e3

SURVEY METHODOLOGY

e3

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2002

•

VOLUME 28

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2002 • VOLUME 28 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

January 2003

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN $47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 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 28, Number 2, December 2002

CONTENTS

In This Issue

This issue of *Survey Methodology* includes papers on a variety of topics including overviews of small area statistics and data quality in statistical offices, survey nonresponse and imputation, survey design, data collection and estimation.

In the first paper of this issue, Brackstone identifies strategies and approaches for the development of small area statistics programs in national statistical offices. The topic of small area estimation will be covered by a number of papers in a special section in the June 2003 issue of *Survey Methodology*. The paper first considers the crucial role of censuses, and discusses issues related to their usefulness for small area statistics. Other potential sources of small area statistics include administrative files and sample surveys, either on their own or combined with census data to provide estimates for the intercensal period or for characteristics not directly covered by the census. Rolling censuses are also discussed, as well as the unique challenges in producing small area business and environmental statistics. Finally, issues of organization of national statistical offices for production and dissemination of small area statistics are considered.

Trewin reviews the practices and approaches used to maintain high quality of output from a national statistical office. Important ingredients include good relations with respondents, skilled and motivated staff, sound statistical and operational methods, and relevance of statistical programs. Current challenges include increasing the use of administrative data sources, effective use of the internet for both collection and dissemination, maintaining knowledge and skills as staff leave, and handling increasing user expectations. This paper is based on a talk presented as the keynote address at Statistics Canada's Symposium 2001.

Thibaudeau presents an innovative approach to the imputation of demographic characteristics in a large scale survey or Census. Instead of relying on the usual approach of either the closest complete record in the processing stream or constructing imputation groups, Thibaudeau proposes a compromise method which uses maximum likelihood estimation based on the conditional probabilities. This approach seeks to create groups that are close in order and in geography to the imputed record. He also presents an interesting Bayesian approach to evaluating the method.

Nandram, Han and Choi consider the problem of analyzing multinomial nonignorable nonresponse data from small areas in the framework of Bayesian inference. This paper extends some earlier work by Stasny by assuming a Dirichlet prior underlying the multinomial probabilities and using a prior distribution on the hyperparameters. The authors apply this model to Body Mass Index data from a complex survey design.

In the Stewart paper, the possible biases introduced by different contact strategies in telephone time-use surveys are investigated. Two contact strategies, convenient-day scheduling, where the designated reference day changes with the contact day, and designated-day scheduling, where the reference day remains fixed, are discussed and compared through simulation studies.

Bell and McCaffrey consider the problem of unbiasedly estimating the variance of coefficients of linear regressions from multi-stage survey data when only a small number of Primary Sampling Units (PSUs) are sampled. After investigating situations where the bias of the linearization variance estimator can be large, a bias reduced linearization variance estimator is proposed. In addition, a Satterthwaite approximation is used to determine the degrees of freedom to be used for tests and confidence intervals in conjunction with the bias reduced linearization variance estimator.

Sirken considers estimation of the volume of transactions that a population of establishments has with a population of households. An approach based on indirect sampling of establishments through the households that they have transactions with is compared to the more typical approach based on direct pps sampling of establishments. Estimators and expressions for the variances are derived and compared for the two methods. Situations where one approach or the other is preferable are explored.

Rivest considers the problem of identifying stratum boundaries. The commonly used Lavallée-Hidiroglou algorithm assumes that the values of the study variable are available and are used in the determination of optimal stratum bounds. In his paper, Rivest relaxes this assumption and modifies the Lavallée-Hidiroglou algorithm to account for a discrepancy between the stratification variable and the study variable through the use of models that link these two variables together. These models are then incorporated into the Lavallée-Hidiroglou algorithm.

In the Lu and Sitter paper, the problem of the sample size being smaller or only slightly larger than the total number of strata is considered. Consequently, conventional methods of sample allocation to strata may not be applicable. One solution for this problem is to use a linear programming technique to minimize the expected lack of desirability of the samples subject to a constraint of expected proportional allocation (EPA). However, as the number of strata increases this solution rapidly becomes expensive in terms of magnitude of computation. In the proposed approach, the amount of computation is reduced substantially at the small cost of approximate EPA for strict EPA.

Renssen and Martinus explore the use of generalized inverse matrices in survey sampling. After reviewing the properties of generalized inverses, they consider the generalized regression estimator when the set of regressors is not of full rank, and they set out a regularity condition under which the estimator is invariant to the choice of generalized inverse. They then present an algorithm for calculating the regression weights, and briefly discuss weighting in the Dutch Labour Force Survey.

M.P. Singh

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 117-123
Statistics Canada

Strategies and Approaches for Small Area Statistics

GORDON J. BRACKSTONE[1]

ABSTRACT

National statistical offices are often called upon to produce statistics for small geographic areas, in addition to their primary responsibility for measuring the condition of the country as a whole and its major subdivisions. This task presents challenges that are different from those faced in statistical programs aiming primarily at national or provincial statistics. This paper examines these challenges and identifies strategies and approaches for the development of programs of small area statistics. The important foundation of a census of population, as well as the primary role of a consistent geographic infrastructure, are emphasized. Potential sources and methods for the production of small area data in the social, economic and environmental fields are examined. Some organizational and dissemination issues are also discussed.

KEY WORDS: Small area statistics; Census; Geography.

1. INTRODUCTION

The mandate of most national statistical offices (NSO) focuses on the monitoring of social, economic, and environmental conditions at the national level, and for the major administrative units (provinces, states, major metropolitan areas) within the country. However, the demand for data at lower geographic levels is always present, especially from local governments and from businesses needing to make investment, marketing, and location decisions that depend on knowledge of local areas. We will use the term "small area statistics" to mean statistics for areas below the level of state, province, or major metropolitan areas – a broad spectrum of areas from large towns, through urban neighbourhoods, to rural villages. In some circles the term "small areas" is used more broadly to refer to any small sub-group or domain of the population, but here we are talking strictly about small geographic areas.

The extent of an NSO's responsibility for small area statistics depends on the division of governmental responsibilities within a country. For example, in some countries local governments are the creation of provinces and the responsibility for supporting their statistical needs may rest with provincial governments. But in many countries, whatever the formal division of powers, it is, *de facto*, the NSO that is expected to respond to the need for small area statistics, either within its own resources or in cooperation with other levels of government. At the very least, it is the NSO that must set the standards and framework for small area data if these are not to become a mishmash of uneven and overlapping statistics incomparable across the country.

With limited budgets an NSO is faced with the difficult trade-off between investment in national statistics and provision of small area detail. How should it choose between covering more subject areas, or existing subject areas in more detail, at the national and provincial levels, and, on the other hand, providing more small area detail for subject areas it is already covering nationally? There is no formula for resolving this problem. The balance struck in any country will be largely a function of national needs, relative powers, and historical tradition, with perhaps some statistical considerations on the margin. Nevertheless, there is a series of measures and approaches that a NSO can consider to maximize the degree to which it can satisfy demands for small area statistics within a limited budget.

Four potential sources of small area statistical data either individually or in combination, account for most production of small area data by statistical agencies. Censuses or complete enumerations of populations are the traditional source. Administrative records, including national registers, that cover all, or almost all, of a defined population are in many respects equivalent to a census. National sample surveys are rarely large enough to produce small area data directly but they do represent a valuable current source of information that can be used, under certain assumptions and in combination with other sources, to produce small area data. And finally, local studies focused on particular small areas will produce small area data, but not for complete sets of small areas. Sources such as satellite imaging or aerial photography can be thought of as censuses or local studies depending on their coverage.

In this paper we first review the important role of the Census of Population, with or without a population register, in the provision of small area socio-economic data (Section 2), and then emphasise the fundamental role of an up-to-date geographic infrastructure to support any production of small area statistics, including especially the census of population (Section 3). We then examine approaches to providing small area data on individuals and families between censuses (Section 4), on business activities (Section 5), and on environmental issues (Section 6). We conclude with some general observations about the dissemination of small area statistics and the management of small area statistics within an NSO.

[1] Gordon J. Brackstone, Informatics and Methodology Field, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: bracgor@statcan.ca.

2. CENSUS OF POPULATION

The census of population, in most countries, plays the central role in the provision of small area data about people, families and households. Based on a complete enumeration of the population (at least for basic characteristics), its estimates are free of the sampling error that limits the ability of sample surveys to produce small area estimates. Provided the individual households are geographically coded to a fine level (*e.g.*, a block or block face), direct tabulations of households can produce statistical aggregates for any geographic area that can be defined, or approximated, in terms of the lowest level of geographic coding.

However, censuses have their drawbacks. They are costly, and therefore they are infrequent. Data from the last census may provide a poor representation of a small area that is undergoing rapid development. In many countries, sampling is utilized in the census for many of the questions. While this introduces sampling error into estimates from the census, these samples are still huge compared with those in a typical sample survey. Furthermore, the samples are typically spread through every enumeration area of the country, so the ability to produce small area estimates is maintained, even though the small areas will need to be somewhat larger than in a true census.

Potentially more serious, with respect to accuracy, are nonsampling errors such as coverage error and response bias. Most censuses miss some people, or count some people twice, and it has been repeatedly shown that those miscounted are generally not typical of the population as a whole. Census estimates may therefore be biased against certain sub-groups of the population. If these subgroups (*e.g.*, certain immigrant groups) tend to be geographically clustered, this can have a serious impact on estimates for some small areas. Response bias arises if a census question is systematically misunderstood by many respondents. Both small area and large area estimates would be affected by such errors.

Countries that maintain a population register have the potential to produce census-like data for small areas more frequently than the traditional 5-10 year cycles of a census. Up-to-date residence registration is clearly a requirement for accurate small area data from such registers. The breadth of data available from a register system may be less than that available through a conventional census, since the former is limited to the characteristics maintained in linkable administrative registers. In some countries the population register may be used as the basis for a census that collects the necessary additional characteristics not available within existing registers Redfern (1989) provides a useful description of practices within Europe in this regard.

Since the Census has the potential to produce estimates for very small areas, rules to protect against direct or residual disclosure of individual data have to be in place. These can include imposing a minimum population on areas for which data will be released, random perturbation of data, suppression of data, or other techniques (Jabine (1993), Zayatz, Steel and Rowland (2000)). NSOs have also to be concerned about privacy issues arising with the publication of small area census data that, while not disclosing any individual responses, do reveal dominant characteristics of an area (*e.g.*, that 90% of the families received unemployment benefits). Such findings cannot be withheld, but they can be selected and presented with sensitivity.

Though a census, with or without a population register, is a source of direct small area data as of census day, the value of such data declines as time passes. However, the role of census data in the provision of small area statistics goes well beyond the direct use of the results from each periodic enumeration. Inter-censally, census data may be used as a benchmark, a sampling frame, or as auxiliary information to be used with other sources of data that are available between censuses. These usages are pursued in section 4. An innovative alternative to the traditional census is described in Section 4.4.

3. GEOGRAPHIC INFRASTRUCTURE

To enable a national census to produce accurate data for small areas, a geographic infrastructure of boundaries and mapping capacity covering the whole country is a prerequisite. Such an infrastructure requires that each dwelling be associated with a precise geographic location on the ground, where the degree of precision determines the fineness with which small areas can be defined. Though modern global positioning technology makes it possible to pinpoint each dwelling to a specific pair of coordinates, it is usually sufficient for statistical purposes to associate each dwelling in an urban area with a block face (*i.e.*, one side of a street between two intersections), or a building in the case of high-rise buildings. In rural areas, the chosen degree of precision will depend on local administrative and natural boundaries, though maximum flexibility is preserved by using precise coordinates for each dwelling.

While necessary for a census, a geographic infrastructure is equally required for the provision of small area statistics from other sources. Essentially each data point, from whatever source, has to be associated with a geographic location at a level detailed enough to allow aggregation into any small areas of statistical interest. For example, if the data source is an administrative register, or a business register, the address in each record must be convertible into a pair of geographic coordinates, or at least into a small area within which the address falls. Since administrative registers often use mailing addresses, a file that converts postal codes into geographic locations is a valuable tool in the development of small area data.

The availability of an accurate up-to-date geographic infrastructure, whether maintained by the NSO or obtained from outside, is essential if a program of small area statistics is to have flexibility in the choice of areas for which statistics are produced.

4. SMALL AREA STATISTICS ON PERSONS AND HOUSEHOLDS - BETWEEN CENSUSES

We turn now to the issue of producing small area data for persons or households inter-censally. Clearly the existence of a current population register makes a fundamental difference to what is possible, and how it can be done. We will confine ourselves to the case where no regularly updated population register exists.

In such circumstances, there are three main classes of approach. The first is to utilize census-like files that come from administrative systems and purport to cover the whole of a well-defined population. The second is to exploit sample survey data and, through additional model assumptions, produce estimates for smaller (though still not very small) areas than is possible through direct survey estimation. The third category is the combination of one or both of these first two approaches with the use of data from the most recent census. In the following paragraphs we review some of the characteristics of these approaches.

4.1 Administrative Files

An example of an administrative file with small area statistical potential is the annual file of individual income tax returns. Other examples, with narrower population definitions, might be drivers' licences, employment insurance recipients, or health insurance records. In the case of tax data, if each record contains a residential address that can be associated with a geographic point or small area, then data can be tabulated directly for small areas, with due regard for confidentiality (as with census data). The characteristics available would generally be restricted to demographic and income variables, and the coverage would be limited to taxfilers. Nevertheless, such a file represents a rich source of annual data for quite small areas. Population coverage can be improved through the imputation of dependents "claimed" on the tax record. In Canada, the coverage of such imputed files is approaching that of the census as coverage increases among low income earners who need to file tax returns to obtain social assistance benefits.

With administrative data in general, the statistician has to take what is available (though some influence on content may be possible in the longer term), reconcile any differences in concepts, definition or coverage between the administrative file and the statistical objectives, and assess any issues of reporting or coding accuracy in the records. Subject to these precautions, administrative data can provide a geographically rich potential source of small area data (Brackstone 1987).

4.2 Sample Survey Data

The problem with sample survey data as a source of small area statistics is sample size. There are frequently insufficient sample cases in the small area to allow a reliable direct estimate to be produced, and sometimes none at all. In large national sample surveys it may be possible to devise sampling strategies that ensure an acceptable level of precision for planned small areas, such as sub-provincial regions, without significantly degrading the reliability of estimates at higher levels (Singh, Gambino and Mantel 1994). But for smaller areas, or for areas of similar size not taken into account during design, reliable estimation will not be possible. Larger samples help, and may allow direct estimation for some of the larger small areas, but budgets usually constrain this approach as a general solution. If no other data sources are available, statisticians can only resort to model-based methods which involve making assumptions about how data for a small area relate to other data. These methods are often described as "borrowing strength" since they borrow information from elsewhere in the sample survey to augment the number of units that contribute to the estimate for a given small area. The borrowing can be from other time periods, from sample units outside the given small area, or from other variables measured on the same sample unit. Some examples follow. Most of these examples will allow some expansion of the range of small area estimates that can be produced from sample surveys with relatively large samples. They cannot magically convert small sample surveys into rich sources of small area data.

1. In a monthly survey, it may be possible to combine data for a small area over a period of consecutive months to produce direct estimates of a multi-month moving average for the area. For example, quarterly estimates may be possible where monthly ones were not.

2. One may be ready to assume that means or proportions estimated for a larger area apply equally to a smaller component area within it. If the size of the small area is known, an estimate can be obtained by multiplying by the assumed mean or proportion. This assumption may be more realistically made within subgroups of the population (*e.g.*, age groups), rather than for the population as a whole. In this case, if the size of each sub-group is known for the small area, a synthetic estimator can be built up by multiplying the sizes by the assumed means and aggregating.

3. If additional related variables are available from the survey, more elaborate models may be set up relating the variable being estimated to these auxiliary variables. The parameters of the model may be estimated at a higher geographic level where there is sufficient sample to estimate them reliably. The model is then applied with the estimated parameters to the data for the given small area.

All of these approaches suffer from the lack of reliable baseline data for each small area. If such data are available, for example from a recent census or from administrative records, then the data may be used in combination to produce more reliable estimates than from either source alone.

4.3 Combined Sources

Methods that combine census or administrative information from the recent past with current sample survey data are borrowing strength from outside the survey. They still require model assumptions. However, these can often be weaker (since they involve assumptions about change from the benchmark, rather than about absolute levels of each small area) and so more acceptable, or more plausible, than in the case of sample survey data alone.

A wide variety of estimation methods (which we won't attempt to describe here) have been developed to handle this situation. Some of these methods can be thought of as estimating change since the most recent benchmark, others as distributing reliable current sample survey estimates among component small areas based on benchmark data, and yet others as recalibrating old benchmark figures to new current estimates. In essence, they all involve some kind of balancing of three kinds of estimates: (a) high variance but unbiased direct current survey estimates for the small area in question; (b) low variance current survey estimates for some surrounding or comparable larger area; and (c) census-type estimates for the same small area from recent administrative data, or a past census, which may contain unknown bias due to the source and the time lag. Any available auxiliary data can be incorporated to improve the accuracy of each component estimate. The way in which these three types of estimates are combined is determined by the choice of model and model parameters.

In summary, the methods of this and the previous section essentially reduce variance by making use of more data, but at the expense of introducing potential bias due to model assumptions that will never be exactly correct. It is very important to analyse the performance of these methods before their use, for example by carrying out the estimation process in a census year when direct estimates are available for comparison, and periodically thereafter. Model checking is becoming an area of increased research activity (Bayarri and Berger 2000). For more detailed descriptions of available methods in this class see, for example, Purcell and Kish (1979); Fay and Herriott (1979); Ghosh and Rao (1994); Singh *et al.* (1994); Schaible (1996); Rao (1999) and Gambino and Dick (2000).

4.4 Rolling Censuses

An innovative alternative to the census is being investigated in at least two countries. The method of producing small area data based on a large rolling sample has long been advocated by Leslie Kish as an alternative to the traditional census (Kish 1990, 1998). The sample survey "rolls" in the sense that over a long period (*e.g.*, a decade) each of the smallest areas for which estimates are required would be included once in the sample so as to provide a direct estimate for that area once each period. Successively larger areas (aggregates of the smallest areas) would be represented more often in the sample, allowing either more reliable or more frequent estimates for those areas. For even larger areas, including provinces and the whole country, the accumulated sample would be sufficient to provide reliable annual, or more frequent, estimates at certain levels of detail. The approach may be considered with or without a periodic census to collect basic demographic data against which to calibrate the inter-censal survey estimates.

The rolling census avoids the need for the assumption of models, but presumes that unbiased estimates of multi-year averages, or asynchronous estimates for different areas of the country, are satisfactory alternatives to the simultaneous point-in-time estimates of the traditional census. Relative cost is also a key factor, especially in the situation where a basic census is also carried out. On the other hand, by producing reliable annual estimates for many of the larger areas, and with much of the content detail of a census, this approach could effectively address the issue that census estimates can be up to 12 years old before the next ones appear. It also responds to mounting concerns over increasing difficulties and costs associated with the conduct of a traditional census.

This approach is being tested in the United States under the name of the American Community Survey (Alexander 1999, 2002) and in France where it is referred to as the "recensement continu" (Isnard 1999; Durr and Dumais 2002).

5. BUSINESS STATISTICS

The problems of producing small area data for businesses are different in many important respects from those encountered for data on persons or households.

