{IJ}^{(n)} = \pi_I^{(n)} \sum_{J \neq I} \pi_{J|I}^{(n)}$$
$$= (n-1)\pi_I^{(n)},$$

which is obvious once we realize that the conditional probability involved represents the probability that population unit J enters a sample of size $n-1$ for which (19) also applies. Additionally, using (19) again, note that

$$\sum_{J \neq I} \pi_J^{(n)} = (n - \pi_I^{(n)}),$$

and therefore,

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)})$$
$$= \pi_I^{(n)} - (\pi_I^{(n)})^2 + (n-1)\pi_I^{(n)} - \pi_I^{(n)}(n - \pi_I^{(n)}).$$

For WR sampling (21) implies:

$$0 = np_I q_I + \sum_{J \neq I} (n(n-1)p_I p_J - n^2 p_I p_J)$$
$$= np_I q_I - np_I \sum_{J \neq I} p_J \quad (22.2)$$

which is immediately seen to apply.

In any case, the most important implication of the above results is that regardless of the sampling scheme, the correlation matrix of the population random variables $\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_N$ is singular. For the practical situations described in the introduction, the most important implication of this fact lies mainly in the use made by many model fitting and estimation procedures of the inverse of the covariance matrix.

4. The First Two Moments of Sample Units

Once the first and second order moments of the vector $\underline{\vartheta}$ have been established, we are in a position to determine the corresponding moments for sub-vectors of different sizes and whose components are randomly chosen, *i.e.*, the sample. To this end, let us define the random variables $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r}$, where r represents the number of different population units in the sample, and whose indices $I_k, 1 \leq k \leq r \leq n,$ can take the value I with probability $\pi_I^{(n)}$. In other words, under the above conditions, we are in the presence of a set of random variables whose indices are random themselves.

4.1 Mean and Variance for WR Sampling

For this case, the probability function of ϑ_{I_i} is given by

$$P(\vartheta_{I_j} = x) = \sum_{I=1}^{N} p_I P(\vartheta_I = x)$$
$$= \sum_{I=1}^{N} p_I \binom{n}{x} p_I^x (1 - p_I)^{n-x}. \quad (23)$$

The first two moments may also be obtained via a conditional argument. The mean of its distribution is given by

$$E(\vartheta_{I_j}) = \sum_{I=1}^{N} p_I E(\vartheta_I) = \sum_{I=1}^{N} np_I\, p_I = n\sum_{I=1}^{N} p_I^2. \quad (24)$$

In turn, its variance is computed using the well known formula

$$V(\vartheta_{I_j}) = V_{I_j}[E(\vartheta_{I_j} \mid I_j)] + E_{I_j}[V(\vartheta_{I_j} \mid I_j)]. \quad (25)$$

In this case, we have

$$E(\vartheta_{I_j} \mid I_j = I) = np_I$$
$$\text{and } V(\vartheta_{I_j} \mid I_j = I) = np_I(1 - p_I). \quad (26)$$

Hence,

$$V_{I_j}[E(\vartheta_{I_j} \mid I_j)] = V_{I_j}(np_{I_j})$$
$$= n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})],$$
$$E_{I_j}[V(\vartheta_{I_j} \mid I_j)] = nE_{I_j}[p_{I_j}(1 - p_{I_j})]$$
$$= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] \quad (27)$$

and therefore

$$V(\vartheta_{I_j})$$
$$= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] + n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})]$$
$$= \sum_{I=1}^{N} np_I^2 \left(1 + (n-1)p_I - \sum_{J=1}^{N} np_J^2\right). \quad (28)$$

For the case of SRS, (24) above results in

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2 = \frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2 \frac{N}{n} = \frac{n}{N}.$$

While (28) yields

$$V(\vartheta_{I_j}) = \sum_{I=1}^{N} n\frac{1}{N^2}\left(1 + (n-1)\frac{1}{N} - \sum_{J=1}^{N} n\frac{1}{N^2}\right) = n\frac{1}{N}\left(1 - \frac{1}{N}\right).$$

4.2 Mean and Variance for WOR Sampling

For this case, the probability function of ϑ_{I_i} is given by

$$P(\vartheta_{I_j} = x) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}\sum_{k=1}^{n}(p_I^{(k)})^x(1 - p_I^{(k)})^{1-x} \quad (29)$$

and therefore

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}E(\vartheta_I) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}\sum_{j=1}^{n}(p_I^{(j)}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2. \quad (30)$$

Using (25) again, we note firstly that

$$E(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)} \text{ and } V(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)}(1-\pi_{I_j}^{(n)})$$

from which we get

$$V[E(\vartheta_{I_j} \mid I_j)] = V(\pi_{I_j}^{(n)}) = E[(\pi_{I_j}^{(n)})^2] - [E(\pi_{I_j}^{(n)})]^2$$

and

$$E[V(\vartheta_{I_j} \mid I_j)] = E[\pi_{I_j}^{(n)}(1-\pi_{I_j}^{(n)})] = E[(\pi_{I_j}^{(n)})] - [E(\pi_{I_j}^{(n)})^2].$$