Whereas the association of each individual with a "usual place of residence" is, for the vast majority of the population, a fairly clear and unambiguous concept (though perhaps becoming less clear with the growth of second residences, the incidence of prolonged absences away from the snow, and more flexible living arrangements), for businesses the question of where, geographically, to attribute various characteristics of a business is less clear in many situations. For single establishment businesses where all the activity takes place in a single location there is no conceptual problem, though there may still be a practical problem if the source of information is an administrative file that provides, say, an accountant's address rather than the place of business. For some variables, such as employment, there may be no major conceptual problem even for larger businesses (except perhaps for those working in the transportation industry, or certain service industries). However, for variables such as revenues and profits there can be real questions about how these should be allocated geographically in multi-establishment businesses. The larger the geographic area the smaller the problem – location within a province doesn't matter if one is only interested in provincial totals. But, in general, geographic attribution rules have to be determined before small area estimates for business activity can be considered, and for some aspects of business activity small area estimates may not make conceptual sense.

While for household surveys the main obstacle to the production of small area estimates is sample size, for business surveys considerations of confidentiality usually constitute the major barrier. The smaller the area, the greater the chance that a particular industry will be dominated by one or a few major companies, thus precluding the provision of estimates for that area due to disclosure risk. Methods for checking statistical output on businesses to recognize potential disclosure risks are fairly well developed (Federal Committee on Statistical Methodology 1994) but require constant attention on the part of the NSO. The confidentiality problem is less of an issue in those industries characterized by small units – which may be the same industries in which the conceptual problems of the previous paragraph are not so severe. In those industries, considerations of sample size may indeed be the limiting factor, in which case the families of methods described in the previous section are available.

A third area of contrast with data on individuals, at least for countries that do not maintain a population register, is the existence of a relatively up-to-date list frame of businesses. This not only provides a base for sampling and a source of some auxiliary data for estimation, but also constitutes a potential source of direct estimates of business demography, at least annually. In many countries the currency of the business register is maintained by receiving transactions from the business tax system, which itself provides an annual census-like source of administrative data on business activity. However, use of tax data still requires careful consideration of the conceptual, geographical and confidentiality issues raised above.

6. ENVIRONMENT STATISTICS

Environment statistics provide yet different challenges for the production of small area statistics. While some environmental issues are national or even global in scope, many are by their nature local. Many sources of pollution are typically localized with their impacts being felt most severely in the neighbourhood of a plant or accident. The socio-economic impacts of broader environmental problems (*e.g.*, loss of fish stocks) are frequently felt in small and often isolated resource-based communities.

Some environment data are collected from households or individuals (*e.g.*, recycling practices, fuel use) and their potential as a source of small area data is subject to the considerations already described in section 4. Other environment data (*e.g.*, waste generation, environmental protection expenditures, use of natural resources) come from businesses and would be governed by the considerations of Section 5. However, a great deal of environment data is obtained from physical surveys (*e.g.*, geological, physiographic, hydrographic), from instrument measurement (*e.g.*, temperature, air quality, water quality, ozone layer thickness), and from direct observation (*e.g.*, land use). Different considerations govern the relation of these data sources to small area data.

Because environment data are no respecters of administrative boundaries, the need for a flexible geographic infrastructure, emphasised in Section 3, is especially important here. Small area geographic identification is needed to regroup data to geographical units that are more suitable for environmental analysis. For example, the production of waste attributable to a certain type of agricultural activity might be aggregated for all of the producers within a river basin. Environmental geographic units are either predefined (ecozones, drainage basins) or dictated by special events (areas covered with different thicknesses of ice, land areas flooded by heavy rains or spring thaws). In some cases, the area studied could be a very small site such as a park.

Physical quantity or quality data can be difficult to aggregate or summarize. In some cases, point source data such as air quality measures cannot be considered representative of any larger geographic unit. Water quality may be summarized or compared by using an indicator, such as the number of days beaches are open for swimming, but not simply as an aggregate or average of water quality readings. For many measures, the focus of interest may be on change over time rather than small area comparisons. In other cases, sampling and estimation techniques may need to make use of spatial analysis techniques such as contouring or interpolation.

The privacy and confidentiality concerns associated with environment data depend on their source. Data collected from households or businesses, even if they involve physical measurements, are protected by the same confidentiality rules as other data from those sources. Direct measurements of the stock of natural resources or the quality of the environment do not raise these concerns. Cartographic representation of spatial patterns may be one way to overcome some of the analytical frustrations of data suppression for small areas. Choropleth maps (maps which show the distribution of variables or characteristics by using colour or shading for ranges of the distribution) can explicitly represent the ranges implicit in rows or columns that would be suppressed in a published table.

Cross-border pollutant flows and their global effects make physical environment data an international issue. Cooperation between neighbouring countries is necessary to ensure that national boundaries do not impede analysis of the impact of physical processes that recognize no such boundaries.

In summary, the small area dimension is particularly important for environment data, not only because a locality is frequently the point of interest, but also because data must often be reaggregated to geographic areas more appropriate for environmental analysis such as ecozones or watersheds.

7. ORGANIZATION AND DISSEMINATION ISSUES

Most NSOs are organized by subject-matter area. The production of small area estimates cuts across subject-matter areas, but requires support from Geography staff for geographic infrastructure, from Methodology staff for estimation and evaluation methods, and perhaps from other staff for analyzing and packaging data across subject areas. The question of how to organize small area estimation within an NSO therefore arises.

Requiring subject-matter areas to manage small area estimation in their areas, with support from methodology and geography staff as needed, is a natural choice since they should be most in touch with the data requirements and data limitations in their subject areas. More of an issue is how to package data for small areas for dissemination to users. Who should be responsible for pulling together data from different subject-matter areas for a particular small area? Should this be a regular program, or something that is done 'on demand'? Here there are different models to choose from – and Statistics Canada has tried most of them over the years.

At some periods in the past a division focussing on regional or urban statistics has existed to provide a regional focus for statistical data. At times, the census program, which is of course the richest source of small area data, has spearheaded the production of small area data profiles. At other times, an inter-divisional project has been used to manage a program of profiles for electoral districts or for other geographic areas. At the same time, regional office staff have played a key role in pulling together information for small areas in response to client requests. None of these arrangements has been ideal. The production of profiles has typically been a labour-intensive task requiring a broad subject-matter understanding and a lot of searching and manipulation of data. Despite the existence of standard geographic areas, the combination of data based on several different geographic bases is usually an issue. Ensuring that data for a large number of small areas are properly matched and collated can be an arduous quality assurance challenge.

Pre-planned profiles on paper were never overly successful. As a result, a strategy of maximizing responsiveness to client demands as they arose was preferred. With recent advances in technology, and broader coverage of small area data in the corporate database, a more automated approach is possible. A component of the Statistics Canada website (www.statcan.ca), called Community Profiles, and largely based on Census of Population data, is our most recent attempt to make small area data more accessible and promises to be a precursor of future directions in this field. Some health data for health districts are already included, and certain other non-census sources of community data are under consideration.

8. CONCLUSIONS

The production of small area statistics by an NSO raises issues that are qualitatively different from those faced in its regular production of national, provincial or other large area data. The statistical theory that makes data based on sufficient individual measurements inherently reliable for large areas (ignoring bias for the moment) begins to break down for smaller areas. Unless a current census or administrative source with full coverage is available, this means that the NSO has to resort to some model-based help in order to provide estimates. Since alternative models can produce different estimates, a degree of arbitrariness is introduced into estimates, and this may be seen by some as undermining the objectivity of a NSO and its methods. The fundamental principle of openness and transparency about methods, including the choice of any models used and the impact of different assumptions, takes on even greater importance in the domain of small area estimation.

On top of this, an NSO should expect that small area estimates will come under more focused scrutiny than do many large area estimates. Though large area estimates receive broader attention, few individuals have the capacity to confirm or refute an estimate at the national level. But at the local level there will be many who think they know what is going on in their town. And typically small area estimation does not work uniformly well for all areas. The argument that a method works well on average will not quell criticism from those areas where it has not worked well – unless it has also worked to the local advantage! The NSO has to be prepared for the double jeopardy of weaker estimates under closer scrutiny.

If that is not enough already, confidentiality considerations loom larger at the small area level. The very fact that estimates are being produced for local areas highlights the potential for identification of individuals even though the NSO has taken sufficient precautions to prevent such disclosure. Some users of small area data for marketing purposes do not help the situation by implying in their advertizing that they can target mail to households based on individual or household characteristics, when they are actually using small area data to distinguish neighbourhoods. Some methods of small area estimation require record linkage which may also raise privacy concerns. Again a policy of openness and careful review of all such applications, at a senior level and before they begin, is necessary to ensure that the public benefit outweighs any privacy invasion.

Despite these potential difficulties, the demand for small area data remains high, technology offers new approaches to the management and dissemination of small area data, and methodological work on small area estimation is an active research area among statisticians. While small area data will generally not be an NSO's first priority, the relevance of its statistical programs will be magnified many times if it is able to cater to the most important small area data needs.

ACKNOWLEDGEMENTS

This paper was originally prepared to introduce a discussion on small area statistics among Heads of National Statistical Offices at the June 2001 Conference of European Statisticians in Geneva. The author wishes to acknowledge the contributions of many individuals at Statistics Canada who provided input to this paper, as well as helpful suggestions from a referee.

REFERENCES

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

ALEXANDER, C.H. (2002). Still rolling: Leslie Kish's "Rolling samples" and The American Community Survey. *Survey Methodology*. 28, 1, 35-41.

BAYARRI, M.J., and BERGER, J.O. (2000). *P* Values for composite null models. *Journal of the American Statistical Association*. 95, 452, 1127-1142.

BRACKSTONE, G. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology*. 13, 1, 29-43.

DURR, J.-M., and DUMAIS, J. (2002). Redesign of the French Census of Population. *Survey Methodology*. 28, 1, 43-49.

FAY, R.E., and HERRIOTT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *Journal of American Statistical Association*. 74, 269-277.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on Statistical Disclosure Limitation Methodology (Statistical Policy Working Paper #22). Washington, D.C., Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

GAMBINO, J., and DICK, P. (2000). Small area estimation practice at Statistics Canada. *Statistics in Transition*. 4, 597-610.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*. 9, 55-93.

ISNARD, M. (1999). *Alternatives to Traditional Census Taking: The French Experience*. Paris: INSEE.

JABINE, T.B. (1993). Statistical disclosure limitation practices of united states statistical agencies. *Journal of Official Statistics*. 9, 2, 427-454.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*. 16, 1, 63-71.

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

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

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*. 25, 2, 175-186.

REDFERN, P. (1989). European experience of using administrative data for censuses of population: the policy issues that must be addressed. *Survey Methodology*. 15, 1, 83-99.

SCHAIBLE, W.L. (1996). (Ed.) *Indirect Estimators in U.S. Federal Programs, Lecture Notes in Statistics*. New York: Springer-Verlag, 108.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*. 20, 3-14.

ZAYATZ, L., STEEL, P. and ROWLAND, S. (2000). Disclosure limitation for Census 2000. *Proceedings of the American Statistical Association, Section on Government Statistics and Section on Social Statistics*. 67-71.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 125-133
Statistics Canada

The Importance of a Quality Culture

DENNIS TREWIN[1]

ABSTRACT

The reputation of a national statistical office (NSO) depends very much on the quality of the service it provides. Quality has to be a core value – providing a high quality service has to be the natural way of doing business. It has to be embedded in the culture of the NSO.

The paper will outline what is meant by a high quality statistical service. It will also explore those factors that are important to ensuring a quality culture in a NSO. In particular, it will outline the activities and experiences of the Australian Bureau of Statistics in maintaining a quality culture.

KEY WORDS: Continuous quality improvement; National Statistical Office.

1. INTRODUCTION

Fellegi (1996) provides a strong argument that the trust in the national statistical agency is how most users judge the quality of its statistical products.

> "Credibility plays a basic role in determining the value to users of the special commodity called statistical information. Indeed, few users can validate directly the data released by statistical offices. They must rely on the reputation of the provider of the information. Since information that is not believed is useless, it follows that the intrinsic value and usability of information depends directly on the credibility of the statistical system. That credibility could be challenged at any time on two primary grounds: because the statistics are based on inappropriate methodology, or because the office is suspected of political biases."

Trust will not happen unless the culture is right. Culture is a word with many meanings but I am interpreting culture as "the way we do things". Core values are important to this. They cannot be just statements hanging on the wall. They have to be understood. They have to be reflected in behaviours, particularly by leaders of organizations.

The Australian Bureau of Statistics (ABS) places great reliance on adherence to its core values. More than anything, they distinguish us from other survey providers in Australia. The core values are:

- Relevance – regular contact with those with policy influence, good statistical planning, which requires a keen understanding of the current and future needs for statistics, are essential, as is the need for statistics to be timely and relatable to other statistics.
- Integrity – our data, analysis and interpretation should always be objective and we should publish statistics from all collections. Our statistical system is open to scrutiny, based on sound statistical principles and practices.
- Access for all – our statistics are for the benefit of all Australians and we ensure that equal opportunity of access to statistics is enjoyed by all users.
- Professionalism – the integrity of our statistics is built on our professional and ethical standards. We exercise the highest professional standards in all aspects of ABS statistics.
- Trust of providers – we have a compact with respondents; they are encouraged to provide us with accurate information and we ensure that the confidentiality of the data provided is strictly protected. We keep the load and intrusion on respondents to a minimum, consistent with meeting justified statistical requirements.

Adherence to core values is just one element of maintaining a quality culture. Part 2 discusses the key steps the ABS uses to maintain a quality culture.

It is now widely recognized that quality is much more than accuracy (*e.g.*, Brackstone 1999 and Carson 2000). In Part 3, the different dimensions of quality are discussed before identifying in Part 4 what I think are some of the major quality challenges for the ABS over the medium term. Many of these will be shared by other national statistical organizations.

2. TOWARDS A HIGH QUALITY STATISTICAL SERVICE

Quality assurance is a responsibility of all staff in the ABS. There is no central "quality management" group although Methodology Division is encouraged to be our conscience on quality issues – a role it takes on with

[1] Dennis Trewin, Australian Statistician, Australian Bureau of Statistics.

enthusiasm, sometimes to the annoyance of others. However, that is a good sign – they are provoking debate on some of the more difficult quality issues. Support from senior management for this type of role is very important.

The key strategies for ensuring a high quality are described under six broad headings.

- A high degree of credibility for the ABS and its outputs.
- Maintaining the relevance of ABS outputs.
- Effective relationships with respondents.
- Processes that produce high quality outputs.
- Regular review and evaluation of statistical activities.
- Staff who are skilled and motivated to assure the quality of ABS outputs.

2.1 A High Degree of Credibility

Credibility is fundamental to the effective use of official statistics. Credibility arises from a system of statistics which provides an objective window upon the condition of a nation's economy and society.

The legislative framework within which the ABS operates is an important pre-condition for the integrity of Australia's official statistics. The Australian Statistician (*i.e.,* the chief executive of the ABS) is guaranteed considerable independence by law. This helps ensure that the ABS is, and is seen to be, impartial and free from political interference. In particular, the independence of the Statistician supports his objectivity in determining the statistical work program and determining what statistics are published. Although the legal authority is there, it still needs to be reflected in the way senior staff behave.

Government statisticians must not just apply professionalism skills to their work; they must also be seen to adhere to high ethical standards, especially with respect to objectivity and integrity. We are frank and open when describing our statistical methods to users; we publish information about our performance – for example, in terms of both sampling and non-sampling errors, and revision histories for key series; we are willing and able to identify and address user concerns regarding quality; we are receptive to objective criticism and prepared to respond quickly even if the problem is one of perception rather than reality. We promote good relationships with the media as they have a major influence on public opinion of the ABS and its outputs. Also, most Australians find out about official statistics through the media. We engage in other user education activities aimed at fostering intelligent use of official statistics.

The fact and perception of ABS objectivity are reinforced by our policies of pre-announcing publication dates for main economic indicators, allowing very limited pre-release of publications (the details of which are in the public domain), and making special data services available on an even handed basis to all.

2.2 Maintaining the Relevance of ABS Outputs

There can be, of course, tension between (on the one hand) being responsive to changing policy needs and (on the other) maintaining the continuity of a system of statistics that can objectively monitor performance. Senior staff of the ABS devote a great deal of attention to maintaining personal contact with key users, to gather intelligence about policy issues and emerging areas of economic, social and environmental concern. This includes regular meetings with the most senior staff of the government agencies responsible for policy. The Directors of our State offices have similar arrangements with State officials. That intelligence feeds into strategic planning and the reviews of national statistical programs.

The ABS has a range of other means for communicating with the users of statistics, to ensure that our products are relevant to their needs. For example, advisory groups representing users and experts in various fields provide valuable guidance to our statistical activities.

There may also be some tensions or trade-offs between the different aspects of quality. The ABS positions itself at the higher accuracy end of the information market, to protect the valuable ABS "brand name". But if, for example, there is an urgent demand for data in a new field, some aspects of quality may be traded off in order to achieve timeliness and relevance. Nevertheless, there is a "bar" below which we will not go. Because it is probable that the new statistics will be used to inform significant decisions or debate, the ABS makes very clear statements about the accuracy of the data to help users understand how they can be used. On occasion, such new statistics may be differentiated from our other products by labelling them "experimental" or releasing as an information or occasional paper, rather than a standard publication. We regard this form of branding as very important to reliable interpretation of our statistics.

2.3 Effective Relationships with Respondents

An official statistical agency must maintain good relations with respondents, especially trust, if it wants them to co-operate and provide high quality data. The ABS approach includes – explaining the importance of the data to government policy, business decisions and public debate; a policy of thoroughly testing all forms before they are used in an actual survey; obtaining the support of key stakeholders; minimizing the load placed on respondents particularly by using administrative data where possible; and carefully protecting privacy and confidentiality.

The ABS monitors and manages the load it imposes on both households and businesses; we have developed 'respondent charters' for both groups. As well, a Statistical Clearing House has been set up within the ABS to coordinate surveys of businesses across government agencies (including the ABS), to reduce duplication and to ensure that statistics of reasonable quality are produced.

All ABS forms and collection methods are tested to ensure that the data we seek are available at reasonable cost to respondents, and the best available methods are used to collect them. For business surveys, our units model, classifications and data items, are designed to be as consistent as possible with the way businesses operate. This now corresponds closely with their reporting for taxation purposes, making it easier to integrate survey data with data collected for taxation purposes. For household surveys, the extensive use of cognitive testing tools within the ABS, and the establishment of a questionnaire testing laboratory, have helped to improve quality and to reduce respondent load. Standards for form design and form evaluation are set out in manuals and are promoted and supported by experts in form design.

The ABS uses efficient survey designs to minimize sample sizes to achieve a specified level of accuracy, and hence total reporting load; we also control selection across collections to spread the load more equitably. To take advantage of current reforms of the Australian taxation system, the ABS is seeking every opportunity to improve the efficiency of our sample designs, through the use of taxation data as benchmarks, as well as using it as a substitute for some of the data now gathered through direct collections. We have changed the business unit structure used in our surveys to make it consistent with the structure used for taxation purposes.

For household surveys, the introduction of computer assisted interviewing has helped to streamline interviewing procedures, reduce respondent load, and improve the quality of data collected.

2.4 Processes that Produce High Quality Outputs

The quality of ABS statistics is underwritten by the application of good statistical methods during all stages of a collection including the design stages. The ABS has a relatively large Methodology Division (about 120 staff) which reports directly to the Australian Statistician. The Division is responsible for ensuring that sound and defensible methods are applied to all collections and compilations. The Methodological Advisory Committee, a group of academic experts, provides independent reviews of our statistical methods.

The ABS puts substantial effort into developing statistical standards, including concepts, data item definitions, classifications, and question modules. All ABS surveys must use these standards. The standards are supported by relevant data management facilities to ensure they are accessible and to make it easier to use standard rather than non-standard approaches.

Sample design and estimation methods are the responsibility of the Methodology Division. Where possible, a "total survey design" is used – accuracy requirements are set according to the intended use of the data, and accuracy is measured in terms of both sampling and non-sampling errors. For example, in business surveys total survey design guides the allocation of resources to the intensive follow up of non-respondents or the editing of questionnaires; the effort for reducing non-sampling errors is optimized according to the impact of errors on overall quality. The cost to data providers is also taken into consideration. The "total survey design" has to be approved by a senior ABS committee before it is implemented.

In recent years, the ABS has made substantial progress by applying standardized best practice across surveys. For example, business surveys based on the business register now draw their frames at a common date each quarter, and use a common estimation method to ensure all collections have a consistent and complete coverage. Standard rules are adopted for frame maintenance, field collection and estimation, and generalized processing facilities are available to support the use of these rules. Standard methods are used to allow for "new businesses" not yet included on the survey frame. The ABS is thereby able to increase the coherence of estimates across different business surveys.

For household surveys, a master sample system has been adopted since the mid 1960's. The system is updated regularly after each five-yearly census, and has been the cornerstone for ensuring the accuracy of statistics collected from household surveys.

Achieving quality in surveys is easier when computer systems support current best practice. The ABS has invested in generalized tools. They have been developed for all major processing steps of both business and household surveys, including sample frame management, data input and editing, imputation, estimation and aggregation.

The ABS embraces a rigorous continuous quality improvement approach wherever appropriate. The Australian Population Census is a classic example of raising quality through a strategy of measuring quality and involving all staff in examining and devising solutions to quality problems. This approach was applied very effectively at the data processing centre for the 1996 and 2001 Censuses. In both cases, the centre achieved significant budget savings, better quality and an improvement in timeliness. Continuous quality improvement is also applied to the coding of businesses on the business register, and to many other ABS processes.

At the output end of collections, each subject group is required to confront its data with other ABS data and with external information, to ensure the coherence of our statistics. The key macroeconomic data have to be "signed off" by the national accountants in meetings established especially for the purpose of clearing the statistics. The national accountants then have an obligation to use this data without further adjustment in the compilation of the accounts, enhancing consistency between the national accounts and source data collections. More generally, confrontation of different data sources is undertaken by our national accountants through use of an 'input-output approach' to compiling national accounts estimates. The new methodology has led to more consistent accounts. Furthermore, the data confrontation and balancing process at detailed levels have helped to identify data deficiencies. Information about quality is fed back to the economic collection groups and is resulting in a more focused approach to improvements in the quality of source data.

One important quality improvement initiative that the ABS has pursued is the development of an Information Warehouse to manage and store all of our publishable data. By drawing together different datasets into a single database, the Warehouse enables our statisticians to confront statistics produced from different collections. Furthermore, all forms of publication, be they paper based or electronic, are to be produced from a single data store, with the objective of ensuring that the same data released in different products, and at different times, are consistent.

Another important element of quality management is documentation. Good documentation supports review activity and facilitates the dissemination of quality information to users, so they can assess the fitness of the data for the purposes they have in mind. As part of the Information Warehouse initiative, the ABS can now enforce standards for documentation of the metadata that describe concepts, definitions, classifications and quality.

A relevant and responsive statistical service must do more than provide data to clients. The ABS has recently strengthened its analytical ability. A team of analysts has been set up to develop new measures of socioeconomic concepts, to explore relationships between variables and to prototype new analytical products. The expanded program of analysis work is expected to deliver significant benefits in the form of insights into data gaps and quality concerns.

2.5 Review and Evaluation of Statistical Activities

Each ABS area is responsible for continuous quality review and improvement. For statistical collection areas, quality management is supported by sets of performance indicators. A standard set of measures has been developed to permit a comparison of quality across collections. Tools are now being developed to calculate these measures as part of our normal survey processes, and the Information Warehouse will allow us to store and display the measures. The key indicators are also included in the annual reports each Branch makes to the ABS Executive for review.

Quality measures are of interest to the users of statistics. The Information Warehouse will improve users' access to information about quality issues. As well, the ABS places high priority on helping users understand the quality of data and their implications for them, and has adopted active education strategies to promote such understanding. As highlighted in Lee and Allen (2001), there is much to do to improve user understanding of quality.

Each ABS household survey now includes an evaluation program which reviews the effectiveness and efficiency of all survey activities and assesses the extent to which the data are used by clients. The Statistical Clearing House conducts a review of each ABS business survey. These initiatives ensure that all collections are subjected to at least a basic evaluation, and brings to light opportunities for improvements to quality and efficiency.

As well as making internal comparisons of performance across its own collection areas, the ABS has established a benchmarking network with overseas statistical agencies; the aim of the network is to share information about survey design, processes and costs. The benchmarking exercise is providing very useful guidance to the ABS's efforts to improve its processes and outputs.

2.6 Skilled and Motivated Staff

The ABS could not provide high quality information to its user community if it did not employ people who bring skills and energy to our statistical work. The staff are responsible for implementing the strategies discussed above. They must take a professional approach and be committed to the development of new methods, to continuous quality improvement, and to the open discussion of methods and quality issues.

Quality improvement and on-going statistical work compete for the time and energies of our staff. The ABS approach is, as far as possible, to integrate quality work with on-going processes and systems. We emphasize to staff that quality management is a corporate priority and ensure that tools and resources are made available to support it. In particular, the ABS is implementing a tighter approach to project management; this is being supported by manuals, systems and training.

Statistical training plays an important role in maintaining and improving quality. The ABS is always searching for new, more effective, approaches to skills development. An important element of our performance management system is a focus on identifying and addressing individuals' development needs.

Relationships with other national and statistical agencies are a very important element of the ABS efforts to improving official statistics. The ABS is committed to using international standards; we take advantage of the wide range of expertise embodied in those standards. On the other hand, there is an obligation for us to make a positive contribution to the development of the standards. In doing so, we try to take account of the interests of the Asia/Pacific region as well as those of Australia. With ever increasing globalization of economic activity and the pursuit of world wide social goals, the compatibility between Australian statistics and those of other countries, is an important element of quality. The ABS maintains strong links with many overseas agencies. We are fortunate that there is a lot in common in the challenges we face and there are great benefits from sharing experiences with other statistical agencies.

3. DIMENSIONS OF QUALITY

Figure 1 is taken from Lee and Allen (2001). Among other things, it neatly summarizes, on the left hand side, three existing frameworks for judging quality. There are some differences with the descriptors used but basically they are providing the same message – there is much more to quality than accuracy. This is now widely accepted although it was not so long ago that discussion of the quality of a statistic focussed on its accuracy and the sampling variability in particular.

There are several messages in the right hand side of Figure 1.

(i) There are many different ways of compiling official statistics – from modelled data/analytical outputs to censuses and sample surveys. In Australia we are making greater use of administrative data, systems of accounts (linked to the national accounts) and model based and other analytical methods to produce statistical outputs, compared with five years ago. The quality challenges differ between the different means of compiling statistics.

(ii) There are several groups of activities associated with statistical outputs – from "frameworks, concepts, standards and classifications" through to "services/dissemination". Each is important in its own right and has its own quality challenge.

(iii) The performance of a National Statistical Office is extremely important to its quality image as recognized in the opening quote of the paper. A number of the elements are specified in Figure 1. All are important. Indeed you cannot have a high performing statistical office unless you rate well against each of these elements; including management and financial performance.

(iv) There are other elements such as institutional settings (*e.g.*, legislation) which are also important.

The main purpose in describing the above is to emphasize that the list of quality challenges for a national statistical office is very large. All have to be tackled in some way – this would not be possible unless you have a quality culture, *i.e.*, attention to quality is the responsibility of all staff. There are many "moments of truth" to genuinely test whether a quality culture exists or not.

4. CURRENT QUALITY CHALLENGES AT ABS

Psychologists say that it is difficult to grasp more than seven points at one time so the remainder of the paper is limited to identifying seven major quality challenges for the ABS.

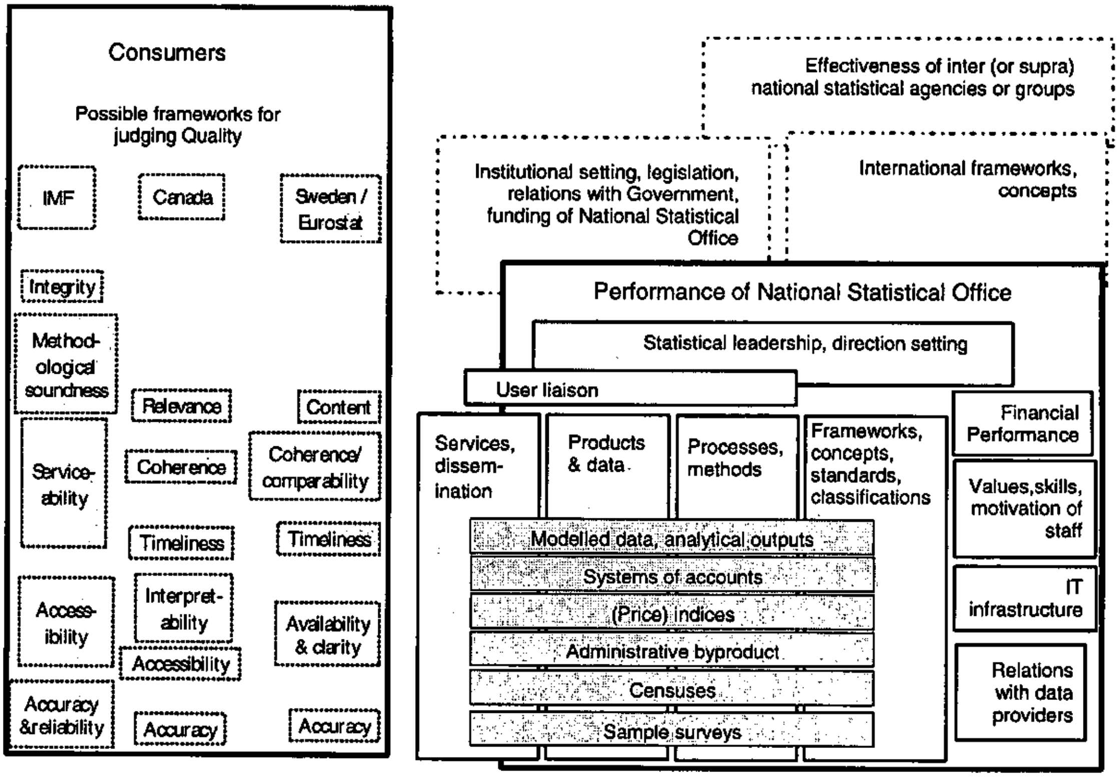
Consumers
Possible frameworks for judging Quality
IMF
Canada
Sweden / Eurostat
Integrity
Method-ological soundness
Relevance
Content
Service-ability
Coherence
Coherence/ comparability
Timeliness
Timeliness
Access-ibility
Interpret-ability
Accessibility
Availability & clarity
Accuracy &reliability
Accuracy
Accuracy
Effectiveness of inter (or supra) national statistical agencies or groups
Institutional setting, legislation, relations with Government, funding of National Statistical Office
International frameworks, concepts
Performance of National Statistical Office
Statistical leadership, direction setting
User liaison
Services, dissem-ination
Products & data
Processes, methods
Frameworks, concepts, standards, classifications
Financial Performance
Values,skills, motivation of staff
IT infrastructure
Relations with data providers
Modelled data, analytical outputs
Systems of accounts
(Price) indices
Administrative byproduct
Censuses
Sample surveys

Figure 1. A Framework for Assessing Quality

(i) The increasing use of large, but imperfect, administrative and transactional data bases for compiling official statistics.

(ii) Increasing user expectations raising the quality "bar".

(iii) Managing the tension between improving business processes (which can mean removing those responsible for statistical outputs from direct involvement with input processes) and maintaining or improving the quality of statistical outputs.

(iv) Quality assurance on electronic outputs.

(v) The presentation of statistics on the internet, including the need to educate the user community on quality of official statistics.

(vi) Managing the transfer of knowledge and skills with an ageing senior management team, many of whom will retire over the next 5 years.

(vii) Use of international statistical standards to maintain comparability where the standard may not be the most appropriate for national statistics.

4.1 Increasing Use of Administrative/Transactional Data Bases

We have used administrative data bases for many years (*e.g.*, vital registrations for births and deaths, customs for trade data) to compile official statistics. Others have been used to develop frameworks for statistical collections. The issues at hand are the increasing availability of these data bases, their under-utilization for statistical purposes, and taking advantage of the potential to link across data bases and ABS collected data sets using a common identifier (*e.g.*, the Australian Business Number for business statistics).

Examples of administrative data bases that are becoming available are extended personal and business income tax data bases, health insurance transactions, and details of those on income support.

Transactional data bases are becoming available, although not in readily accessible form. Data bases of particular interest to the ABS are scanner data bases from retail outlets and eftpos (*i.e.*, electronic fund transfers between customers and retailers) data bases.

There are some particular advantages in using administrative or transactional data bases:

- they reduce the compliance cost we impose on respondents
- they are often "censuses" and therefore provide scope for producing detailed data sets (*e.g.*, by geography)
- they often have a longitudinal element (*e.g.*, tax data) to support this form of analysis
- they often contain an identifier which facilitates analysis across data sets (e.g., *the Australian Business Number will facilit*ate analysis across business tax data sets, customs data, and ABS surveys)
- they might be cheaper than directly collected data sets.

There are negatives of course – for example, the definitions may not be consistent with the preferred statistical concepts; less attention may have been given to incoming quality; and they may be out of date. Managing privacy aspects is a particularly important element. Although our motives are entirely honourable, and are in the public interest, matching data bases is a sensitive issue and ignored at our peril. Many of our users, particularly those in the academic community, are not as sensitive to these concerns.

There is also the question of whether the ABS should produce the statistical outputs or the agency responsible for the data sets. A number of issues come into consideration – the importance of the outputs to the national statistical service, costs, the extent to which quality can be managed and the basic question of whether the administrative agency is prepared to give up custodianship. Only the most important data sets will be brought into the ABS for compiling official statistics; for the others, we will work with the administrative agency to help them deliver "fit for purpose" statistical outputs into the public domain.

What have been our key responses to this important quality management issue?

- We are developing protocols for the publication and management of data from administrative sources. Associated with this is the promotion and support of good statistical and data management practices.
- For each statistical field, we are preparing information development plans in conjunction with other stakeholders which identify those areas of greatest importance and set out specific activities which will lead to increased availability of non-ABS data, particularly quality management issues.
- We are actively promoting good practice in information management.
- A major investment project has been the greater utilization of taxation data to provide cost- effective statistics.
- We are investigating methods for assuring the quality of the very large but imperfect data sets that are available through administrative and transactional data holdings.

4.2 Increasing User Expectations

User expectations on quality are changing – they are much higher than what they were as recently as 5-10 years ago. This trend is likely to continue. The increasing globalization of financial markets will mean that key macroeconomic statistics have international, as well as national prominence.

There is a perception that statistics have become more volatile. In some cases they have because the underlying phenomenon has become more volatile. However, we do not believe statistical measurement methods are a significant contributing factor – in most cases methodological developments have led to improvements although the perception may be different. For example, the volatility in the key national accounts series is considerably less than what it was 10-15 years ago yet this is quite different to the perception of some users.

We also receive more criticism of inaccuracies in very detailed data (*e.g.*, Population Census tables) than previously. Again, it is not that the quality is deteriorating – it is that the expectation is higher.

We have to accept that "the bar is rising" and do what we can to improve quality to the expected level. That is not always possible of course so managing expectations is important. This can be done by:

- providing good explanations of the strengths and weaknesses of particular data sets;
- talking to key users whenever possible about the strengths and weaknesses of data series;
- responding to their informed criticism (seek partnerships in improving quality *e.g.*, in our detailed foreign trade statistics we openly seek feedback from users on the quality of the statistics); and
- providing as much explanation as possible for statistics that might seem unusual or different to expectation.

4.3 Improving Business Processes

Like several statistical organizations, the ABS is looking at how it might use new technologies, and other elements such as increased access to taxation data, to improve the efficiency of its business statistics processes.

We are also investigating the business processes associated with household surveys, particularly as increased use is made of computer assisted interviewing (CAI). However, in this section the paper will concentrate on the changes we are making to the way we manage business statistics to describe this particular quality challenge.

A team was set up to look at the possibilities. As a consequence, a number of significant changes were agreed to – this is to be known as the Business Statistics Innovation Program. We are looking at revised business processes that will be in place for at least 10 years and will yield a significant return on the investments required to set it up. We will:

- extend the responsibilities of the Business Register Unit to capture and store taxation data with a direct link to the Business Register through the Australian Business Number (ABN). The ABN is now allocated through the taxation registration scheme and is available with most business transaction data bases. The data will be stored in a way that it can be used by the various ABS statistical areas to compile statistics directly from taxation data or in combination with ABS survey data;
- improve the way we manage business respondents – this will include some preference in how they provide data to us;
- set up an input data warehouse, with the Australian Business Number as the link across the various data sets;
- establish a business statistics processing environment based around the input data warehouse; and
- increase centralization of a number of the functions associated with compiling business statistics.

We can see the positives in these developments – more efficient delivery of business statistics, enhanced use of taxation data and other administrative data, data bases that support a wider range of statistical analysis. However, it will reduce the level of contact that statistical output areas have with their input data sources. What impact will that have on quality? What strategies can we deploy to mitigate the impact? These are important questions that we will have to answer. It is the main risk we will have to manage in implementing the Business Statistics Innovation Program.

4.4 Quality Assurance on Electronic Outputs

Great care is taken on the quality of our paper products. This has been built on many years of experience. Our record is good and the quality assurance processes well embedded in the way we go about our business. Yet, more and more of user community receive their data in electronic form only. They will make analyses based on these outputs often leading to important decisions being made. It is just as embarrassing to us to have errors in electronic outputs as to have them in paper outputs.

Our quality assurance procedures for electronic outputs are not as sophisticated, but they are evolving. The key responses have been as follows:

- Our data warehouse supports the storage of all the objects associated with the dissemination with a particular set of statistics, including data cubes and meta data.
- Statistical areas are asked to approve each object – they are individually developing their own techniques for quality assurance (but sharing ideas on best practice).
- A publishing system has been developed to support the simultaneous release of all outputs. If they are delivered from the same set of objects, there is less chance of inconsistency between the outputs.

4.5 The Presentation of Statistics on the Internet

Ultimately the user can only make judgements about the fitness of a statistical output for their purposes. These vary of course and what might be fit for one purpose may not be for another. There is an obligation on us to provide a range of supporting information on data outputs, including that on quality, so that the statistical users can make their own judgements on fitness of use. There are a number of existing, well proven practices relating to declarations about the quality of statistics. These activities are now a routine part of existing dissemination practices. They include:

- Concepts, Sources and Methods publications that describe in detail the methods used to compile major statistical outputs. These are available on our web site as well as on other media.
- An assortment of Information and Working Papers, and feature articles in publications, which are used to draw attention to issues specific to particular outputs or changes that are being made to their compilation methods.
- A policy of "no surprises" when there are significant changes to the methods used for the compilation of statistical series. As well as Information Papers *etc*, if there are important changes to statistical series, we embark on a program of seminars and bilateral discussions with key users to explain the changes and the reasons for their changes.
- Material on methods is included in all our publications. The ordering and physical presentation of this information is according to agreed standards. These were developed following research undertaken for us by a communications consultant on how our users use the material in statistical publications.
- The analysis section of our publications includes material that explains, among other things, large or unusual movements in our statistical series. Often this will be based on information that is only available to ABS staff through their contact with respondents or their intimate knowledge of the methods used in compiling statistics. Our User Groups have advised that this is one of the most valuable forms of analysis that we can undertake.