Hence, the variance is given by

$$V(\vartheta_{I_j}) = E(\pi_{I_j}^{(n)}) - E^2(\pi_{I_j}^{(n)}) = E(\pi_{I_j}^{(n)})[1-E(\pi_{I_j}^{(n)})] = \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\left[1-\left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\right]. \quad (31)$$

Once again, in order to exemplify these results, let us turn to SRS. Expression (30) becomes

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2 = \frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2\frac{N}{n} = \frac{n}{N}. \quad (32)$$

Whereas (31) results in

$$V(\vartheta_{I_j}) = \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\left[1-\left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\right] = \frac{n}{N}\left(1-\frac{n}{N}\right). \quad (33)$$

4.3 The Covariance Between Sample Units

In order to establish the covariance between different sample units we resort to a simple extension to (25),

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \mathrm{Cov}_{I_j, I_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)] + E_{I_j, I_j}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)]. \quad (34)$$

In this case, we have that

$$E(\vartheta_{I_j} \mid I_j = I) = \pi_I^{(n)} \quad (35)$$

and

$$E(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j = I, I_k = J) = \pi_{IJ}^{(n)} \quad (36)$$

while the covariance between brackets on the right-hand side of (34) is easily seen to equal

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j = I, I_k = J) = \pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_j^{(n)}. \quad (37)$$

From (35) and (36), we obtain

$$\mathrm{Cov}_{I_jI_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)] = E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}) - E_{I_j}(\pi_{I_j}^{(n)})E_{I_k}(\pi_{I_k}^{(n)}) \quad (38)$$

whereas from (37) we get

$$E_{I_jI_k}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)] = E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}). \quad (39)$$

Finally, adding these last two expressions we arrive at the desired covariance

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - [E_{I_j}(\pi_{I_j}^{(n)})][E_{I_k}(\pi_{I_k}^{(n)})] = \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{\substack{J=1\\J\neq I}}^{N}(\pi_{IJ}^{(n)})^2 - \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)^2. \quad (40)$$

In the SRS/WR (40) results in

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{J=1}^{N}\left(\frac{n(n-1)}{N^2}\right)^2 - \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)^2 = \frac{n(n-1)}{N^2} - \frac{n^2}{N^2} = -\frac{n}{N^2}, \quad (41)$$

while for the WOR case the covariance can be seen to equal

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{\substack{J=1\\J\neq I}}^{N}\left(\frac{n(n-1)}{N(N-1)}\right)^2 - \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)^2 = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\left(\frac{n(N-n)}{N^2(N-1)}\right). \quad (42)$$

It should be stressed that for SRS, regardless of whether it takes place with or without replacement, the correlation coefficients are given by

$$\mathrm{Corr}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{-1}{(N-1)}, \qquad (43)$$

independently of the sample size.

Furthermore, we have that, as the value of n approaches that of N in WOR sampling, both $\pi_I^{(n)}$ and $\pi_{IJ}^{(n)}$ approach one. In particular, when $n = N$, the values of expressions (31) and (40) become zero.

5. The Correlation Matrix for Sample Units

Once we realize that none of the expressions in (28), (31) and (40) depend on any of the arbitrary indices used to differentiate population units, it should become clear that the $r \times r$ correlation matrix for the random vector $\underline{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r})$, where $r \leq n$, may be written as:

$$\mathrm{Corr}(\underline{\theta}) = R_r(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}. \qquad (44)$$

It should be noted that the elements of $R_r(\rho)$ in (44) depend only on the inclusion probabilities which, for any sample size, may be fully computed from recursion (7), and expressions (8) and (10). In other words, they do not depend on any unknown population parameters to be estimated nor on the values of the variables to be measured on the sample units.

6. Final Remarks

In theory, the efficiency of every estimation procedure will experience some gain whenever explicit allowance for the correlation between sample units is made. This would certainly be the case for linear as well as for some instances of maximum-likelihood estimation.

On the other hand, it should be emphasized that $R_n(\rho)$ may become singular as the sample size n approaches the population size N; this is the case for SRS ($R_N(-1/(N-1))$ as well as for WOR sampling in general. Therefore, numerically, many estimation procedures which rely on the inverse or the determinant of R, rather than on the correlation matrix itself, may also benefit from replacing the simplifying assumption of independence between observations by a more realistic one of correlated observations whenever sample sizes are large relative to population sizes. Instances where this can happen are given by some stages in multi-stage sampling (*e.g.*, number of households in a block) and by large country-wide surveys.

Survey Methodology, December 2005
Vol. 31, No. 2, pp. 239-243
Statistics Canada, Catalogue No. 12-001-XPB

Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling

Changbao Wu [1]

Abstract

We present computational algorithms for the recently proposed pseudo empirical likelihood method for the analysis of complex survey data. Several key algorithms for computing the maximum pseudo empirical likelihood estimators and for constructing the pseudo empirical likelihood ratio confidence intervals are implemented using the popular statistical software R and S-PLUS. Major codes are written in the form of R/S-PLUS functions and therefore can directly be used for survey applications and/or simulation studies.

Key Words: Confidence interval; Bi-section algorithm; Empirical likelihood; Newton-Raphson procedure; Stratified sampling; Unequal probability sampling. $\boldsymbol{x}$

1. Introduction

One of the major challenges in applying advanced and often sophisticated statistical methods for real world surveys is the computational implementation of the method. Practical considerations often rule out the use of methods which are theoretically sound and attractive but are computationally formidable.

The empirical likelihood method first proposed by Owen (1988) is one of the major advances in statistics during the past fifteen years. In addition to its data driven and range respecting feature in estimation and testing, its nonparametric and discrete nature is particularly appealing for finite population problems. Indeed an early version of the method, the so-called scale-load estimators, was used in survey sampling by Hartley and Rao back in 1968. The more recent investigation of the method in survey sampling has resulted in a series of research papers and generated noticeable interests among survey statisticians to further explore the method. Wu and Rao (2004) contains a brief summary on the recent development of the pseudo empirical likelihood (PEL) method in survey sampling.

Progress on algorithmic development for the PEL method has also been made. A modified Newton-Raphson procedure for computing the maximum PEL estimators under non-stratified sampling was proposed by Chen, Sitter and Wu (2002). The procedure was further modified by Wu (2004a) to handle stratified sampling designs.

In this article we present computational algorithms for computing the maximum PEL estimators and for constructing the related PEL ratio confidence intervals for complex surveys under a unified framework, with particular interest in implementing those algorithms using R and S-PLUS. The software package R, a friendly programming environment and compatible to the popular commercial statistical software S-PLUS, is attracting more and more users from the statistical community. What is advantageous about using R is that it is available free for research use and the package may be easily downloaded from the web. It is hoped that this article will bridge the current gap between theoretical developments and practical applications of the PEL method and will generate more research activities in this direction to make fully practical use of the PEL method a reality.

The algorithm for computing the maximum PEL estimator under non-stratified sampling and some notes on its implementation in R/S-PLUS are presented in section 2. The algorithm of Wu (2004a) for stratified sampling is discussed in section 3. Construction of the PEL ratio confidence intervals involves profiling the pseudo empirical likelihood ratio statistic and is detailed in section 4. All R functions or sample codes are included in the Appendix. They can also be downloaded from the author's personal homepage http://www.stats.uwaterloo.ca/~ cbwu/paper. html. These functions and codes had been tested in the simulation study reported in Wu and Rao (2004) and were observed to perform very well.

2. Non-Stratified Sampling

Consider a finite population consisting of N identifiable units. Associated with the i^{th} unit are values of the study variable, and a vector of auxiliary variables, . The vector of population means $\overline{\boldsymbol{X}} = N^{-1}\sum_{i=1}^{N} \boldsymbol{x}_i$ is known. Let $\{(y_i, \boldsymbol{x}_i), i \in s\}$ be the sample data where s is the set of units selected using a complex survey design. Let $\pi_i = P(i \in s)$ be the inclusion probabilities and $d_i = 1/\pi_i$ be the design weights.

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. E-mail: cbwu@uwaterloo.ca.

The pseudo empirical maximum likelihood estimator of the population mean $\bar{Y}=N^{-1}\sum_{i=1}^{N} y_i$ is computed as $\hat{\bar{Y}}_{\mathrm{PEL}}=\sum_{i\in s}\hat{p}_i\, y_i$ where the weights $\hat{p}_i$ are obtained by maximizing the pseudo empirical log likelihood function

$$l_{ns}(\boldsymbol{p})=n^{*}\sum_{i\in s} d_i^{*}\log(p_i) \qquad (2.1)$$

subject to the set of constraints

$$0<p_i<1,\ \sum_{i\in s}p_i=1 \ \text{ and } \ \sum_{i\in s}p_i\,\boldsymbol{x}_i=\bar{\boldsymbol{X}}. \qquad (2.2)$$

The original pseudo empirical likelihood function proposed by Chen and Sitter (1999) is $l(\boldsymbol{p})=\sum_{i\in s} d_i\,\log(p_i)$. The pseudo empirical likelihood function $l_{ns}(\boldsymbol{p})$ given by (2.1) was used by Wu and Rao (2004), where $d_i^{*}=d_i/\sum_{i\in s} d_i$ are the normalized design weights and n^{*} is the effective sample size. The point estimator $\hat{\bar{Y}}_{\mathrm{PEL}}=\sum_{i\in s}\hat{p}_i\, y_i$ remains the same for either version of the likelihood function. The rescaling used in $l_{ns}(\boldsymbol{p})$ facilitates the construction of the PEL ratio confidence intervals.

Using a standard Lagrange multiplier argument it can be shown that

$$\hat{p}_i=\frac{d_i^{*}}{1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})} \quad \text{for} \quad i\in s, \qquad (2.3)$$

where the vector-valued Lagrange multiplier, $\boldsymbol{\lambda}$, is the solution to

$$g_1(\boldsymbol{\lambda})=\sum_{i\in s}\frac{d_i^{*}(\boldsymbol{x}_i-\bar{\boldsymbol{X}})}{1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})}=0.$$

The major computational task here is to find the solution to $g_1(\boldsymbol{\lambda})=0$. This can be done using the modified Newton-Raphson procedure proposed by Chen *et al.* (2002). The modification involves checking at each updating stage that the constraint $1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})>0$ (*i.e.*, $p_i>0$) is always satisfied. Without loss of generality, we assume $\bar{\boldsymbol{X}}=0$ (if not, replace $\boldsymbol{x}_i$ by $\boldsymbol{x}_i-\bar{\boldsymbol{X}}$ throughout). The modified procedure is as follows.