We believe that our key users have a reasonable understanding of the quality of the statistics they use. However the increased reliance on electronic dissemination poses new challenges. In one sense this move provides a wonderful opportunity to present a range of information on quality that is easily accessible through a few well-designed "clicks". But because information about the quality of the statistics is "not in your face" like it can be in hard copy publications it is easier for users to avoid the key messages that you are trying to convey. The real challenge for us is to develop methods for presenting quality in a way that is not easy for users to avoid the main messages we want to convey.

One means of doing this may be to provide separate messages that draw attention to particular information you want to transmit on quality. These could be automatically activated as particular statistical series are accessed or could be delivered by a separate email message. Research is required into the most effective means.

Lee and Allen (2001) have described some of our research work to date on this issue . The work is still at the exploratory stage. Things that are being investigated are:

- Usability testing of how users prefer to access information on quality.
- Showing leadership and developing user education programs on how to use information on quality. A trial version of the is now available.
- The development of four prototype tools to assist users understand the quality of particular statistics. The four prototype tools are "Quality Issue Summaries", "Quality Measures", "Data Accuracy" and "Integrated Access to Data and Metadata".

More details are available in Allen (2001).

4.6 Managing the Transfer of Knowledge and Skills

Like several other national statistical organizations, many of the ABS management team, and other senior staff, are aged in their 50's. Some have retired in recent years. Others are expected to over the next few years. If managed correctly, this is a great opportunity to refresh the organization through providing new blood to management positions. These will normally be younger staff who will bring new ideas and energy into the management team.

On the other hand, experience and know-how will be lost. Both sides of this equation need to be managed carefully. Our strategy is as follows.

- We have developed special programs for those staff with potential. Specifically, they undertake a leadership and management development program which has been specially customized for the ABS. Staff are chosen for these programs by senior managers. You cannot select yourself to be a participant in the program. Furthermore, after staff have completed the program they can be expected to be chosen for a special assignment or rotated to a new position. The underlying philosophy is that the best way of learning is to obtain a variety of work experiences. A very high proportion of recent promotions to senior management positions have been participants in these programs. So far this has helped us to adequately cover the gaps created by a larger number of retirements than in the past.

- We retain links with retired ABS staff through a variety of informal and formal means (*e.g.*, social functions, including them on the distribution list for ABS News, *etc*). Their knowledge is accessible if required.
- We have placed a stronger emphasis on knowledge management, using the facilities of our groupware product (Lotus Notes), means that key parts of our work are well documented and easily accessible.
- We have made substantial moves to standardize methods and systems meaning there is less dependence on local knowledge.
- For some key positions (*e.g.*, Director of National Accounts) we ensure shadowing of work prior to the retirement of the incumbent.

To date we have managed this transition well. We have been able to adequately fill vacant senior positions and at the same time refresh the organization by promoting staff with fresh ideas. There is a need to remain adroit.

4.7 Use of International Standards

Our starting position is that where international standards exist we should use them. This has not always been the case. For example, although our industrial classification has been loosely based on ISIC, and a concordance developed with ISIC, the classification is largely homegrown reflecting the specific interests of Australia and New Zealand. We have agreed to use the 2007 version of ISIC, at least for the upper two levels, with variations at lower levels only where there are specific circumstances that justify it.

There are often pressures on us to divert from international standards. Sometimes this is to make the Australian situation look better. In other cases, such as with the ILO unemployment definition, the pressure is because the international definition does not seem to reflect the real situation in Australian circumstances. We resist these pressures but it is important that we have a well documented international standard as a reference point to justify our position. Nevertheless, where diversions from the international standard are made on an exception basis, they need to be well documented with a clear explanation of the reason. In cases where there is a need to have information on a basis other than the international standard our position is that we should publish statistics on both bases. The headline figure would still reflect international standard as increasingly the Australian situation is being compared with that of other countries and it is important that it is done on a comparable basis. For example, this approach is being taken to satisfy the demand for underemployment data and to reduce criticisms of the ILO unemployment definition.

There is a tension that needs to be managed but if we are serious about the importance of international comparisons it is imperative that international standard is the main guiding light in developing the concepts, sources and methods used in Australia. For these reasons we regard it as a priority to make a significant contribution to the development and revision of international standards.

5. CONCLUSION

We would all agree that attention to quality is a fundamental aspect of our operation. In this paper, we have attempted to show that there are many dimensions to quality. This same message is clear from the frameworks for quality that have been developed by other organizations, such as the IMF, Statistics Canada and Statistics Sweden. The consequence is that a quality organization depends on the actions of all its staff as all can have an impact on quality in one way or another. It cannot be left to a work group with designated responsibility for quality. Therefore, quality can only happen if there is a genuine quality culture within the organization. The paper attempts to describe how we achieve this within the ABS. Nevertheless, it is important to have someone who performs the role of the corporate conscience on quality. We have given this responsibility to the Methodology Division and made the Chief part of the ABS Executive team so that it is easier for key messages to be conveyed to the senior managers. Among other things they draw attention to the most important risks to quality or behaviours they see as contrary to our corporate objectives.

ACKNOWLEDGEMENT

Section 2 was partly based on a document produced by Frank Yu (ABS) for the 9th Conference of South Asian Statistican Offices.

REFERENCES

ALLEN, B. (2001). Qualifying Quality – Issues of Presentation and Education. *Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective.* Statistics Canada.

BRACKSTONE, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*. 25, 2, 129-149.

CARSON, C. (2000). Towards a framework for assessing data quality. The Proceedings of the Statistical Quality Seminar, Jeju Korea, Korean National Statistical Office and International Monetary Fund.

FELLEGI, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*. 64, 165-197.

LEE, G., and ALLEN, B. (2001). Educated Use of Information about Quality. *Bulletin of the International Statistical.*

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 135-143
Statistics Canada

Model Explicit Item Imputation for Demographic Categories

YVES THIBAUDEAU[1]

ABSTRACT

We propose an item imputation method for categorical data based on a MLE derived from a conditional probability model (Besag 1974). We also define a measure for the item non-response error that is useful to evaluate the bias relative to other imputation methods. To compute this measure, we use Bayesian iterative proportional fitting (Gelman and Rubin 1991; Schafer 1997). We implement our imputation method for the 1998 dress rehearsal of Census 2000 in Sacramento, and we use the error measure to compare item imputations between our method and a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) at aggregate levels. Our results suggest that our method gives additional protection against imputation biases caused by heterogeneities between domains of study, relative to the hot-deck.

KEY WORDS: Nearest Neighbor; Conditional probability approach; Bayesian iterative proportional fitting.

1. INTRODUCTION AND BACKGROUND

Let S represent a demographic categorical count requested from a census, or needed to compute a survey statistic, and suppose S can be computed from the records of a survey file f, when the records are complete. Also, suppose f is ordered in such a way that proximity in the order of f corresponds to geographical proximity. Consider the situation where f includes records with unreported items. We propose to estimate S with $d\ (A(f))$, where $A(f)$ is an imputation method that produces a complete survey file, and $d(\cdot)$ estimates S by replacing the unreported items with their values imputed with $A(f)$. $A(f)$ is based on a likelihood that models transitions between two neighbors in f, and associations between the items to be imputed and the relevant domains of study (Cochran 1977, page 34) defined by partitions of the population. $A(f)$ is meant as an advantageous alternative to the popular sequential hot-deck (Kovar and Whitridge 1995), which is a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) that attempts to minimize geographical distance between a unit with unreported items and a suitable imputation donor, while also guaranteeing the distributional homogeneity of the observed and the imputed items with respect to each domain of study. When the domains of a same partition tend not to geographically overlap, borrowing imputation items from a near-by neighbor preserves homogeneity. But, when small domains tend to be dispersed within large domains, the methodologist faces a dilemma. Then, she must choose between hot-deck rules that lead to borrowing the imputed items from geographically close units, leaving the possibility of imputation biases reflecting the local heterogeneity between domains, and domain-specific rules, which guarantee distributional homogeneity by domain, but may not minimize geographical distance. $A(f)$ is an alternative designed to preserve domain integrity, while also simulating the distributional profile of an imputation donor sharing some characteristics with a geographical neighbor. We motivate the design of $A(f)$ with examples and a theoretical description. In this section we review a classification of current hot-deck methods for item imputation with their operating principles, so that we can properly compare them with $A(f)$ in later sections. We also give details on the dress rehearsal of Census 2000 in Sacramento, our test bed throughout the paper.

Fay (1999), and Sande (1981) identify the sequential hot-deck (SHD) as the first category of hot-decks, which we call the "pure" SHD. They add a second category, the fixed-cell hot-deck (FCHD), which we call the pure FCHD. Fay defines a third category of hot-decks: the nearest neighbor hot-deck (NNHD). Chen and Shao (1997, 2000) give an abstract definition of the NNHD in terms of a measure of proximity $|\ \ |$, based on a covariate x. With the NNHD, a "donor" is any unit such that $|x_r - x_d|$ is minimal, where x_r corresponds to the receiving unit (receiver), and x_d corresponds to the provider of the imputations (donor). By constructing the appropriate measure, and defining a suitable x, we recover both the pure SHD and the pure FCHD as special cases of the NNHD. The pure SHD imputes a receiver item by replacing it with the corresponding item from the closest unit for which it was reported, in the order of f. The pure FCHD relies only on the value of variables that we call the class variables to divide the units between post-strata that are homogenous with respect to the items to be imputed. A donor is chosen at random from the same post-stratum as that of the receiver, irrespective of the order of f.

Fay (1999), and Fay and Town (1998) propose the concept of exchangeability to validate the NNHD. For categorical data two units in f are exchangeable if they are uncorrelated and identically distributed, given the

[1] Yves Thibaudeau, Mathematical Statistician, Statistical Research Division, US Census Bureau, 4700 Silver Hill Road, Stop Code 9100, Washington, DC 20233-9100. E-mail: yves.thibaudeau@census.gov.

information available prior to imputing. The operational assumption of the NNHD is that a unit and its nearest neighbor(s) are exchangeable. For the pure SHD it means two contiguous units in *f* are exchangeable. For the pure FCHD it means that units sharing the same values for their class variables anywhere in *f* are exchangeable. We define a third instance of the NNHD, which we call the hybrid sequential hot deck (HSHD). To guarantee exchangeability the HSHD requires proximity both in terms of the order of *f*, and in terms of the class variables.

We use the term "nearest neighbor" in the abstract sense of the NNHD, unless specified otherwise. We use the terms "closest neighbor" to designate the nearest neighbor of the pure SHD, and "closest complete neighbor" to mean the survey unit with no unreported items that is closest in the order of *f*. In the case of the Sacramento dress rehearsal, the Census Bureau uses a HSHD to estimate householder counts by tenure, race, origin (Hispanic origin), and sex. The householder, usually an adult, is unique for each housing unit, and is determined by the ages, relationships, and order of the persons on the census questionnaire. The HSHD substitutes unreported items with the values of these items corresponding to the last householder who reported them and is in the same post-stratum (Treat 1994). The sorted order of *f* maintains the proximity of geographical neighbors. The intent behind the HSHD is to define nearest neighbors who are close, both in geography and "in kind". Throughout the paper, we continue to use the term householder, although its meaning may extend to a generic survey unit.

The design of the HSHD is well suited for item imputation in populations geographically clustered by domain. Then the need for class variables is limited. But difficulties arise when the geographical boundaries between the domains begin to blur. Designing a HSHD with good discrimination power in those conditions is an attempt at walking a fine line between specifying enough class variables to account for heterogeneities between domains, and specifying too many, which could yield post-strata so narrowly defined in terms of domain that they don't capture the local geographical character of the receivers. Complicating the situation is the fact that the demographic composition of the population may change as the geography changes, and thus a particular scheme for the HSHD might need to be revised, as the geography changes. In the face of these difficulties $A(f)$ is innovative in the sense that, instead of searching for an ideal nearest neighbor, it generates imputations through a model-based simulation that integrates information relating to the local geography, as well as to domain partitions. $A(f)$ integrates both kind of information by calibrating the parameters of a log-linear model on the basis of the strength of the correlations between the covariates and the variables subject to imputation. Our parameter estimation strategy is the same as that of Zanutto and Zaszlavsky (1995a, b). However, because they have access to a representative sample of complete non-respondents, these authors can obtain estimates of the imputation probabilities by implementing a one-step EM algorithm (Dempster, Laird and Rubin 1977). In our situation, we don't assume access to a representative sample, and we implement the full EM algorithm. Implicitly we make an assumption of items "missing at random" (MAR) (Little and Rubin 1987, page 16).

To analyze the results obtained with $A(f)$, and to compare them with those of the HSHD, we derive error measures related to $A(f)$ based on approximations computed using a Bayesian algorithm first introduced by Gelman and Rubin (1991). There are fundamental objections to Bayesian methodologies. Fay (1992) shows that variance estimation based on multiple imputations (Rubin 1996) can lead to inflated estimates of variance, whereas in the same situation the jackknife estimator (Rao and Shao 1992) avoids biases. Meng (1994) suggests that Fay's example stems from a poor communication between an imputer who has specific model information, and an analyst who only has knowledge of the estimation process. In the language of Meng, this situation is uncongenial. While requirements for coordination between imputer and analyst are restrictive, imputation based on exchangeability also has dangerous pitfalls, as we show in section 2. In addition the Bayesian approach allows for asymptotic approximations of error measures through mechanical algorithms, while a strict frequentist approach might require tedious expansions, as we show in section 5.

Our objective is to present $A(f)$, and to show its comparative advantages over the HSHD, using the Sacramento dress rehearsal as an example. In this case *f* contains records for the 138,271 physically enumerated householders (Kostanich 1999), of whom 90,156 returned a census questionnaire by mail or were visited by an enumerator at a first attempt, and 48,115 were selected in a sample. We implement our method at the level of the tract, a connected unit of geography containing on average 1,300 householders in *f*.

The paper is organized as follows. In section 2 we illustrate the difficulties of designing a HSHD methodology that guarantees exchangeability. In section 3, we define $A(f)$, and in section 4 we present a likelihood for the model parameters. In section 5, we show how to implement $A(f)$ and derive a measure of error to make comparisons with the HSHD. Section 6 presents and motivates the basic model for the dress rehearsal, and section 7 gives results for both $A(f)$ and the HSHD in this case. In section 8, we summarize the differences and we make recommendations.

2. ASSESSING EXCHANGEABILITY WITH RESPECT TO A PARTITION BY DOMAINS OF STUDY

We illustrate the difficulties inherent in designing a HSHD that preserves exchangeability between domains of study (Cochran 1977, page 34) with an example, where tenure (ownership) is the measurement of interest, and the relevant domains of study are defined by race. To impute tenure, the Census Bureau uses the class variable "household type" to post-stratify *f* in five post-strata defined by the presence/absence of a live-in spouse for the householder, and the size of the household (1, 2, 3+) (Wilson 1998). The intent is to define post-strata that establish distributional homogeneity in terms of ownership at the level of the post-stratum, rendering the domain boundaries of a relevant partition uninformative within each post-stratum.

We examine the post-stratum comprising all the householders without a live-in spouse, and living in households of 3 or more. We call it post-stratum 3. For the purpose of this example, we have removed from *f* all the householders with unreported tenure, and each nearest neighbor is exclusive to a single householder. Table 1 gives householder frequencies for eight exhaustive race-tenure categories for post-stratum 3. Table 1 also gives the rate of ownership for their nearest neighbors, cross-classified by their race and by the same eight race-tenure categories of the corresponding householders. We observe that, on average, when a householder is either in the Black-owner or in the Black-renter category, his nearest neighbor is at least 25% more likely to be an owner when this nearest neighbor is White, than when he is Black. It is tempting to explain this differential rate by geographical differences. However, table 2, which shows the rates of ownership of the householders in post-stratum 3, cross-classified by their own race and that of their nearest neighbors, reveals that in fact Blacks with White nearest neighbors have a slightly lower rate of ownership than Blacks with Black nearest neighbors. What this means is that, if the probability of not reporting tenure is constant for all Blacks, then imputing their tenure by substituting the tenure of their nearest neighbor overestimate ownership for Blacks in post-stratum 3.

These distributional disparities between householders and their nearest neighbors reflect a lack of exchangeability. A McNemar test leads to a formal rejection of the exchangeability hypothesis. There are 1,784 Black householders with White nearest neighbors. In 1,187 instances, tenure is tied. Among the 597 non-tied cases, the owner is White in 396 cases. Under the exchangeability hypothesis, ownership goes to either race with probability one-half. But the proportion of Whites among the owners is eight standard deviations above one-half. This example illustrates the difficulties in designing a valid NNHD that maintains exchangeability. In the next section we present our imputation method, which is devised for this type of situation.

3. AN IMPUTATION METHOD BASED ON DEMOGRAPHIC TRANSITION PROBABILITIES

Besag (1974) describes the conditional probability approach to spatial processes. This approach gives a framework for probabilistically modeling the values of "sites", in terms of the values of their "neighbours" to construct a spatial process. Besag (1974) also suggests making a unilateral approximation to simplify this construction. Then, the value of each site depends only on a finite number of "predecessors". This approach is natural in our situation since *f* provides a unilateral ordering of householders who play the roles of sites and predecessors, in turn. Specifically, we construct a first-order process where each householder is a site, and the complete closest neighbor is

Table 1
Number of Householders and Rates of Ownership of the Nearest Neighbors in Post-Stratum 3 by Race of the Nearest Neighbor and Joint Race and Tenure of the Householder

	Race-Tenure Category of the Householder							
	White Owner	White Renter	Black Owner	Black Renter	Asian Owner	Asian Renter	Other Owner	Other Renter
Number of Householders in Post-Stratum 3	3,347	5,197	1,319	3,630	872	1,196	681	1,637
Rate of Ownership of the White Nearest Neighbors	0.556	0.564	0.562	0.299	0.561	0.287	0.540	0.163
Rate of Ownership of the Black Nearest Neighbors	0.379	0.189	0.427	0.211	0.443	0.202	0.471	0.158
Rate of Ownership of the Asian Nearest Neighbors	0.589	0.332	0.667	0.320	0.668	0.262	0.535	0.302
Rate of Ownerships of the Other Nearest Neighbors	0.423	0.251	0.497	0.237	0.595	0.177	0.463	0.152

Table 2
Rates of Ownership of the Householders in Post-Stratum 3 by Race of the Householder and Race of the Nearest Neighbor

	Race of the Nearest Neighbor			
	White	Black	Asian	Other
Rate of Ownership of the White Householders	0.415	0.358	0.384	0.337
Rate of Ownership of the Black Householders	0.257	0.264	0.304	0.267
Rate of Ownership of the Asian Householders	0.441	0.441	0.400	0.360
Rate of Ownership of the Other Householders	0.309	0.297	0.337	0.234

its only predecessor. In this set-up, the value of a site is the state of a householder, which we define shortly. We refer to the conditional probability for the value of a site given that of its predecessor as the transition probability from the state of the closest complete neighbor to the state of the householder. Our imputation methodology is based on the MLE of the transition probabilities at the level of a tract. In this section we describe the imputation methodology, and in the next section we introduce a likelihood for the transition probabilities.