Step 0: Let $\boldsymbol{\lambda}_0=\boldsymbol{0}$. Set $k=0$, $\gamma_0=1$ and $\varepsilon=10^{-8}$.

Step 1: Calculate $\boldsymbol{\Delta}_1(\boldsymbol{\lambda}_k)$ and $\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$ where

$$\boldsymbol{\Delta}_1(\boldsymbol{\lambda})=\sum_{i\in s} d_i^{*}\frac{\boldsymbol{x}_i}{1+\boldsymbol{\lambda}'\boldsymbol{x}_i}$$

and

$$\boldsymbol{\Delta}_2(\boldsymbol{\lambda})=\left\{-\sum_{i\in s} d_i^{*}\frac{\boldsymbol{x}_i\,\boldsymbol{x}_i'}{(1+\boldsymbol{\lambda}'\boldsymbol{x}_i)^2}\right\}^{-1}\boldsymbol{\Delta}_1(\boldsymbol{\lambda}).$$

If $\|\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)\|<\varepsilon$, stop the algorithm and report $\boldsymbol{\lambda}_k$; otherwise go to Step 2.

Step 2: Calculate $\boldsymbol{\delta}_k=\gamma_k\,\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$. If $1+(\boldsymbol{\lambda}_k-\boldsymbol{\delta}_k)'\boldsymbol{x}_i\le 0$ for some i, let $\gamma_k=\gamma_k/2$ and repeat Step 2.

Step 3: Set $\boldsymbol{\lambda}_{k+1}=\boldsymbol{\lambda}_k-\boldsymbol{\delta}_k$, $k=k+1$ and $\gamma_{k+1}=(k+1)^{-1/2}$. Go to Step 1.

In the original algorithm presented by Chen *et al*. (2002), their step 2 also checks a related dual objective function. While this is necessary for the theoretical proof of convergence of the algorithm, it is not really required for practical applications.

The R function Lag2(u,ds,mu) can be used for finding the solution to $g_1(\boldsymbol{\lambda})=0$ when the vector of auxiliary variables $\boldsymbol{x}$ is of dimension m and $m\ge 2$. When $\boldsymbol{x}$ is univariate, an extremely simple and stable bi-section method to be described shortly should be used. Let n be the sample size. The three required arguments are the $n\times m$ data matrix u, the $n\times 1$ vector of design weights ds and the $m\times 1$ population mean vector mu. The output of the function Lag2(u,ds,mu) returns the value of $\boldsymbol{\lambda}$ which is the solution to $g_1(\boldsymbol{\lambda})=0$.

The function Lag2(u,ds,mu) will fail to provide a solution if (i) the mean vector $\bar{\boldsymbol{X}}$ is not an inner point of the convex hull formed by $\{\boldsymbol{x}_i,\ i\in s\}$, or (ii) the matrix $\sum_{i\in s} d_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i'$ is not of full rank. In case (i) the pseudo empirical maximum likelihood estimator does not exist. This happens with probability approaching to zero as the sample size n goes to infinity; in case (ii) one may consider to remove some components of the $\boldsymbol{x}$ variables from the set of constraints (2.2) to eliminate the collinearity problem.

When the $\boldsymbol{x}$ variable is univariate, so is the involved Lagrange multiplier $\boldsymbol{\lambda}$. In this case we need to solve $g_2(\lambda)=\sum_{i\in s} d_i^{*}\,x_i/(1+\lambda x_i)=0$ for a scalar λ, assuming $\bar{X}=0$. A unique solution exists if and only if $\min\{x_i,\ i\in s\}<0<\max\{x_i,\ i\in s\}$. The solution, if exists, lies between $L=-1/\max\{x_i,\ i\in s\}$ and $U=-1/\min\{x_i,\ i\in s\}$. Noting that $g_2(\lambda)$ is a monotone decreasing function for $\lambda\in(L,\,U)$, the most efficient and reliable algorithm for solving $g_2(\boldsymbol{\lambda})=0$ is the bi-section method. The function Lag1(u,ds,mu) does exactly this, where the required arguments are u $=(x_1,\ldots,x_n)$, ds $=(d_1,\ldots,d_n)$ and mu $=\bar{\boldsymbol{X}}$. The output returns the solution to $g_2(\lambda)=0$.

The function Lag1(u,ds,mu) can be used in conjunction with the model-calibrated pseudo empirical likelihood (MCPEL) approach of Wu and Sitter (2001) to handle cases where the $\boldsymbol{x}$ variable is high dimensional. The MCPEL approach involves only a single dimension reduction variable derived from a multiple linear regression model and the related Lagrange multiplier problem is always of dimension one.

3. Stratified Sampling

Let $\{(y_{hi},\,\boldsymbol{x}_{hi}),\ i\in s_h,\ h=1,\ldots,H\}$ be the sample data from a stratified sampling design. Let $d_{hi}^{*}=d_{hi}/\sum_{i\in s_h} d_{hi}$ be the normalized design weights for stratum h, $h=1,\ldots,H$. The pseudo empirical likelihood function

under stratified sampling defined by Wu and Rao (2004) is given by

$$l_{st}(\boldsymbol{p}_1, ..., \boldsymbol{p}_H) = n^* \sum_{h=1}^{H} W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}), \quad (3.1)$$

where $W_h = N_h / N$ are the stratum weights and n^* is the total effective sample size as defined in Wu and Rao (2004). The value of n^* is not required for point estimation but this scaling constant is needed for the construction of confidence intervals. Let $\overline{\boldsymbol{X}}$ be the known vector of population means for auxiliary variables. The maximum pseudo empirical likelihood estimator of the population mean $\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h$ is defined as $\hat{\overline{Y}}_{\text{PEL}} = \sum_{h=1}^{H} W_h \sum_{i \in s_h} \hat{p}_{hi} y_{hi}$ where the $\hat{p}_{hi}$ maximize $l_{st}(\boldsymbol{p}_1, ..., \boldsymbol{p}_H)$ subject to the set of constraints

$$p_{hi} > 0, \sum_{i \in s_h} p_{hi} = 1, \ h = 1, ..., H$$

and

$$\sum_h W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}. \quad (3.2)$$

The major computational difficulty under stratified sampling is caused by the fact that the subnormalization of weights (*i.e.,* $\sum_{i \in s_h} p_{hi} = 1$) occurs at the stratum level while the benchmark constraints (*i.e.,* $\sum_h W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}$) and the constrained maximization of the PEL function are taken at the the population level. The algorithm proposed by Wu (2004a) for computing the $\hat{p}_{hi}$ proceeds as follows: let $\boldsymbol{x}_{hi}$ be augmented to include the first $H-1$ stratum indicator variables and $\overline{\boldsymbol{X}}$ be augmented to include $(W_1, ..., W_{H-1})$ as its first $H-1$ components. In the case of no benchmark constraints involved, the augmented $\boldsymbol{x}$ variable will consist of the $H-1$ stratum indicator variables only and $\overline{\boldsymbol{X}} = (W_1, ..., W_{H-1})$. It follows that the set of constraints (3.2) is equivalent to

$$p_{hi} > 0, \sum_{h=1}^{H} W_h \sum_{i \in s_h} p_{hi} = 1$$

and

$$\sum_{h=1}^{H} W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}, \quad (3.3)$$

where the $\boldsymbol{x}$ variable is now augmented. Let $\boldsymbol{u}_{hi} = \boldsymbol{x}_{hi} - \overline{\boldsymbol{X}}$. It is straightforward by using a standard Lagrange multiplier argument to show that

$$\hat{p}_{hi} = \frac{d_{hi}^*}{1 + \boldsymbol{\lambda}' \boldsymbol{u}_{hi}},$$

with the vector-valued $\boldsymbol{\lambda}$ being the solution to

$$g_3(\boldsymbol{\lambda}) = \sum_h W_h \sum_{i \in s_h} \frac{d_{hi}^* \boldsymbol{u}_{hi}}{1 + \boldsymbol{\lambda}' \boldsymbol{u}_{hi}} = 0.$$

The modified Newton-Raphson procedure of section 2 for solving $g_1(\boldsymbol{\lambda}) = \boldsymbol{0}$ can be used for solving $g_3(\boldsymbol{\lambda}) = \boldsymbol{0}$. The key computational step under stratified sampling designs is to prepare the data file into suitable format so that the R function Lag2(u,ds,mu) for non-stratified sampling can directly be called. Sample R codes for doing this are included in the Appendix.

4. Construction of PEL Ratio Confidence Intervals

While the computational algorithms for the maximum PEL estimator under non-stratified and stratified sampling designs are somewhat different, the search for the lower and the upper boundary of the pseudo empirical likelihood ratio confidence interval for $\overline{Y}$ involves the same type of profile analysis. Under non-stratified sampling designs, the $(1-\alpha)-$level PEL ratio confidence interval of $\overline{Y}$ is constructed as

$$\{\theta \mid r_{ns}(\theta) < \chi_1^2(\alpha)\}, \quad (4.1)$$

where $\chi_1^2(\alpha)$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom. The pseudo empirical log likelihood ratio statistic $r_{ns}(\theta)$ is computed as

$$r_{ns}(\theta) = -2\{l_{ns}(\tilde{\boldsymbol{p}}) - l_{ns}(\hat{\boldsymbol{p}})\},$$

where the $\hat{\boldsymbol{p}}$ maximize $l_{ns}(\boldsymbol{p})$ subject to the set of "standard constraints" such as (2.2) and the $\tilde{\boldsymbol{p}}$ maximize $l_{ns}(\boldsymbol{p})$ subject to the "standard constraints" plus an additional one induced by the parameter of interest, $\overline{Y}$, *i.e.*

$$\sum_{i \in s} p_i y_i = \theta. \quad (4.2)$$

To compute $\tilde{\boldsymbol{p}}$ one needs to treat (4.2) as an additional component of the "standard constraints" for each fixed value of θ so that the maximization process is essential the same as before.