Consider a population of householders in f representing a tract. Let Ψ represent a set of C categorical variables that characterize each householder. The variables are labeled $1, ..., C$, and have respectively $K_1, ..., K_C$ categories. Let $\Psi^\times$ denote the Cartesian product of the categorical variables in Ψ. Then, $\Psi^\times$ is the state space of the householder and has K states, where $K = \Pi_{i\in\Psi} K_i$. Similarly, let Ξ be the set of E categorical variables defining the closest complete neighbor in f. The variables are labeled $1, ..., E$, and have $F_1, ..., F_E$ categories. $\Xi^\times$ is the state space of the closest complete neighbor and has F states, where $F = \Pi_{i\in\Xi} F_i$. The items represented in Ξ are also represented in Ψ. Let the state of the householder be $s \in \Psi^\times$, where s is a vector whose components represent the variables in Ψ. Similarly, $t \in \Xi^\times$ is the state of the closest complete neighbor. Under the assumptions above, let $P(s\,|\,t)$ represents the transition probability from t to s in the order of f. Now suppose a householder only reported the categorical variables in a subset $Z \subset \Psi$. Let $v \in Z^\times$ be the vector of reported variables. Let $\sigma(\Psi, Z, v) \subset \Psi^\times$ be the subset containing all the values of s, such that s agrees with v on the variables in Z. Define

$$P(s|t, Z, v) = \frac{P(s|t)}{\sum_{u\in\sigma(\Psi,Z,v)} P(u|t)}; \quad s\in\sigma(\Psi, Z, v). \tag{1}$$

To impute the items in the set difference $\Psi - Z$ according to $A(f)$, we roll dice weighted by the values of the MLE of $P(s\,|\,t, Z, v)$, for each householder in marginal state v and with closest complete neighbor in state t. Under our assumptions, the MLE of $P(s\,|\,t, Z, v)$ contains all the information available from f on the unreported items. In the next section we formulate a likelihood for $P(s\,|\,t, Z, v)$.

4. A LIKELIHOOD FOR THE TRANSITION PROBABILITIES

Let $N(t, Z, v)$ be the number of householders who only reported the items defining the marginal state v involving only the items in $Z \subset \Psi$, and with closest complete neighbor in state t. Let N be a vector with the $N(t, Z, v)$'s as its components, at the level of a tract. Let $P = [P(s\,|\,t)]$ be the vector comprising the $P(s\,|\,t)$'s ordered lexicographically by t and s. Based on the assumptions described above, we have the following likelihood for the transition probabilities.

$$L(N;P) = \prod_{t\in\Xi^\times}\prod_{Z\subset\Psi}\prod_{v\in Z^\times}\left(\sum_{s\in\sigma(\Psi,Z,v)} P(s|t)\right)^{N(t,Z,v)}; \quad P\in\Theta_P. \tag{2}$$

The running indices in (2) are t, Z, v, and s. If every item is reported, then Ψ is the only instance of Z with $N(t, Z, v) \neq 0$, for some t and v. In that case (2) is analogous to the likelihood of the transition probabilities of a first-order Markov chain (Bishop, Fienberg and Holland 1975 page 263). In general, we model Θ_P as a log-linear subspace. For this purpose it is more convenient to work with an expression equivalent to (2) that has a simpler algebraic representation. We introduce the nuisance parameter $U = [U(t)]$, where U is a probability vector, that is $\sum_{t\in\Xi^\times} U(t) = 1$, and $0 < U(t) < 1$, for all $t \in \Xi^\times$. U represents the prevalences of the states of the closest complete neighbors. Let $Q(s,t) = U(t) \times P(s\,|\,t)$, and $Q = [Q(s,t)]$. Then Q is a probability vector with $K\times F$ components lexicographically ordered by t and s. We set up Θ, the parameter space of Q, as a hierarchical log-linear model (Agresti 1990, page 143; Bishop, Fienberg and Holland 1975, page 67). Then, if we design Θ so that it includes the interactions of all orders between the variables in Ξ, (2) is equivalent to the following likelihood in terms of Q.

$$L^*(N;Q) = \prod_{t\in\Xi^\times}\prod_{Z\subset\Psi}\prod_{r\in Z^\times}\left(\sum_{s\in\sigma(\Psi,Z,r)} Q(s,t)\right)^{N(t,Z,r)}; \quad Q\in\Theta. \tag{3}$$

That is, if Θ has the architecture described above, a specific choice for Θ unambiguously defines Θ_P in (2), and since the items of the closest complete neighbor are always reported, the factorization $L(N\,;P) = L^*(N;Q)\times R(N;U)$ holds, for some $R(;)$. (3) is easier to manipulate than (2) since it corresponds to the likelihood of the cell probabilities associated with a partially classified contingency table (Little and Rubin 1987, page 181). Under mild conditions on the non-response mechanism (for example, strictly positive and constant probabilities for each response configuration (Thibaudeau 1988)) the likelihoods in (2) and (3) are identifiable and asymptotically unimodal. Multimodality is theoretically possible for finite samples, but it does not appear to occur in the cases studied in the paper, where the proportions of unreported items are small.

5. FINDING THE MLE AND DERIVING MEASURES FOR THE NON-RESPONSE ERROR

In this section, we recall how to compute $\hat{P}$, the MLE of P, and we derive measures of errors for $A(f)$ and another predictor $\hat{S}(s)$, which we term the "MLE" of the expected value of $S(s)$, which is the actual count of householders in state s at the tract level. An error measure for $\hat{S}(s)$ will be useful in section 7 to evaluate the imputation results obtained with $A(f)$ relative to those with the HSHD. We compute $\hat{P}$ by maximizing (3), in terms of Q, with the EM algorithm. Because of the factorization described in section 4, this maximum also yields $\hat{P}$.

To derive measures of error in predicting $S(s)$ for a given s, consider all the triples of the form (t, Z, v) in (1) that are observed in the sample (*i.e*, the tract) for which it is possible, but due to item non-response it is not known, that one or more householders corresponding to such a triple are in state s. Let $\Lambda(s)$ be the number of such triples. We index these triples with $\lambda = 1, ..., \Lambda(s)$. Let $\delta(\lambda)$ be the number of householders corresponding to triple λ, and let $\rho_\lambda(s)$ be the probability that such a householder is indeed in state s, where $\rho_\lambda(s)$ is derived from P. Let $\Delta(s,\lambda)$ be the unknown number of householders who are indeed in state s among the $\delta(\lambda)$ candidates. Based on our model we have $S(s) = S_{\text{obs}}(s) + \sum_{\lambda=1}^{\Lambda(s)} \Delta(s,\lambda)$, where $S_{\text{obs}}(s)$ is the number of householders who reported being in state s and $\Delta(s,\lambda)$ is Binomial$(\delta(\lambda), \rho_\lambda(s))$. Furthermore, let $\hat{S}(s) = S_{obs}(s) + \sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda)\hat{\rho}_\lambda(s)$, where $\hat{\rho}_\lambda(s)$ is the MLE of $\rho_\lambda(s)$. If we treat the λ's as independent predictors, like in a regression situation, and since $\hat{P}$ is asymptotically normal with mean P, we have the following large sample approximation for the MSE of $\hat{S}(s)$ in predicting $S(s)$.

$$E\left[\left(\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda)\hat{\rho}_\lambda(s) - \Delta(s,\lambda)\right)^2 | P\right]$$

$$\approx \mathrm{V}\left(\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda)\hat{\rho}_\lambda(s) \,|\, P\right) + \mathrm{V}\left(\sum_{\lambda=1}^{\Lambda(s)} \Delta(s,\lambda) \,|\, P\right). \quad (4)$$

Let V_p and V_ε be the first and second variances on the RHS of (4). Gelman and Rubin (1991), Larsen (1996), and Schafer (1997, page 324) introduce data augmentation Bayesian iterative proportional fitting (DABIPF) to simulate posterior and predictive distributions associated with log-linear models with data missing at random. We can use DABIPF to approximate model-consistent estimators for V_p and $V_\varepsilon + V_p$ through simulations of the posterior distribution of $\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda)\rho_\lambda(s)$ and the predictive distribution of $S(s)$ respectively. Furthermore, we approximate the MSE of the demographic counts obtained imputing with $A(f)$ by adding another V_ε to $V_\varepsilon + V_p$ in (4) to account for the additional noise of the "dice roll" involved in $A(f)$.

6. MODELING AND SENSITIVITY ANALYSIS

6.1 A Conditional Independence Model for Sacramento

Using the notation of section 3, the householder variables in Ψ are race, origin, tenure, and sex. The categories for race are White, Black, Asian, and Other. For origin they are Hispanic and non-Hispanic. For tenure they are owner and renter. For sex they are male and female. The neighbor variables in Ξ are race, origin, and tenure. The categories for race of the neighbor are Black and non-Black. The categories for origin and tenure of the neighbor are the same as for the householder. We design Θ in (3), by selecting interactions between the variables in Ψ and Ξ. To ensure equivalence between (2) and (3), we select the interactions of all orders between the variables in Ξ. We attempt to maintain through the imputations the correlation between successive householders in f in terms of each item in Ξ. Thus we include each interaction associating an item in Ξ to the corresponding item in Ψ. We complete the model by selecting consistency associations: We include the six interactions representing the associations involving a pair of items in Ψ. The resulting contingency table has 256 cells, and the log-linear model has thirty free parameters.

This model leads to a conditional independence transition structure. For example, conditional on the race of the closest complete neighbor, the race of the householder is independent of the tenure of the closest complete neighbor. Conditional independence allows us to combine neighbor information obtained from multiple neighbors to produce a synthetic closest complete neighbor. This approach ensures that we can use all the information available from the closest neighbor, even if he is not complete. With this approach, the correlation structure among the items of the householder is maintained whenever only one item per householder is imputed. In Sacramento, among 138,271 householders, approximately 0.1% did not report sex, 3.5% did not report race, 2.9% did not report origin, and 7.6% did not report tenure. Furthermore, race and origin are missing jointly for 0.49% of the householders, race and tenure 0.48%, origin and tenure 0.69%. Given these low rates of jointly missing items, we expect our model to do well.

6.2 Sensitivity Analysis and Evaluation

In section 7 we use the standard error of the predictive distribution of $S(s)$ to approximate $\sqrt{V_\varepsilon + V_p}$, the error of $\hat{S}(s)$ in predicting $S(s)$, as derived in (4), and we assume asymptotic normality of $\hat{S}(s) - S(s)$. The accuracy of this approximation depends on the accuracy of the approximation of the distribution of the MLE $\hat{P}$ with the posterior distribution of P. This later approximation is accurate asymptotically when the model holds, but we still need to verify the extent to which this asymptotic result is applicable when the sample is finite. To do so we examine the sensitivity of the posterior distribution of P under prior changes. A low sensitivity implies that the posterior distribution of P is a good approximation of the distribution of $\hat{P}$. We focus on the posterior distribution for the conditional probability that origin is Hispanic, conditional on each race. An increase of .1 in the value of α, the prior parameter of the constrained Dirichlet family (Schafer 1997, page 346), which is the natural family for (3), is equivalent to observing three additional Hispanics and three additional Non-Hispanics of each race. Table 3 gives the posterior modes and standard deviations (SD) of the posterior density of the conditional probability that origin is Hispanic given each race, for four choices of α, for a specific tract X. Figure 1 shows the posterior of the conditional probability given race is White. This posterior is stable under prior disturbances and we expect it to give a good approximation for the distribution of the corresponding MLE. On the other hand, Figure 2, which shows the posterior of the conditional probability given race is Black, displays a high sensitivity, suggesting that our proposed asymptotic approximation is less accurate in this case. This is not surprising in light of the facts that, for Blacks, the MLE of the conditional probability is close to 0 and the domain (race) size is smaller (among the 1,583 householders in tract X, there are 1,087 Whites, 179 Blacks, 56 Asians, 172 Others, while 89 did not report race). In the next section we focus on cases where the conditional probabilities are not near 0 or 1, and the size of the domain is large. We retain the choice α = .01 for the prior, which is approximately Jeffrey's prior on the marginal conditional probabilities that define the model. It is beyond the scope of the paper to address the difficulties when the domain is small and/or the MLE is near 0/1.

Table 3
MLE, Posterior Mode (approximate), and Standard Deviation for the Conditional Probabilities of Origin Being Hispanic Given Race for Four Choices of Prior Distribution

Race	MLE	Mode α=.01	S.D. α=.01	Mode α=.1	S.D. α=.1	Mode α=.5	S.D. α=.5	Mode α=1	S.D. α=1
White	.1784	.178	.01195	.184	.01247	.180	.01219	.188	.01186
Black	.07428	.0690	.02272	.081	.02330	.120	.02428	.160	.02782
Asian	.09113	.105	.04086	.108	.04550	.195	.04881	.276	.04952
Other	.9662	.966	.01171	.964	.01347	.950	.01495	.930	.01666

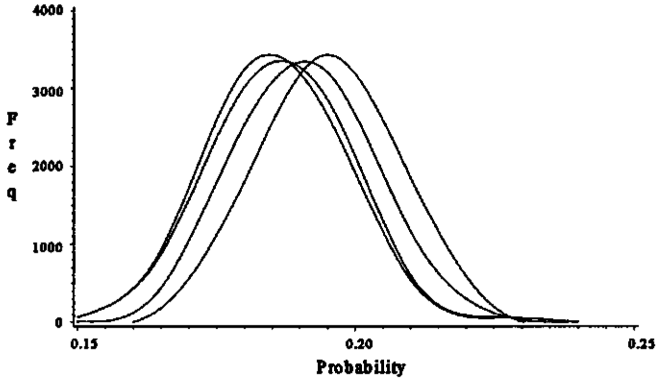

Figure 1. Posterior Distribution
Prob. Origin is Hispanic – White Householder

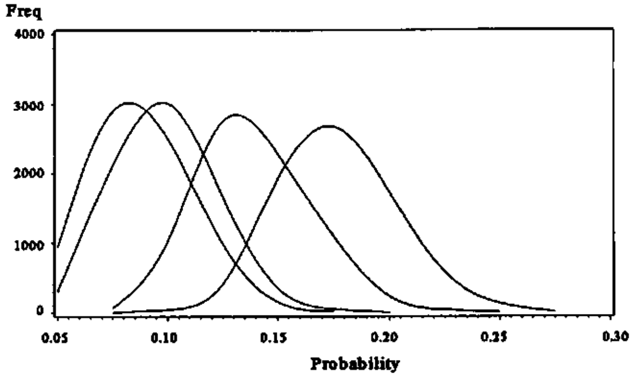

Figure 2. Posterior Distribution
Prob. Origin is Hispanic – Black Householder

7. RESULTS FOR THE SACRAMENTO DRESS REHEARSAL

Table 4 gives count estimates at the level of Sacramento derived with $A(f)$ based on the model of section 6.1 fitted for each of the 102 tracts, as well as count estimates obtained with the HSHD. Table 4 also gives error measurements based on a sequence of 2000 DABIPF iterations with 2000 burn-in iterations, for each of the 102 tracts in Sacramento (see appendix A for convergence), serving to approximate $\sqrt{V_\varepsilon + V_p}$ derived from (4). We call $\sqrt{V_\varepsilon + V_p}$ the prediction error of the MLE. We estimate $\sqrt{V_\varepsilon}$ separately by "rolling dice" loaded with the MLE. We call $\sqrt{V_\varepsilon}$ the model residual error. We use $\sqrt{2\,V_\varepsilon + V_p}$, which we call the total imputation error, to express the error of $A(f)$ in estimating the true count. If we assume $\hat{S}(s)$ is positively correlated with the HSHD, the prediction error of the MLE can be used as an upper bound for the standard error of the distance between the count estimates corresponding to the MLE and the HSHD. For the Black owners, this distance is severely incompatible with the hypothesis that the MLE and the HSHD have the same expectation. This is no surprise in light of the results of section 2.

Interestingly, the results of table 4 can serve to improve the performance of the HSHD. Since tenure is unreported twice as often as race, our results for the Black owners suggest improving the HSHD by including race as a class variable for the imputation of tenure with the HSHD. Table 5 shows results obtained with this re-engineered HSHD, and exchangeability of tenure between nearest neighbors based on this new post-stratification is more plausible than for the original scheme.

Table 4
Population Counts and Uncertainty Measures for Sacramento

	Imputed Count With HSHD	Imputed Count With Model	MLE of the Expected Count	Model Residual Error	Prediction Error of the MLE	Total Imputation Error
All	138,271	138,271	138,271.0	0.0	0.0	0.0
White	89,032	88,914	88,927.7	31.5	35.2	47.2
Black	19,962	19,943	19,952.9	14.9	16.5	22.3
Asian	17,405	17,421	17,426.2	14.0	14.9	20.5
Other	11,872	11,993	11,964.1	29.8	33.5	44.8
Hispanic	21,024	21,050	21,038.1	10.3	10.6	14.7
Non-Hispanic	117,247	117,221	117,232.8	10.3	10.6	14.7
Owner	70,054	70,022	70,026.3	42.8	43.3	60.9
Renter	68,217	68,249	68,244.7	42.8	43.3	60.9
White Hispanic	9,068	8,972	8,991.1	29.9	33.6	45.0
White Non-Hispanic	79,964	79,942	79,936.6	15.4	15.7	22.0
Black Hispanic	605	612	608.6	11.0	12.6	16.7
Black Non-Hispanic	19,357	19,331	19,344.3	10.8	10.7	15.2
Asian Hispanic	518	515	516.5	10.0	11.5	15.2
Asian Non-Hispanic	16,887	16,906	16,909.7	10.4	10.3	14.6
Other Hispanic	10,833	10,951	10,921.9	29.7	33.3	44.6
Other Non-Hispanic	1,039	1,042	1,042.3	3.5	3.4	4.9
White Owner	47,722	47,767	47,770.5	37.8	41.3	56.0
White Renter	41,310	41,147	41,157.3	39.0	41.4	56.9
Black Owner	7,661	7,538	7,542.3	19.6	20.7	28.5
Black Renter	12,301	12,405	12,410.6	21.1	22.5	30.8
Asian Owner	9,810	9,853	9,872.8	18.4	18.6	26.1
Asian Renter	7,595	7,568	7,553.4	18.2	18.8	26.1
Other Owner	4,861	4,864	4,840.7	24.4	28.2	37.3
Other Renter	7,011	7,129	7,123.4	25.4	28.6	38.2
Hispanic Owner	9,409	9,434	9,402.2	19.5	20.9	28.6
Hispanic Renter	11,615	11,616	11,629.9	20.1	21.4	29.4
Non-Hispanic Owner	60,645	60,588	60,618.0	38.9	39.4	55.4
Non- Hispanic Renter	56,602	56,633	56,614.8	38.7	39.6	55.4

Table 5
HSHD with Race as an Additional Class Variable

	Imputed Count with HSHD	Imputed count with HSHD re-engineered with Race as a Class Variable	Imputed Count with Model	MLE of the Expected Count	Prediction Error of the MLE
White owner	47,722	47,687	47,767	47,770.5	41.3
Black Owner	7,661	7,573	7,538	7,542.3	20.7
Asian Owner	9,810	9,851	9,853	9,872.8	18.6
Other Owner	4,861	4,840	4,864	4,840.7	28.2
Owner	70,054	69,951	70,022	70,026.3	43.3

8. CONCLUSION

In section 2 we have shown that the HSHD may fail to retrieve exchangeable householders, producing a bias relative to a situation where exchangeability holds. As more evidence that $A(f)$ partly corrects this relative bias, we compare the observed and the imputed cross-product ratios (Bishop, Fienberg and Holland 1975, page 14) between two races (Black, White) and the two tenures. We look at the cross product ratio involving:

1. Only observed householders.
2. Householders with tenure imputed with the HSHD.
3. Householders with tenure imputed with $A(f)$.

There are 73 tracts where all these cross-product ratios can be measured. 2. The HSHD produces cross-product ratios smaller than those observed for 53 tracts. $A(f)$ displays more symmetry as it produces cross-product ratios smaller than observed only for 43 tracts. A sign test confirms that $A(f)$(p =.064) is more in sync with the observations than the HSHD (p =.0001).