Let $(\hat{L}, \hat{U})$ be the interval given by (4.1). Our proposed bi-section method in searching for $\hat{L}$ and $\hat{U}$ is based on following observations:

i) The minimum value of $r_{ns}(\theta)$ is achieved at $\theta = \sum_{i \in s} \hat{p}_i y_i = \hat{\overline{Y}}_{\text{PEL}}$. In this case $\tilde{\boldsymbol{p}} = \hat{\boldsymbol{p}}$ and $r_{ns}(\theta) = 0$.

ii) The interval $(\hat{L}, \hat{U})$ is bounded by $(y_{(1)}, y_{(n)})$ where $y_{(1)} = \min\{y_i, i \in s\}$ and $y_{(n)} = \max\{y_i, i \in s\}$.

iii) The pseudo empirical likelihood ratio function $r_{ns}(\theta)$ is monotone decreasing for $\theta \in (y_{(1)}, \hat{\overline{Y}}_{\text{PEL}})$ and monotone increasing for $\theta \in (\hat{\overline{Y}}_{\text{PEL}}, y_{(n)})$.

Conclusion iii) can be reached by noting that $l_{ns}(\hat{\boldsymbol{p}})$ does not involve θ and $l_{ns}(\tilde{\boldsymbol{p}}) = n^* \sum_{i \in s} d_i^* \log(\tilde{p}_i)$ is typically a concave function of θ. It is also possible to show this by directly checking $dr_{ns}(\theta)/d\theta$. For instance, in the case of no auxiliary information involved, the "standard constraints" are $p_i > 0$ and $\sum_{i \in s} p_i = 1$. The $\hat{p}_i$ are given by d_i^* and $\hat{\overline{Y}}_{\text{PEL}} = \sum_{i \in s} d_i^* y_i$. The $\tilde{p}_i$ are computed as

$$\tilde{p}_i = \frac{d_i^*}{1+\lambda(y_i-\theta)}, \qquad (4.3)$$

where the λ is the solution to

$$\sum_{i\in s} \frac{d_i^*(y_i-\theta)}{1+\lambda(y_i-\theta)} = 0. \qquad (4.4)$$

Using (4.3) and (4.4), and noting that $\sum_{i\in s} d_i^* / (1+\lambda(y_i-\theta)) = 1$, it is straightforward to show that

$$\frac{d}{d\theta} r_{ns}(\theta) = 2n^* \sum_{i\in s} \frac{d_i^* \{(d\lambda/d\theta)(y_i-\theta)-\lambda\}}{1+\lambda(y_i-\theta)} = -2n^*\lambda.$$

By re-writing $d_i^*(y_i-\theta)$ as $d_i^*(y_i-\theta)\,[\{1+\lambda(y_i-\theta)\} - \lambda(y_i-\theta)]$ and after some re-grouping in (4.4) we get

$$\lambda \sum_{i\in s} \frac{d_i^*(y_i-\theta)^2}{1+\lambda(y_i-\theta)} = \sum_{i\in s} d_i^* y_i - \theta.$$

It follows that $dr_{ns}(\theta)/d\theta = -2n^*\lambda < 0$ if $\theta < \sum_{i\in s} d_i^* y_i = \hat{\bar{Y}}_{\mathrm{PEL}}$ and $dr_{ns}(\theta)/d\theta > 0$ otherwise.

Sample codes for finding $(\hat{L}, \hat{U})$ where no auxiliary variable is involved are included in the Appendix. In this case $\hat{p}_i = d_i^*$ and $\hat{\bar{Y}}_{\mathrm{PEL}} = \sum_{i\in s} d_i^* y_i = \hat{\bar{Y}}_H$ is the Hajek estimator for $\bar{Y}$. The profiling process involves finding λ for each chosen value of θ and evaluating the PEL ratio statistic $r_{ns}(\theta)$ against the cut-off value from the χ_1^2 distribution under the desired confidence level $1-\alpha$. With auxiliary information, one needs to modify the computation of $r_{ns}(\theta)$ for each fixed θ. The bi-section search algorithm for finding $\hat{L}$ and $\hat{U}$ remains the same.

The value of the effective sample size n^* is required for computing the PEL ratio statistic $r_{ns}(\theta)$. For non-stratified sampling designs it is computed as $n^* = \hat{S}_y^2 / \hat{V}(y)$ where

$$\hat{S}_y^2 = \frac{1}{N(N-1)} \sum_{i\in s} \sum_{j>i} \frac{(y_i - y_j)^2}{\pi_{ij}},$$

and

$$\hat{V}(y) = \frac{1}{N^2} \sum_{i\in s} \sum_{j>i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2,$$

where $e_i = y_i - \hat{\bar{Y}}_{\mathrm{HT}}$ and $\hat{\bar{Y}}_{\mathrm{HT}} = N^{-1}\sum_{i\in s} d_i\, y_i$. See Wu and Rao (2004) for further detail. Computation of n^* involves the second order inclusion probabilities π_{ij} which may impose a real challenge if a πps sampling scheme is used. In the simulation study reported in Wu and Rao (2004), the Rao-Sampford πps sampling method was used. R functions for selecting a πps sample using this method as well as for computing the related second order inclusion probabilities can be found in Wu (2004b). Similar R functions are also available in an add-on R package called "pps", written by J. Gambino (2003), which can be downloaded from the R homepage http://cran.r-project.org/ by clicking the packages option.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks an associate editor for helpful comments which lead to improvement of the paper.

Appendix: R/S-PLUS Codes

A1. R Function for solving $g_1(\lambda) = 0$.

Let m be the number of auxiliary variables involved and $m \geq 2$. There are three required arguments in the function Lag2(u,ds,mu):

(1) u: the $n \times m$ data matrix with $\boldsymbol{x}_i$ as its i^{th} row, $i = 1, \ldots, n$.

(2) ds: the $n \times 1$ vector of design weights consisting of $d_1, \ldots, d_n$.

(3) mu: the $m \times 1$ population mean vector $\bar{\boldsymbol{X}}$.

The output of the function is the solution to $g_1(\lambda) = 0$.

```
Lag2<-function(u,ds,mu)
{
  n<-length(ds)
  u<-u-rep(1,n)%*%t(mu)
  M<-0*mu
  dif<-1
  tol<-1e-08
  while(dif>tol){
    D1<-0*mu
    DD<-D1%*%t(D1)
    for(i in 1:n){
      aa<-as.numeric(1+t(M)%*%u[i,])
      D1<-D1+ds[i]*u[i,]/aa
      DD<-DD-ds[i]*(u[i,]%*%t(u[i,]))/aa^2
          }
    D2<-solve(DD,D1,tol=1e-12)
    dif<-max(abs(D2))
    rule<-1
    while(rule>0){
      rule<-0
      if(min(1+t(M-D2)%*%t(u))<=0) rule<-rule+1
      if(rule>0) D2<-D2/2
          }
    M<-M-D2
  }
  return(M)
}
```

A2. R Function for solving $g_2(\lambda) = 0$.

When the x variable is univariate, the solution to $g_2(\lambda) = 0$ can be found through a simple and reliable bi-section method. The three required arguments for the function Lag1(u,ds,mu) are $\mathrm{u} = (x_1, \ldots, x_n)$, $\mathrm{ds} = (d_1, \ldots, d_n)$ and $\mathrm{mu} = \bar{\boldsymbol{X}}$. The output is the solution to $g_2(\lambda) = 0$.

```
Lag1<-function(u,ds,mu)
{
  L<--1/max(u-mu)
  R<--1/min(u-mu)
  dif<-1
  tol<-1e-08
  while(dif>tol){
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    dif<-abs(glam)
  }
  return(M)
}
```

A3. Sample code for stratified sampling.

We need to call the function Lag2(u,ds,mu) from nonstratified sampling. The key step is to prepare the data file into suitable format. Let

(1) n = $(n_1, ..., n_H)$ be the vector of stratum sample sizes.

(2) x be the data matrix with $\boldsymbol{x}_{hi}$ as row vectors, $i = 1, ..., n_h, h = 1, ..., H$.

(3) ds = $(d_{11}^*, ..., d_{1n_1}^*, ..., d_{H1}^*, ..., d_{Hn_H}^*)$, where d_{hi}^* are the normalized initial design weights for stratum h.

(4) X be the vector of known population means.

(5) W = $(W_1, ..., W_H)$ be the vector of stratum weights (*i.e.*, $W_h = N_h / N$).

The following sample codes show how the solution to $g_3(\boldsymbol{\lambda}) = \mathbf{0}$ is found (M from the second last line of the following code) and how the $\hat{p}_{hi}$'s are computed (phi from the last line).

```
###
nst<-sum(n)
k<-length(n)-1
ntot<-rep(0,k)
  ntot[1]<-n[1]
  for(j in 2:k) ntot[j]<-ntot[j-1]+n[j]
ist<-matrix(0,nst,k)
  ist[1:n[1],1]<-1
  for(j in 2:k) ist[(ntot[j-1]+1):ntot[j],j]<-1
uhi<-cbind(ist,x)
mu<-c(W[1:k],X)
whi<-rep(W[1],n[1])
  for(j in 2:(k+1)) whi<-c(whi,rep(W[j],n[j]))
dhi<-whi*ds
M<-Lag2(uhi,dhi,mu)
phi<-as.vector(ds/(1+(uhi-rep(1,nst)%*%t(mu))%*%M))
###
```

A4. Sample code for finding the PEL ratio confidence interval.

The search for the lower boundary (LB) and the upper boundary (UB) of the PEL ratio confidence interval needs to be carried out separately. The following codes show how this is done for the case of no auxiliary information. With auxiliary information, one needs to modify the computation of the involved pseudo empirical likelihood ratio statistic (elratio) accordingly. Let

(1) a = $1-\alpha$ be the confidence level of the desired interval.

(2) ys = $(y_1, ..., y_n)$ be the sample data.