In general, we expect the HSHD to give good count estimates when the householders tend to geographically coalesce by domain of study. But difficulties arise in a situation where domains of study exhibiting substantial distributional dissimilarities are geographically integrated. In such a situation, implementing the HSHD requires accurate parsing of the class variables. Frugality is tantamount when specifying class variables, but at the same time the price to pay for omitting a crucial variable can be substantial. Thus the designer of the HSHD has little room for error. By contrast, although model misspecification certainly remains a danger, the user of $A(f')$ has more freedom to posit several domain partitions without impeding on the ability of $A(f)$ to adjust the imputations for the local geographical character, based on information from the closest complete neighbor. $A(f)$ will be useful to impute categorical measurements when the impact of the relevant domain partitions on the measurements is not known a priori, and some of the relevant domains may define small subpopulations dispersed within the entire population. Then, based on policy considerations, $A(f)$ can be applied directly, or to help parse the class variables of the HSHD, as we did in section 7.

A referee notes that a comparison with a procedure based on an unbiased sample, building on the method of Zanutto and Zaslavsky (1995a,b), would be a defining test for $A(f)$. This procedure would require collecting information from the item non-respondents on a scale sufficiently large to ensure bias detection, and we should take advantage of any such opportunity to perform a test of this type. Unfortunately, because of limited resources, samples containing this information are seldom collected. Nevertheless, we are hopeful that the analysis of the returns from Census 2000 aided with procedural information can provide new insights on the reliability of $A(f)$.

ACKNOWLEDGEMENTS

The author is indebted to William Winkler for his guidance. The author is grateful to two referees for their discernment, to Eric Slud, Don Malec and Joseph Schafer for essential discussions, and to Andrew Gelman and Don Rubin for providing their unpublished paper. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

APPENDIX A – CONVERGENCE OF DABIPF

We ran two chains of 8,000 iterations each, with over-dispersed starting points, for the case $\alpha = 0.01$, for tract X. We computed $\sqrt{\hat{R}}$ (Gelman and Rubin 1992) for $Q(s,t)$ in (3), for sequences of 1,000, 2,000, and 4,000 iterations, after burn-in lags of 1,000, 2,000, and 4,000 iterations respectively. After 2,000 iterations, with 2,000 burn-in iterations, we observed that $\sqrt{\hat{R}} \leq 1.010$ in all studied cases, including those in table 3. We think this level of accuracy is acceptable for approximating modes and variances.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis.* Wiley-Intersience.

BESAG, J. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society, B.* 36, 2.

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

CHEN, J., and SHAO, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. *Proceedings for the Section on Survey Research Methods,* American Statistical Association. 365-369.

CHEN, J., and SHAO, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics.* 16, 2.

COCHRAN, W.G. (1977). *Sampling Techniques.* 3rd Edition. Wiley.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B.* 39, 1-22.

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

FAY, R.E. (1999). Theory and application of nearest neighbor imputation in census 2000. *Proceedings for the Section on Survey Research Methods,* American Statistical Association. 112-121.

FAY, R.E., and TOWN, M.K. (1998). Variance estimation for the 1998 census dress rehearsal. *Proceedings for the Section on Survey Research Methods,* American Statistical Association. 605-610.

GELMAN, A., and RUBIN, D.B. (1991). Simulating the Posterior Distribution of Loglinear Contingency Table Models. Unpublished Technical Report, Harvard University.

GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science.* 7, 4.

KOSTANICH, D.L. (1999). DSSD Census 2000 Dress Rehearsal Memorandum Series #A, US Bureau of the Census.

KOVAR, J.G., and WHITRIDGE, P.J. (1995). Imputation of Business Survey Data. *Business Survey Methods,* (Eds. Cox, D. Binder, Chinnappa, Christianson, M. Colledge and Kott). Wiley.

LARSEN, M.D. (1996). *Bayesian Approaches to Finite Mixture Models.* Doctoral Dissertation, Department of Statistics, Harvard University.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley.

MENG, X.M. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 9, 4.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 79, 4.

RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 91, 434.

SANDE, I.G. (1981). Imputation in surveys: coping with reality. *Survey Methodology*. 7, 21-43.

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

THIBAUDEAU, Y. (1988). *Approximating the Moments of a Multimodal Posterior Distribution with the Method of Laplace*. Doctoral Dissertation, Department of Statistics, Carnegie Mellon University.

TREAT, J.B. (1994). *Summary of the 1990 Census Imputation Procedures for the 100 % Population and Housing Items*. DSSD REX Memorandum Series BB-11, US Bureau of the Census.

WILSON, E.B. (1998). Communication to Dan E. Philip. Housing and Household Economics Statistics Division, US Bureau of the Census.

ZANUTTO, E., and ZASLAVSKY, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 608-613.

ZANUTTO, E., and ZASLAVSKY, A.M. (1995b). Models for imputing nonsample households with sampled nonresponse follow-up. *Proceedings of the Annual Research Conference*, U.S. Department of Commerce, Bureau of the Census. 673-686.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 145-156
Statistics Canada

A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas

BALGOBIN NANDRAM, GEUNSHIK HAN and JAI WON CHOI[1]

ABSTRACT

The analysis of survey data from different geographical areas, where the data from each area are polychotomous, can be easily performed using hierarchical Bayesian models even if there are small cell counts in some of these areas. However, there are difficulties when the survey data have missing information in the form of nonresponse especially when the characteristics of the respondents differ from the nonrespondents. We use the selection approach for estimation when there are nonrespondents because it permits inference for all the parameters. Specifically, we describe a hierarchical Bayesian model to analyze multinomial nonignorable nonresponse data from different geographical areas, some of them can be small. For the model, we use a Dirichlet prior density for the multinomial probabilities and a beta prior density for the response probabilities. This permits a "borrowing of strength" of the data from larger areas to improve the reliability in the estimates of the model parameters corresponding to the smaller areas. Because the joint posterior density of all the parameters is complex, inference is sampling based and Markov chain Monte Carlo methods are used. We apply our method to provide an analysis of body mass index (BMI) data from the third National Health and Nutrition Examination Survey (NHANES III). For simplicity, the BMI is categorized into three natural levels, and this is done for each of eight age-race-sex domains and thirty-four counties. We assess the performance of our model using the NHANES III data and simulated examples, which show our model works reasonably well.

KEY WORDS: Latent variable; Metropolis-Hastings sampler; Nonignorable nonresponse; Selection approach; Small area.

1. INTRODUCTION

The nonresponse rates in many surveys have been increasing steadily (De Heer 1999; Groves and Couper 1998), making the nonresponse problem more important For many surveys the responses are polychotomous. For example, in the third National Health and Nutrition Examination Survey (NHANES III), we can estimate the proportions of persons belonging to three levels of body mass index (BMI), although BMI is a continuous variable. The purpose of this paper is to describe a new hierarchical Bayesian model to study nonignorable multinomial nonresponse for small areas, and to apply it to the NHANES III BMI data.

Rubin (1987) and Little and Rubin (1987) describe two types of models which differ according to the ignorability of response. In the ignorable nonresponse model the distribution of the variable of interest for a respondent is the same as the distribution of that variable for a nonrespondent with the same values of the covariates. In addition, the parameters in the distributions of the variable and response must be distinct (see Rubin 1976). All other nonresponse models are nonignorable. We use both ignorable and nonignorable nonresponse models for our data because there are no nonrespondents for some domains.

Crawford, Johnson and Laird (1993) used nonignorable nonresponse models to analyze data from the Harvard Medical Practice Survey. Stasny, Kadane, and Fritsch (1998) used a Bayesian hierarchical model for the probabilities of voting guilty or not on a particular trial when the views of nonrespondents differ from those of respondents in various death-penalty beliefs. Park and Brown (1994) used a pseudo-Bayesian method (Baker and Laird 1988), and Park (1998) applied a method in which prior observations are assigned to both observed and unobserved cells to estimate the missing cells of a multi-way categorical table under nonignorable nonresponse. Our approach differs from these authors. We describe small area estimation for multinomial data, and we use Markov chain Monte Carlo methods to implement the methodology. This permits the inclusion of all sources of variability in our models.

There are two approaches to model nonresponse. The selection approach is used for the hypothetical complete data, and a nonresponse model is added conditional on the hypothetical data. This approach was developed to study sample selection problems (*e.g.*, Heckman 1976 and Olson 1980). In the pattern mixture approach the respondents and the nonrespondents are modeled separately, and the final answer is obtained by a probabilistic mixture of the two. We use the selection approach for our problem.

Stasny (1991) used an empirical Bayes model to study victimization in the National Crime Survey, and she followed the selection approach. This analysis pools binomial data from several domains, and some of them have small counts. Essentially this is an exercise in small area estimation. A related method was presented by Albert and

[1] Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA (balnan@wpi.edu); Geunshik Han, Division of Computer and Information Science, Hanshin University, Osan, Korea, (gshan@hucc.hanshin.ac.kr); Jai Won Choi, National Center for Health Statistics, Room 915, 6525 Belcrest Road, Hyattsville, MD 20782, USA, (jwc7@cdc.gov).

Gupta (1985), who used an approximation to obtain a Bayesian approach for a population with a single domain (see also Kaufman and King 1973). That is, unlike Stasny (1991), these latter authors did not perform. small area estimation, and their analysis in a single domain do not use data from other domains.

Since the Bayesian approach can incorporate other information about nonrespondents, the Bayesian method is appropriate for the analysis of nonignorable nonresponse (Little and Rubin 1987 and Rubin 1987). However the main difficulty is how to describe the relationship between the respondents and nonrespondents. Using the selection approach within the framework of Bayes empirical Bayes (see Deely and Lindley 1981), Stasny (1991) estimated the hyper-parameters by maximum likelihood methods and then assumed them known, thereby suppressing some variability. We extend this approach in two directions.

First, we consider multinomial data obtained independently from several geographical areas. It is worthy to note that Basu and Pereira (1982) considered multinomial nonresponse data from a single domain using a multinomial Dirichlet model when the hyper-parameters are assumed known. Recently, Forster and Smith (1998) used graphical multinomial Dirichlet log-linear models to analyze data from the panel survey in British general election. Again the hyper-parameters are assumed known, and a model with a single domain is used. Secondly, we obtain a full Bayesian approach for multinomial nonignorable nonresponse data from several areas. We do not estimate the hyper-parameters using the data.

As a summary, we develop a multinomial nonignorable nonresponse model which is used for pooling data over many small areas, and we note that it can be used in other applications. The rest of the paper is organized as follows. In section 2 we describe the NHANES III. In section 3 we discuss the Bayesian model for nonignorable nonresponse. In particular, a three-stage Bayesian hierarchical multinomial model is applied to the NHANES III data to investigate the nonresponse problem. In section 4 we describe an analysis of the NHANES III data in which we include a regression analysis to combine all the age-race- sex domains. In section 5 we describe a simulation study to assess the performance of our model. Finally, section 6 has the conclusion.

2. NHANES III DATA AND NONRESPONSE

The NHANES III is one of the periodic surveys used to assess an aspect of health of the U.S. population (National Center for Health Statistics 1994). Our research is motivated by nonresponse of body mass index (BMI) in the NHANES III. The data for our illustration come from this survey, and were collected from October 1988 to September 1994. In section 2.1 we describe the actual data, and in section 2.2 we describe the data we analyze.

2.1 NHANES III Data

The NHANES III consists of two parts. The first part is the interview of the sampled individuals for their personal information and the second part is the examination of those sampled. One or more persons from the sampled households were placed into a number of subgroups depending on their age, race and sex. Some subgroups were sampled at different rates. Sampled persons were asked to come to a mobile examination center (MEC) for a phyzsical examination, Those who did not come were visited by the examiner for the same purpose. Details of the NHANES III sample design are available (National Center for Health Statistics 1992). We incorporate design features associated with clustering in our model.

The main reasons for NHANES III nonresponse are "not interested", "no time/work conflict", "concerns/suspicious", "don't bother me" and "health reasons". The nonresponse rate of younger individuals is very high because the parents, especially older mothers of an only child, were extremely protective of their babies, and would not allow them to leave their homes for the MECs. Field workers often observe that obese persons tend to avoid the medical examination. So that nonresponse might be nonrandom and hence require some special attention.

NHANES III data are adjusted by multistage ratio weightings for the data to be consistent with the population (Mohadjer, Bell and Waksberg 1994). The ratio is the proportion of persons in the sample to the number of persons who completed interview and examination. Weighting with nonresponse ratio is one of these stages. In nonresponse ratio estimation, the proportions of nonrespondents in the multinomial cells are the same as those for the respondents (*i.e.*, ignorable nonresponse). In this case since the proportions are of interest, no adjustment is required. Clearly, this ratio estimation can be incorrect when these two groups are different. Therefore there is a need to consider the adjustment by a method other than ratio adjustment. In this paper we investigate a Bayesian method as an alternative to ratio weighting for nonignorable nomesponse.

NHANES III nonresponse also occurs at several levels in the survey: interview and examination. The interview nonresponse arises from sample individuals who did not respond for the interview. Some of those who were already interviewed did not come to the MEC, missing all or part of the examinations. In this paper, our population consists of those individuals who would have agreed to take the physical examination in the MECs. Thus, nonrespondents are those individuals who agreed to take the physical examination, and did not show up at the MECs. More specifically, since we are considering item response, the nonrespondents are those individuals who agreed to come to the MECs and their heights and/or weights were not measured.

Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin (1996) attempted a comprehensive multiple imputation project on the NHANES III data for many variables. The purpose was to impute the nonresponse data to provide several data sets for public use. Unfortunately, one of the limitations of the project was that "the procedure used to create missingness corresponds to a purely ignorable mechanism; the simulation provides no information on the impact of possible deviations from ignorable nonresponse." Another limitation is that the procedure did not include geographical clustering. Our purpose is different; we do not provide imputed public-use data.

2.2 Data Used for Illustration

Our data have two age groups (younger than 45 years, 45-, and 45 years or older, 45+), two race groups (white and non-white) and, of course, two groups for sex (male and female). Thus, there are eight age-race-sex domains.

One of the variables of interest in the NHANES III is BMI, an index of weight adjusted for height (Kg / m^2), that broadly categorizes obesity within age-race-sex groups (Kuczmarski, Carrol, Flegal and Troiano 1997) as low body fat (level 1: BMI < 20), healthy body fat (level 2: 20 ≤ BMI < 25), hefty or unhealthy (level 3: BMI ≥ 25). We use this broad classification for each of the eight age-race-sex groups.

Rather than a categorical data analysis, one can also provide an analysis that treats BMI as a continuous variable. While some information is lost by discretizing the BMI values, an analysis using continuous models for BMI will also be approximate and there is a need to search for an appropriate transformation. In the final analysis, a doctor only needs to know what proportions of the public belong to different levels of BMI, so he or she can tell his patient's standing in obesity.

The analysis of BMI data using categorical data methods is not uncommon. For example, Malec, Davis and Cao (1999) described a Bayes empirical Bayes analysis of the NHANES III data. They classified an individual older than 20 years as normal if her/his BMI is below a certain gender specific threshold. This is an application of a Bayesian analysis of binary data. However, their classification is somewhat restricted (see Kuczmarski *et al.* 1997). By considering multinomial data, we have generalized the analysis of Malec *et al.* (1999). In fact, they did not provide a nonignorable nonresponse model.

Unlike Schafer *et al.* (1996), we include clustering at the county level, although there is a need to include clustering at the household level. For the complete data there are 6,440 households. Of these households 52.1% contributed one person to the sample, 22.5% two persons, and 21.4% at least three persons. We have calculated the correlation coefficient for the BMI values based on pairing the members within households (see Rao 1973 page 199). It is 0.19 which indicates that as a first approximation the clustering within households can be ignored.

Table 1 shows the number of respondents for each BMI level for each age-race-sex domain and 34 counties (population at least 500,000). The pattern of respondents differs greatly by age. The nonresponse rate for the older group (45+) is negligible. Therefore the main concern about nonresponse must be given to the younger group (45-). There is also higher response rate among females than males. We note that the selection procedure is not random over the single population of males and females.

Table 1
Number of individuals in each BMI level and number of nonrespondents (Non) by age, race and sex over all 34 counties

				BMI		
Age	Race	Sex	1	2	3	Non
45-	W	M	1,098	651	597	558
		F	845	434	380	233
	B	M	1,198	713	665	574
		F	745	463	524	214
45+	W	M	46	439	1,014	3
		F	51	223	365	4
	B	M	79	470	942	8
		F	48	169	552	6

Note: BMI (1=less than 20; 2 = at least 20 and smaller than 25;
3 = greater than 25)
Age (Younger than 45 years = 45-; 45 years or older = 45+)
Race (White = W; all others = B)
Sex (Male = M; Female = F)

Table 2
Number of individuals in each BMI level and number of nonrespondents (Non) for eight examples (Ex) of small age-race-sex domains from different counties

					BMI Level		
Ex	Age	Race	Sex	1	2	3	Non
1	45-	W	M	1	3	1	14
2			F	3	4	1	0
3		B	M	5	5	6	10
4			F	3	1	1	1
5	45+	W	M	1	2	6	0
6			F	1	3	4	0
7		B	M	3	3	5	0
8			F	2	0	1	1

Note: BMI (1=less than 20; 2 = at least 20 and smaller than 25;
3 = greater than 25)
Age (Younger than 45 years = 45-; 45 years or older = 45+)
Race (White = W; all others = B)
Sex (Male = M; Female = F)

One important aspect of our work is on small area estimation. Because we consider inference for each age- race-sex domain separately over the the geographical areas (counties), the samples from some of these areas can be very small. Thus, small area estimation techniques are required to estimate the parameters corresponding to these smaller areas. Specifically, we need to "borrow strength" from the larger areas to make the estimates for the smaller areas more reliable. Table 2 presents eight examples to show the need for small area techniques. We have selected eight counties that have small domains; all the cell counts are at most 6 and many of them are as small as 1 (one of them is 0 for 45+). We will present overall estimates and the estimates for the first four examples (45-). Note that in comparison to the cell counts, the nonrespondents are large for two of them (14 and 10 nonrespondents).

We note that the purpose is not a comprehensive analysis of the NHANES III data although it forms an approximate analysis for these data. Our method is general enough to analyze multinomial nonresponse data from many areas, some of which can be small. It is for these small areas that we develop this modeling technique. Thus, in this paper we use the NHANES III data to illustrate our method.

Our method considers each domain separately with a "borrowing of strength" across the 34 areas (counties) to analyze the BMI data. Thus, there are eight separate analyses, each with 34 areas, and some of them are small. We use a hierarchical multinomial nonresponse model to analyze data of this form. The small cell counts, substantial nonrespondents and multinomial data make the methodology much more practical. Our methodology is also extended to incorporate all the domains simultaneously through logistic models.

3. METHODOLOGY FOR HIERARCHICAL MULTINOMIAL MODEL

We propose a model for each of the eight age-race-sex domains but for all counties taken simultaneously. However, the models fall into two broad classes. We will use a nonignorable nonresponse model for the younger group and an ignorable nonresponse model for the older group since the nonresponse rate for the older group is negligible. Of course, it is worthwhile to compare the ignorable nonresponse model and the nonignorable nonresponse model for the younger group. We will show how to combine the groups later using logistic regression, although this is not the key issue of this paper.

For each age-race-sex group, the k^{th} individual in the i^{th} county belongs to one of J BMI levels. Then for the k^{th} individual in i^{th} county, the characteristic variable at the j^{th} BMI level is defined as follows,

$$\mathbf{x}_{ik} = (x_{i1k}, \ldots, x_{ijk}, \ldots, x_{iJk})', \, i = 1, \ldots, c, \, k = 1, \ldots, n_i,$$

where each $x_{ijk} = 0$ or 1, $j = 1, \ldots, J$, and $\sum_{j=1}^{J} x_{ijk} = 1$. The response variable, y_{ijk}, is defined for each age-race-sex domain

$$y_{ijk} = \begin{cases} 1, & \text{if individual } k \text{ belonging to BMI level } j \text{ in county } i \text{ responded} \\ 0, & \text{if individual } k \text{ belonging to BMI level } j \text{ in county } i \text{ did not respond.} \end{cases}$$

We use a probabilistic structure to model the $\mathbf{x}_{ik}$ and y_{ijk}. In our application, there are $c = 34$ counties and $J = 3$ BMI levels.