(3) ds = $(d_1^*, ..., d_n^*)$ be the normalized design weights.

(4) YEL = $\sum_{i \in s} \hat{p}_i \, y_i$ (in this case $\hat{p}_i = d_i^*$).

(5) nss be the estimated effective sample size n^*.

```
###
tol<-1e-08
cut<-qchisq(a,1)
###
t1<-YEL
t2<-max(ys)
dif<-t2-t1
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t2-t1
        }
UB<-(t1+t2)/2
###
t1<-YEL
t2<-min(ys)
dif<-t1-t2
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t1-t2
        }
LB<-(t1+t2)/2
###
```

References

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2004b). R/S-PLUS Implementation of pseudo empirical likelihood methods under unequal probability sampling. Working paper 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.

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

Wu, C., and Rao, J.N.K. (2004). Pseudo empirical likelihood ratio confidence intervals for complex surveys. Working paper 2004-06, Department of Statistics and Actuarial Science, University of Waterloo.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2005.

A.K Adhikary, *ISI Kolkata*
M. Battaglia, *ABT Associates*
J.-F. Beaumont, *Statistics Canada*
N. Billor, *Auburn University*
C. Boudreau, *Medical College of Wisconsin*
K. Brewer, *Australian National University*
F. Butar Butar, *Sam Houston State University*
D. Cantor, *Westat*
S.R. Chowdhury, *Westat, Inc.*
S.L. Christ, *University of North Carolina*
J. Chromy, *RTI International*
R. Courtemanche, *Institut de la statistique du Québec*
A. Dessertaine, *EDF R&D-OSIRIS - CLAMART*
P. Dick, *Statistics Canada*
P. Duchesne, *Université de Montréal*
J. Dumais, *Statistics Canada*
J. Eltinge, *Bureau of Labour Statistics*
M. Feder, *Research Triangle Institute*
R. Fisher, *U.S. Census Bureau*
O. Frank, *Stockholm University*
S.G. Heeringa, *Institute for Social Research, University of Michigan*
S. Haslett, *Massey University*
D. Heng-Yan Leung, *Singapore Management University*
K. Jae Kwang, *Yonsei University*
F. Jenkins, *Westat*
J. Jiang, *University of California at Davis*
J.K. Kim, *Yonsei University*
M. Kovačević, *Statistics Canada*
S. Laaksonen, *University of Helsinki and Statistics Finland*
P. Lahiri, *University of Maryland*
F. Lapointe, *Institut de la statistique du Québec*
M.D. Larsen, *Iowa State University*
P. Lavallée, *Statistics Canada*
H. Lee, *Westat, Inc.*
R. Lehtonen, *University of Jyvaskyla*
N.T. Longford, *SNTL*
L. Magee, *McMaster University*
T. Maiti, *Iowa State University*
D. Malec, *United States Bureau of the Census*
B. Mandall, *Ohio State University*
S. Matthews, *Statistics Canada*
D. Marker, *Westat, Inc.*
D. McCaffrey, *RAND*
C.E. M'Lan, *University of Connecticut*
J. Moore, *U.S. Bureau of the Census*
R. Munnich, *University of Tubingen*
J. Opsomer, *Iowa State University*
O. Phillips, *Statistics Canada*
M. Pratesi, *University of Pisa, Italy*
J. Reiter, *Duke Univesity*
R.H. Renssen, *Statistics Netherlands*
G. Roberts, *Statistics Canada*
I. Şchiopu-Kratina, *Statistics Canada*
C.J. Schwarz, *Simon Fraser University*
A. Scott, *University of Aucklan*
J. Sedransk, *Case Western University*
R. Sitter, *Simon Fraser University*
M. Sinclair, *U.S. Department of Labor*
A. Singh, *Statistics Canada*
T.W. Smith, *NORC*
J. Stec, *InteCap, Inc*
D.G. Steel, *University of Wollongong, Australia*
L. Stokes, *Southern Methodist University*
E. Stuart, *Mathematica Policy Research, Inc.*
A. Jr. Tersine, *United States Bureau of the Census*
R. Thomas, *Carleton University*
N. Thomas, *Pfizer, Inc.*
C. Tucker, *United States Bureau of Labor*
J. van der Brakel, *Statistics Netherlands*
S.L. Vartivarian, *Mathematica Policy Research, Inc.*
J. Wang, *Merck Research Labs, Merck & Co., Inc.*
X. Wang, *Southern Methodist University*
C. Wu, *University of Waterloo*
R. Yucel, *University of Massachusetts*
W. Yung, *Statistics Canada*
E. Zanutto, *University of Pennsylvania*
H. Zheng, *Massachusetts General Hospital and Harvard Medical School*

Acknowledgements are also due to those who assisted during the production of the 2005 issues: Francine Pilon-Renaud and Roberto Guido (Dissemination Division), Marc Bazinet (Marketing Division) and François Beaudin (Official Languages and Translation Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier, Nancy Flansberry and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

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 21, No. 2, 2005

Volume 21, No. 3, 2005

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 33, No. 2, June/juin 2005

Volume 33, No. 3, September/septembre 2005


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

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. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 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, page 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.