3.1 Ignorable and Nonignorable Nonresponse Models

For both ignorable and the nonignorable nonresponse models, we have

$$\mathbf{x}_{ik} \mid \mathbf{p}_i \overset{\text{iid}}{\sim} \text{Multinomial } (1, \mathbf{p}_i) \qquad (1)$$

where p_{ij} is the probability that an individual in the i^{th} county belongs the j^{th} BMI level. Next, we describe the remaining portions of the ignorable and the nonignorable models.

First, we describe the ignorable nonresponse model. Let π_i denote the probability that an individual within the i^{th} county responds (*i.e.*, the probability of responding depends only on the county). Then, we assume that

$$y_{ijk} \mid \pi_i \overset{\text{iid}}{\sim} \text{Bernoulli } (\pi_i). \qquad (2)$$

At the second stage, letting $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1J})'$, we take

$$\mathbf{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet } (\boldsymbol{\mu}_1 \tau_1), \qquad (3)$$

$$\pi_i \mid \mu_{21}, \tau_{21} \overset{\text{iid}}{\sim} \text{Beta } (\mu_{21} \tau_{21}, (1 - \mu_{21}) \tau_{21}) \qquad (4)$$

where

$$p(\mathbf{p}_i \mid \boldsymbol{\mu}_1, \tau_1) \prod_{j=1}^{J} p_{ij}^{\mu_{1j}\tau_1 - 1} / D(\boldsymbol{\mu}_1 \tau_1), \; 0 < p_{ij} < 1, \sum_{j=1}^{J} p_{ij} = 1$$

and

$$D(\boldsymbol{\mu}_1 \tau_1) = \prod_{j=1}^{J} \Gamma(\mu_{1j} \tau_1) \Big/ \Gamma(\tau_1), 0 < \mu_{1j} < 1, \sum_{j=1}^{J} \mu_{1j} = 1.$$

The components of $\boldsymbol{\mu}_1$ are the prior means of the corresponding components of the $\mathbf{p}_i$, and τ_1 can be interpreted as a prior sample size. Similar interpretations can be given for μ_{21} and τ_{21} for π_i. Thus, assumption (3) expresses similarity among the cell proportions $\mathbf{p}_i$ and (4) expresses similarity among the response probabilities π_i. It is this structure that causes the "borrowing of strength" across the c counties.

Second, we describe the nonignorable nonresponse model. Let π_{ij} denote the probability that an individual within the i^{th} county responds in the j^{th} BMI level (*i.e.*, the probability of responding depends not only on the county but also on the BMI level). Then, we assume that

$$y_{ijk} \mid \{\mathbf{x}_{ik} = (x_{i1k}, \ldots, x_{iJk}), \pi_{ij}\} \overset{\text{iid}}{\sim} \text{Bernoulli } (\pi_{ij}) \quad (5)$$

where $x_{ijk} = 1$, $x_{ij'k} = 0$, $j \neq j'$ for $j, j' = 1, 2, \ldots, J$. Letting $\boldsymbol{\mu}_3 = (\mu_{31}, \mu_{32}, \ldots, \mu_{3J})'$, at the second stage we also take

$$\mathbf{p}_i \mid \boldsymbol{\mu}_3, \tau_3 \overset{\text{iid}}{\sim} \text{Dirichlet} (\boldsymbol{\mu}_3 \tau_3) \qquad (6)$$

and

$$\pi_{ij} \mid \mu_{4j}, \tau_{4j} \overset{\text{iid}}{\sim} \text{Beta} (\mu_{4j} \tau_{4j}, (1 - \mu_{4j}) \tau_{4j}), \, j = 1, \ldots, \text{J}. \; (7)$$

Like the assumptions in (3) and (4), the assumptions in (6) and (7) express similarity among the counties. We note that the response parameters π_{ij} are weakly identifiable (*i.e.*, unreliable estimates). However, the selection model works to our advantage, because the joint density of $\mathbf{x}_{ik}$ and $\mathbf{y}_{ik} = (y_{i1k}, ..., y_{iJk})'$ connects the p_{ij} and π_{ij}. In fact, this is an advantage over the pattern mixture approach.

To ensure a full Bayesian analysis, at the third stage we take the prior densities for the hyper-parameters as follows. For the ignorable nonresponse model, the prior densities are

$$\boldsymbol{\mu}_1 \sim \text{Dirichlet } (1, 1, ..., 1), \mu_{21} \sim \text{Beta } (1, 1),$$
$$\tau_1 \sim \text{Gamma } (\eta_1^{(0)}, \nu_1^{(0)}) \text{ and } \tau_{21} \sim \text{Gamma } (\eta_{21}^{(0)}, \nu_{21}^{(0)}),$$

where (letting t denote either τ_1 or τ_{21}, a either $\eta_1^{(0)}$ or $\eta_{21}^{(0)}$, and b either $\nu_1^{(0)}$ or $\nu_{21}^{(0)}$) $\tau \sim \text{Gamma } (a, b)$ means that $f(t) = b^a t^{a-1} e^{-bt} / \Gamma(a)$, $t > 0$ and $f(t) = 0$ otherwise. The hyper-parameters $\eta_1^{(0)}, \nu_1^{(0)}, \eta_{21}^{(0)}$ and $\nu_{21}^{(0)}$ are to be specified. The corresponding part of the nonignorable nonresponse model is

$$\boldsymbol{\mu}_3 \sim \text{Dirichlet } (1, 1, ..., 1), \mu_{4j} \sim \text{Beta } (1, 1),$$
$$\tau_3 \sim \text{Gamma } (\eta_3^{(0)}, \nu_3^{(0)}) \text{ and}$$
$$\tau_{4j} \sim \text{Gamma } (\eta_{4j}^{(0)}, \nu_{4j}^{(0)}), \ j = 1, ..., J.$$

Again, the hyper-parameters $\eta_3^{(0)}, \nu_3^{(0)}, \eta_{4j}^{(0)}, \nu_{4j}^{(0)}$, $j = 1, ..., J$, are to specified. It is possible to use other prior densities such as shrinkage priors, but it is likely that these will provide similar inference as our sensitivity analysis indicates in section 4.

It is an attractive property of the hierarchical model that it introduces correlation among the variables. For example, in our application (1), (2), (3) and (4) make the $(\mathbf{x}_{ij}, y_{ij})$ equi-correlated across the individuals within the i^{th} area. This is the clustering effect within the areas. Such an effect can be obtained directly, but it will not be as simple as in a hierarchical model. A further benefit of the hierarchical model is that it takes care of extraneous variations among the areas, and this effect can be obtained directly by using random effects model. But in our case, this will loose the natural multinomial data structure.

Let r_i be the number of respondents in county i and y_{ij} the number of respondents having the j^{th} BMI level in the i^{th} county. Then r_i and y_{ij} are random variables; $n_i - r_i$ is the number of nonrespondents. Since the number of nonrespondents at the j^{th} BMI level is unknown, we denote them by the latent variables z_{ij} (see the tree diagram in Figure 1). If we can tell what the z_{ij} are, our nonresponse problem will be solved. Of course, under the assumption of ignorable nonresponse, they can be estimated easily using ratio estimation. The z_{ij} are useful because under the assumption of nonignorable nonresponse they simplify the sampling based method to obtain estimates of the parameters of interest.

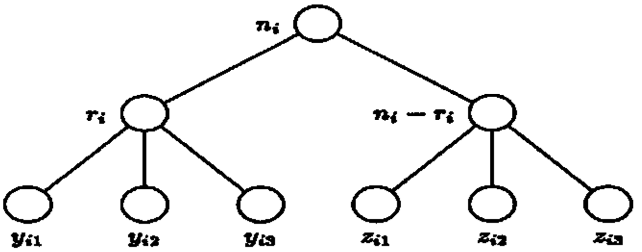

Figure 1. Latent nonignorable response tree diagram. From a sample of n_i individuals, there are r_i respondents of which y_{ij} belong to category j, $j = 1, 2, 3$. Among the $(n_i - r_i)$ nonrespondents z_{ij} individuals belong to category j, where z_{ij} are latent variables.

The likelihood function for the ignorable nonresponse model is

$$f(\mathbf{y}, \mathbf{r} \mid \mathbf{p}, \boldsymbol{\pi}) = \prod_{i=1}^{c} \left\{ \binom{n_i}{r_i} \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i} \right\}$$
$$\times \prod_{i=1}^{c} \left\{ \binom{r_i}{y_{i1}, ..., y_{iJ}} \prod_{j=1}^{J} \left\{ p_{ij}^{y_{ij} + n_i - r_i} \right\} \right\}.$$

Here the likelihood function has two distinct parts, one for p_{ij} and the other for the π_i. Using Bayes' theorem the joint posterior density of all the parameters is

$$f(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \mu_{21}, \tau_{21} \mid \mathbf{y}, \mathbf{r})$$
$$\propto \prod_{i=1}^{c} \left\{ \left\{ \prod_{j=1}^{J} p_{ij}^{y_{ij} + n_i - r_i} \right\} \left\{ \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i} \right\} \right.$$
$$\times \left\{ \prod_{j=1}^{J} p_{ij}^{\mu_{1j}\tau_1 - 1} \right\} \Big/ D(\boldsymbol{\mu}_1 \tau_1)$$
$$\left. \times \left\{ \frac{\pi_i^{\mu_{21}\tau_{21} - 1} (1 - \pi_i)^{(1 - \mu_{21})\tau_{21} - 1}}{B(\mu_{21}\tau_{21}, (1 - \mu_{21})\tau_{21})} \right\} \right\}$$
$$\times \left\{ \tau_1^{\eta_1^{(0)} - 1} \exp\left(-\nu_1^{(0)} \tau_1\right) \right\} \left\{ \tau_{21}^{\eta_{21}^{(0)} - 1} \exp\left(-\nu_{21}^{(0)} \tau_{21}\right) \right\}. \tag{8}$$

Similarly, the augmented likelihood function (*i.e.*, including the $\mathbf{z}_i$) for the nonignorable nonresponse model is

$$f(\mathbf{y}, \mathbf{r}, \mathbf{z} \mid \mathbf{p}, \boldsymbol{\pi}) = \prod_{i=1}^{c} \left\{ \binom{n_i}{r_i} \binom{r_i}{y_{i1}, ..., y_{iJ}} \binom{n_i - r_i}{z_{i1}, ..., z_{iJ}} \right.$$
$$\left. \times \prod_{j=1}^{J} \left\{ (\pi_{ij} p_{ij})^{y_{ij}} ((1 - \pi_{ij}) p_{ij})^{z_{ij}} \right\} \right\}$$

and using Bayes' theorem the joint posterior density of all the parameters is

$$f(\mathbf{p}, \boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\mu}_3, \tau_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4 \mid \mathbf{y}, \mathbf{r})$$

$$\propto \prod_{i=1}^{c} \left\{ \binom{n_i - r_i}{z_{i1}, \cdots, z_{iJ}} \prod_{j=1}^{J} (\pi_{ij} p_{ij})^{y_{ij}} \left((1-\pi_{ij}) p_{ij}\right)^{z_{ij}} \right.$$

$$\times \prod_{j=1}^{J} p_{ij}^{\mu_{3j}\tau_3 - 1} \Big/ D(\boldsymbol{\mu}_3 \tau_3) \prod_{j=1}^{J}$$

$$\left. \times \left\{ \frac{\pi_{ij}^{\mu_{4j}\tau_{4j}-1}(1-\pi_{ij})^{(1-\mu_{4j})\tau_{4j}-1}}{B(\mu_{4j}\tau_{4j}, (1-\mu_{4j})\tau_{4j})} \right\} \right\}$$

$$\times \left\{ \tau_3^{\eta_2^{(0)}-1} \exp\left(-\nu_2^{(0)}\tau_3\right) \right\} \prod_{j=1}^{J} \left\{ \tau_{4j}^{\eta_{4j}^{(0)}-1} \exp\left(-\nu_{4j}^{(0)}\tau_{4j}\right) \right\}.$$

We consider inference about the p_{ij}, the proportion of individuals at the j^{th} BMI level in the i^{th} county, and the probability of responding,

$$\delta_i = \sum_{j=1}^{J} \pi_{ij} p_{ij}, \; i = 1, \ldots, c.$$

However, the joint posterior densities in (8) and (9) are complex, and can not be used to make inference analytically. Thus, we use a Markov chain Monte Carlo algorithm to obtain estimates of the posterior distribution of the parameters. Our method is to use a Metropolis-Hastings (MH) sampler to get samples from (8) and (9) and then to use these samples to make posterior inferences about $\mathbf{p}_i$ and δ_i.

3.2 Computations

For the ignorable nonresponse model, it is convenient to represent the posterior density function as

$$f(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \mu_{21}, \tau_{21} \mid \mathbf{y}, \mathbf{r})$$

$$= \prod_{i=1}^{c} \{ f_1(\mathbf{p}_i \mid \mathbf{y}, \mathbf{r}, \boldsymbol{\mu}_1, \tau_1) f_2(\boldsymbol{\pi}_i \mid \mathbf{y}, \mathbf{r}, \mu_{21}, \tau_{21}) \}$$

$$\times f_3(\boldsymbol{\mu}_1, \tau_1, \mu_{21}, \tau_{21} \mid \mathbf{y}, \mathbf{r})$$

where $f_1(\cdot)$ is Dirichlet density,

$$\mathbf{p}_i \mid \mathbf{y}_i, \mathbf{r}_i, \boldsymbol{\mu}_1, \tau_1 \overset{\textbf{ind}}{\sim} D(\mathbf{y}_i + n_i - r_i + \boldsymbol{\mu}_1 \tau_1),$$

$f_2(\cdot)$ is beta density,

$$\pi_i \mid \mathbf{y}_i, \mathbf{r}_i, \mu_{21}, \tau_{21} \overset{\textbf{ind}}{\sim} \text{Beta}(r_i + \mu_{21}\tau_{21}, n_i - r_i + (1-\mu_{21})\tau_{21})$$

and

$$f_3(\boldsymbol{\mu}_1, \tau_1, \mu_{21}, \tau_{21} \mid \mathbf{y}, \mathbf{r})$$

$$\propto \prod_{i=1}^{c} \left\{ D(\mathbf{y}_i + n_i - r_i + \boldsymbol{\mu}_1 \tau_1) / D(\boldsymbol{\mu}_1 \tau_1) \right\} p(\boldsymbol{\mu}_1, \tau_1)$$

$$\times \prod_{i=1}^{c} \left\{ \frac{B(r_i + \mu_{21}\tau_{21}, n_i - r_i + (1-\mu_{21})\tau_{21})}{B(\mu_{21}\tau_{21}, (1-\mu_{21})\tau_{21})} \right\} p(\mu_{21}, \tau_{21})$$

with $p(\boldsymbol{\mu}_1, \tau_1)$ and $p(\mu_{21}, \tau_{21})$ the prior distributions. Hence, f_1 and f_2 are obtained through the Gibbs kernel, while for f_3 we use the MH algorithm (Nandram 1998).

For the nonignorable nonresponse model, it is convenient to represent the posterior density function as

$$f(\mathbf{p}, \boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\mu}_3, \tau_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4 \mid \mathbf{y}, \mathbf{r})$$

$$= \prod_{i=1}^{c} \left\{ \left\{ \prod_{j=1}^{J} f_j(\pi_{ij} \mid \mathbf{y}, \mathbf{r}, \mathbf{z}, \mu_{4j}, \tau_{4j}) \right\} f_{J+1}(\mathbf{p}_i \mid \mathbf{y}, \mathbf{r}, \mathbf{z}, \boldsymbol{\mu}_3, \tau_3) \right\}$$

$$\times f_{J+2}(\boldsymbol{\mu}_3, \tau_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4, \mathbf{z} \mid \mathbf{y}, \mathbf{r}),$$

where $f_1(\cdot), \ldots, f_J(\cdot)$ are beta densities,

$$\pi_{ij} \mid y_{ij}, r_{ij}, z_{ij}, \mu_{4j}, \tau_{4j} \overset{\textbf{ind}}{\sim} \text{Beta}(y_{ij} + \mu_{4j}\tau_{4j}, z_{ij} + (1-\mu_{4j})\tau_{4j}),$$

$f_{J+1}(\cdot)$ is a Dirichlet density,

$$\mathbf{p}_i \mid \mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\mu}_3, \tau_3 \overset{\textbf{ind}}{\sim} D(\mathbf{y}_i + \mathbf{z}_i + \boldsymbol{\mu}_3 \tau_3)$$

and $f_{J+2}(\cdot)$ is given by

$$f_{J+2}(\boldsymbol{\mu}_3, \tau_3, \mu_4, \tau_4, \mathbf{z} \mid \mathbf{y}, \mathbf{r})$$

$$\propto \prod_{i=1}^{c} \left\{ \binom{n_i - r_i}{z_{i1}, \ldots, z_{iJ}} \left\{ D(\mathbf{y}_i + \mathbf{z}_i + \boldsymbol{\mu}_3 \tau_3) / D(\boldsymbol{\mu}_3 \tau_3) \right\} p(\boldsymbol{\mu}_3, \tau_3) \right.$$

$$\left. \times \prod_{j=1}^{J} \left\{ \frac{B(y_{ij} + \mu_{4j}\tau_{4j}, z_{ij} + (1-\mu_{4j})\tau_{4j})}{B(\mu_{4j}\tau_{4j}, (1-\mu_{4j})\tau_{4j})} \right\} \right\} p(\boldsymbol{\mu}_4, \boldsymbol{\tau}_4)$$

with $p(\boldsymbol{\mu}_3, \tau_3)$ and $p(\boldsymbol{\mu}_4, \tau_4)$ the prior distributions. Thus, $f_1, \ldots, f_{J+1}$ are obtained through the Gibbs kernel, while f_{J+2} is obtained using the MH algorithm (Nandram 1998). We obtain the latent variables z_{ij} through one of the conditional posterior densities of the MH algorithm. A sketch of the procedure is given in Appendix 1.

We drew 5,500 iterates, threw out the first 500, and took every fifth (obtained by trace plots). This strategy was satisfactory to wash out the autocorrelation among the iterates and to have good jumping probabilities (0.25-0.50) for the Metropolis steps. For the computation, first we set the hyper-parameters $\eta_1^{(0)}, \nu_1^{(0)}, \eta_{21}^{(0)}, \nu_{21}^{(0)}, \eta_3^{(0)}, \nu_3^{(0)}, \eta_{4j}^{(0)}, \nu_{4j}^{(0)}$, $j = 1, ..., J$ equal to 0. Then we ran our MH algorithm to obtain posterior samples of $\tau_1, \tau_{21}, \tau_3$ and τ_{4j}, $j = 1, ..., J$. To ensure proper posterior densities, we estimate $\eta_1^{(0)}, \nu_1^{(0)}, \eta_{21}^{(0)}, \nu_{21}^{(0)}, \eta_3^{(0)}, \nu_3^{(0)}, \eta_{4j}^{(0)}, \nu_{4j}^{(0)}$, $j = 1, ..., J$, by fitting the gamma priors on the posterior samples for $\tau_1, \tau_{21}, \tau_3$ and τ_{4j}, $j = 1, ..., J$. These values are shown in Table 3. Finally, with these proper priors we ran our algorithm to obtain posterior samples. Specifically, we obtained M = 1,000 iterates $(\mathbf{p}_i^{(h)}, \delta_i^{(h)})$, $h = 1, ..., M$, $i = 1, ..., c$. Inference about the $\mathbf{p}_i$, δ_i and any function of them can be made using these iterates in a straightforward manner.

Table 3
Estimates of $\eta^{(0)}$ and $\nu^{(0)}$ corresponding to the gamma densities on τ_1, τ_{21} for 45+ and $\tau_3, \tau_{41}, \tau_{42}, \tau_{43}$ for 45- by race and sex

			Age					
			45-				45+	
Race	Sex		τ_3	τ_{41}	τ_{42}	τ_{43}	τ_1	τ_{21}
W	M	$\eta^{(0)}$	3.698	2.341	3.085	2.685	4.408	3.941
		$\nu^{(0)}$	.036	.071	.201	.163	.009	.052
	F	$\eta^{(0)}$	4.200	3.294	2.481	1.819	4.788	4.384
		$\nu^{(0)}$	.030	.059	.072	.017	.008	.019
B	M	$\eta^{(0)}$	4.948	2.922	3.156	2.404	5.971	4.376
		$\nu^{(0)}$	.068	.096	.169	.147	.107	.036
	F	$\eta^{(0)}$	3.745	3.084	1.893	2.350	3.292	4.488
		$\nu^{(0)}$	.055	.036	.049	.116	.009	.036

4. AN ANALYSIS OF THE NHANES III DATA

In this section we illustrate our methodology using the BMI data from NHANES III. First, we study our estimates based on summary measures over the counties. Specifically, we use the weighted posterior distributions of the p_{ij},

$$\tilde{p}_j = \sum_{i=1}^{c} n_i p_{ij} \Big/ \sum_{i=1}^{c} n_i, j = 1, 2, 3$$

and the weighted posterior distribution of the δ_i

$$\tilde{\delta} = \sum_{i=1}^{c} n_i \delta_i \Big/ \sum_{i=1}^{c} n_i$$

for each of the eight age-race-sex domains. Then, for the first four examples in Table 2 we show small area effects.

We also show how to relate the p_{ijk} and the π_{ij} to age, race and sex using linear and nonlinear logistic regression models

4.1 Data Analysis

First, we performed a sensitivity analysis to assess the specifications of $\eta^{(0)}$ and $\nu^{(0)}$. We compared three choices of hyper-parameters $\Omega = (\eta^{(0)}, \nu^{(0)})$ to check the sensitivity of the specification of the hyper-parameters on inference. Our first choice is 4 times of Ω, *i.e.*, $4\Omega = (4\eta^{(0)}, 4\nu^{(0)})$; our second choice is the hyper-parameters without any change, *i.e.*, $\Omega = (\eta^{(0)}, \nu^{(0)})$; and our third choice is one fourth of Ω *i.e.*, $\Omega/4 = (\eta^{(0)}/4, \nu^{(0)}/4)$.

Table 4 shows the simulation results for the sensitivity to the inference of $\tilde{p}_j$ for the younger group (45-). The point estimates and standard deviations of the proportions are very similar over the three choices of hyper-parameters. Similarly, Table 5 shows the simulation results for $\tilde{p}_j$ for the older group (45+). The point estimates for males are very similar over the three choices of the hyper-parameters, but there are small changes in the point estimates for females from 4Ω to Ω. The standard deviations are increased when Ω decreases for the females, but no substantial changes are detected for males. Generally, the nonignorable nonresponse model performs better than the ignorable nonresponse model, as the nonignorable nonresponse model is not sensitive to choices of the hyper-parameters.

Table 4
Sensitivity of $\tilde{p}_j$ for choice of $\eta_3^{(0)}, \nu_3^{(0)}, \eta_{4j}^{(0)}$ and $\nu_{4j}^{(0)}$, $j = 1, ..., 4$ for the younger group (45-) for the three BMI levels

Race	Sex	$\tilde{p}_1$	std($\tilde{p}_1$)	$\tilde{p}_2$	std($\tilde{p}_2$)	$\tilde{p}_3$	std($\tilde{p}_3$)
(a) 4Ω							
W	M	.428	.022	.216	.019	.356	.022
	F	.476	.025	.232	.020	.292	.024
B	M	.419	.020	.212	.016	.369	.020
	F	.434	.026	.185	.023	.381	.027
(b) Ω							
W	M	.427	.022	.211	.020	.362	.025
	F	.476	.026	.223	.024	.301	.031
B	M	.419	.020	.208	.017	.373	.022
	F	.435	.025	.178	.026	.387	.029
(c) $\Omega/4$							
W	M	.427	.022	.210	.021	.364	.027
	F	.475	.026	.220	.026	.304	.034
B	M	.419	.020	.206	.018	.375	.024
	F	.435	.025	.177	.028	.388	.029

Note 1: $\Omega = (\eta_3^{(0)}, \nu_3^{(0)}, \eta_{41}^{(0)}, \nu_{41}^{(0)}, \eta_{42}^{(0)}, \nu_{42}^{(0)}, \eta_{43}^{(0)}, \nu_{43}^{(0)})$.

Note 2: The nonignorable nonresponse model is applied to the younger group.

Table 5
Sensitivity of $\tilde{p}_j$ for choice of $\eta_1^{(0)}, \nu_1^{(0)}, \eta_{21}^{(0)}, \nu_{21}^{(0)}$ for the older group (45+) for the three BMI levels

Race	Sex	$\tilde{p}_1$	std($\tilde{p}_1$)	$\tilde{p}_2$	std($\tilde{p}_2$)	$\tilde{p}_3$	std($\tilde{p}_3$)
(a) 4Ω							
W	M	.030	.005	.306	.018	.664	.018
	F	.081	.002	.436	.004	.483	.004
B	M	.053	.011	.317	.017	.630	.018
	F	.075	.005	.201	.004	.724	.006
(b) Ω							
W	M	.031	.005	.292	.016	.677	.016
	F	.063	.002	.443	.006	.494	.005
B	M	.053	.011	.316	.019	.631	.020
	F	.066	.012	.237	.018	.697	.019
(c) $\Omega/4$							
W	M	.031	.005	.293	.018	.676	.019
	F	.073	.015	.359	.011	.568	.019
B	M	.053	.010	.317	.018	.630	.019
	F	.065	.013	.221	.022	.714	.025

Note 1: $\Omega = (\eta_1^{(0)}, \nu_1^{(0)}, \eta_{21}^{(0)}, \nu_{21}^{(0)})$.

Note 2: The ignorable nonresponse model is applied to the older group.

Table 6
Point estimates and 95% credible intervals for the weighted probability of response, $\tilde{\delta} = \sum_{i=1}^{c} n_i \delta_i / \sum_{i=1}^{c} n_i$, for three choices of Ω and the younger group

		4Ω			Ω			$\Omega/4$		
Race	Sex	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval
W	M	.775	.016	(.744, .805)	.769	.017	(.735, .801)	.767	.018	(.732, .799)
	F	.855	.017	(.821, .886)	.855	.020	(.810, .887)	.853	.022	(.806, .887)
B	M	.786	.016	(.752, .817)	.780	.018	(.740, .813)	.778	.018	(.739, .811)
	F	.880	.013	(.854, .902)	.878	.015	(.845, .903)	.876	.015	(.838, .903)

Note: See the note to Table 1.

Table 6 shows point estimates of the probability of responding $\tilde{\delta}$, and their 95% credible intervals for three choices of Ω. The probabilities of responding for males are lower than those for females, and this pattern remains the same for three choices of Ω. If a similar survey is conducted in the future, we should increase the sample size by 1.30 = (1/.769) times for white males and 1.17 = (1/.855) times for white females (*e.g.*, if complete data are required from 1,000 households, the interviewer needs to contact 1,300 white males).

In Table 7 we present 95% credible intervals for the $\tilde{p}_j$ for the three BMI levels. For the younger group, $\tilde{p}_1$ of BMI level 1 is the highest, and $\tilde{p}_2$ of BMI level 2 is the lowest. The lower bounds for $\tilde{p}_1$ and $\tilde{p}_3$ are similar for the younger group except for white females, and those for $\tilde{p}_2$ are similar except for the non-white females. For the older group, $\tilde{p}_3$ of BMI level 3 is highest, and $\tilde{p}_1$ of BMI level 1 is lowest. Specifically $\tilde{p}_1$, $\tilde{p}_2$ are high and $\tilde{p}_3$ is low for the white males.

Table 7
95% credible intervals for the weighted proportions, $\tilde{p}_j = \sum_{i=1}^{c} n_i p_{ij} / \sum_{i=1}^{c} n_i$ by age, race and sex

			95% credible interval		
Age	Race	Sex	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$
45-	W	M	(.382, .470)	(.174, .252)	(.314, .412)
		F	(.425, .525)	(.171, .269)	(.243, .371)
	B	M	(.381, .455)	(.176, .241)	(.333, .419)
		F	(.385, .482)	(.130, .230)	(.329, .442)
45+	W	M	(.022, .041)	(.255, .326)	(.643, .710)
		F	(.059, .068)	(.431, .451)	(.486, .505)
	B	M	(.035, .076)	(.282, .352)	(.592, .670)
		F	(.040, .093)	(.206, .265)	(.661, .731)

Note 1: The nonignorable nonresponse model is applied to the younger group.

Note 2: The ignorable nonresponse model is applied to the older group.

As suggested by a referee, we have looked at the results for older white females (45+) in Table 7 in greater detail. From Table 1 the observed proportions in the three BMI levels are .079, .347 and .568. However, the 95% credible intervals for the population proportions in Table 7 are (.059, .068), (.431, .451) and (.486, .505) respectively. That is, while the observed proportions are close to the intervals, none of these intervals contains the observed proportions. We can explain this phenomenon in the following manner. The data for older white females (45+) are very sparse. For the 34 counties the quartiles of the observed counts in the three BMI levels are (0,1,3), (3,6,10) and (5,9,14) respectively. Thus, when the ignorable nonresponse model is fit to the 34 counties, there is shrinkage not only across the counties but also across the BMI levels. Consequently, the largest proportion tends to be smaller and the smallest proportion tends to be larger, and since the three proportions must add up to one, the second proportion must also "shrink" somewhat. In addition, consider the sensitivity analysis in Table 5. We can approximate 95% credible intervals for $\tilde{p}_1$, $\tilde{p}_2$ and $\tilde{p}_3$, by using the posterior mean ± 2 × standard deviation. The intervals at 4Ω and Ω do not contain the observed proportions, but the intervals at $\Omega/4$ do. Therefore, because of the sparseness of tha data, there is some sensitivity to inference for older white females (45+) with respect to the prior misspecification of Ω. These results are expected within the small area context, when there are sparse data.

We use the first four examples in Table 2 to illustrate small area estimation. As it can be imagined, it is too cumbersome to present all the estimates for the 34 counties and the 8 domains. Table 8 shows the posterior means, standard deviations and 95% credible intervals for the p_{ij} and the δ_i.

First, we compare the estimates of the p_{ij} from the ignorable and nonignorable nonresponse models. The estimates from the two models are generally different with the intervals for the nonignorable nonresponse model wider than those for the ignorable nonresponse model.

Second, we consider the estimates (based on the nonignorable nonresponse model) of p_{ij} for the individual counties in Table 8 with the overall averages, the $\tilde{p}_j$ in Table 7. As expected, when the $\tilde{p}_j$ are obtained, there is an overall reduction in variability because of the extra smoothing, thereby making the intervals for the smaller domains relatively much wider. In fact, all the intervals for the small domains contain the intervals for $\tilde{p}_j$.

Finally, in Table 8 we consider the estimates of $\tilde{p}_{ij}$ for the individual counties with the overall average, $\tilde{p}_j$ in Table 7. The message is similar to that for the p_{ij}.

However, we note that the first example is an exception where the credible interval for $\delta_i(.459, .773)$ is almost completely to the left side of the credible interval for $\tilde{\delta}(.735, .801)$. Thus, there is much shrinkage for this example which is due to the relatively large number of nonrespondents, 14 in this county for white males 45-.

Table 8
Comparaison of the ignorable (ig) and the nonignorable (nig) nonresponse models for the four examples (Ex) corresponding to small domains using the cell probabilities (p_j) and the probability of responding (δ)

Ex	Model		p_1	p_2	p_3	δ
1	ig	avg	.444	.308	.248	
		std	.073	.067	.067	
		CI	(.297, .593)	(.193, .450)	(.125, .386)	
	nig	avg	.450	.276	.273	.637
		std	.093	.079	.082	.081
		CI	(.256, .638)	(.137, .444)	(.133, .448)	(.459, .773)
2	ig	avg	.480	.308	.213	
		std	.075	.066	.062	
		CI	(.324, .619)	(.193, .452)	(.097, .344)	
	nig	avg	.493	.263	.244	.879
		std	.074	.065	.062	.041
		CI	(.338, .628)	(.141, .406)	(.121, .394)	(.782, .948)
3	ig	avg	.420	.306	.274	
		std	.071	.063	.063	
		CI	(.276, .561)	(.192, .437)	(.161, .416)	
	nig	avg	.438	.252	.310	.741
		std	.079	.072	.074	.058
		CI	(.283, .591)	(.116, .406)	(.186, .483)	(.607, .836)
4	ig	avg	.448	.263	.288	
		std	.089	.075	.081	
		CI	(.278, .620)	(.127, .424)	(.138, .468)	
	nig	avg	.430	.261	.308	.874
		std	.100	.086	.091	.046
		CI	(.217, .619)	(.104, .453)	(.145, .517)	(.768, .948)

Note: For each parameter avg = posterior mean; std = posterior standard deviation; CI = 95% credible interval

4.2 Linear and Nonlinear Logistic Regression Models

Let q_{ijl} denote the probability that a respondent in $l^{\text{th}}(l = 1, 8)$ age-race-sex group in the i^{th} county belongs to the j^{th} BMI level. (We add the subscript l to the p_{ij} to denote the domains.) Letting $v_{ijl} = \log\{\sum_{\delta=1}^{j} q_{i\delta l} / (1 - \sum_{\delta=1}^{j} q_{i\delta l})\}$, $j = 1, ..., J-1$, we take

$$v_{ijl} = (\theta_j - (\mu_i + \alpha_l)) \Big/ \psi_i \tag{10}$$

subject to the constraints $\sum_{i=1}^{c} \mu_i = 0, \sum_{j=1}^{J-1} \theta_j = 0$, $\sum_{l=1}^{8} \alpha_l = 0$, and $\sum_{i=1}^{c} \ln \psi_i = 0$. The parameters $\theta_j, \mu_l, \alpha_l$ and ψ_i in (10) have posterior distributions whose properties are inherited from the posterior distributions of q_{ijl}. Each iterate of the MH algorithm provides a value for q_{ijl} which is used in (10), and a nonlinear least squares problem is solved using an iterative method to get the values of $\theta_j, \mu_i, \alpha_l$ and ψ_i (see Appendix 2). Alternatively, we can also use the much simpler linear logistic model in which the ψ_i in (10) are taken equal to unity. In this case, the least squares estimators of $\theta_j, \varphi_i, \mu_i$ and α_l exist in closed form at the h^{th} iteration of MH algorithm. Specifically, for $\varphi_i = 0$, we have the least squares estimates $\hat{\mu}_i = \bar{v}... - \bar{v}_{i..}$, $\hat{\theta}_j = \bar{v}_{.j.}$, $\hat{\alpha}_l = \bar{v}... - \bar{v}_{..l}$, where

$$\bar{v}... = \sum_{i=1}^{c} \sum_{j=1}^{J-1} \sum_{l}^{8} v_{ijl} / 8c\,(J-1),$$

$$\bar{v}_{i..} = \sum_{j=1}^{J=1} \sum_{l=1}^{8} v_{ijl} / 8(J-1),$$

$$\bar{v}_{.j.} = \sum_{i=1}^{c} \sum_{l=1}^{8} v_{ijl} / 8c$$

and $\bar{v}_{..l} = \sum_{i=1}^{c} \sum_{j=1}^{J-1} v_{ijl} / c\,(J-1)$. The nonlinear least squares problem is solved using an iterative method to get the values of $\hat{\theta}_j, \hat{\varphi}_i, \hat{\mu}_i$ and $\hat{\alpha}_l$.

We present 95% credible intervals for θ_1, θ_2 and $\alpha_1, ..., \alpha_8$ for the younger and older groups by regression type in Table 9. For the cut-points θ_j, θ_1 gives a large negative effect compared to θ_2. The relative measure $\alpha_l (l = 1, ..., 4)$ of the younger group gives a negative effect, while the relative measure $\alpha_l (l = 5, ..., 8)$ of the older group gives positive effects. The 95% credible intervals for linear and nonlinear estimates are essentially the same.

We also relate the probability of response, $\delta_i = \sum_{j=1}^{J} \pi_{ij} p_{ij}$, to race and sex using linear and nonlinear logistic regression models for the younger group. The 95% credible intervals for θ and $\alpha_1, ..., \alpha_4$ for the young group by regression type are shown in Table 10. Credible intervals for all α_l for the nonlinear model are shorter than those for the linear model. However, for the nonlinear model the credible interval for θ is wider than and on the right of that for the linear model.

Table 9
Comparaison of 95% credible intervals for θ_1, θ_2 and $\alpha_1, ..., \alpha_8$ for both younger and older groups by regression type

	Linear	Nonlinear
θ_1	(-1.743, -1.469)	(-1.731, -1.466)
θ_2	(0.028, 0.196)	(0.025, 0.193)
α_1	(-1.167, -0.751)	(-1.159, -0.751)
α_2	(-1.395, -0..939)	(-1.385, -0.937)
α_3	(-1.127, -0.723)	(-1.119, -0.728)
α_4	(-1.112, -0.659)	(-1.103, -0.658)
α_5	(1.198, 1.514)	(1.188, 1.498)
α_6	(0.513, 0.689)	(0.506, 0.685)
α_7	(0.715, 1.210)	(0.725, 1.225)
α_8	(0.809, 1.310)	(0.803, 1.300)

Table 10
Comparaison of 95% credible intervals for θ and $\alpha_1, ..., \alpha_4$ for the younger group by regression type

	Linear	Nonlinear
θ	(1.455, 1.729)	(1.664, 2.174)
α_1	(0.165, 0.592)	(0.146, 0.523)
α_2	(-0.535, 0.014)	(-0.467, 0.007)
α_3	(0.078, 0.546)	(0.079, 0.484)
α_4	(-0.704, -0.165)	(-0.638, -0.169)

5. A SIMULATION STUDY

We describe a small simulation study to assess the performance of our multinomial nonignorable nomesponse model. We focus on the probability of responding.

We use the observed data from younger white males to obtain the posterior means of p_{i1}, p_{i2}, p_{i3} and $\pi_{i1}, \pi_{i2}, \pi_{i3}$ for each county. These are taken to be the true (t) values which we denote by $p_{i1}^{(t)}, p_{i2}^{(t)}, p_{i3}^{(t)}$ and $\pi_{i1}^{(t)}, \pi_{i2}^{(t)}, \pi_{i3}^{(t)}$. Thus, the true probability of responding in the i^{th} county is $\delta_i^{(t)} = \sum_{j=1}^{3} p_{ij}^{(t)} \pi_{ij}^{(t)}$ and the weighted probability of responding is $\tilde{\delta}^{(t)} = \sum_{i=1}^{c} n_i \delta_i^{(t)} / \sum_{i=1}^{c} n_i$. In our simulated examples, we used the n_i as in the BMI data for younger white males, and we kept the $p_{ij}^{(t)}$ fixed throughout. However, we varied the π_{ij} in the following manner. We kept π_{i1} fixed at $\pi_{i1}^{(t)}$, and we denote the vector of the π_{i1} by π_1. The 34 values of the $\pi_{i1}^{(t)}$ range from .73 to .83. Then, we set $\pi_2 = \alpha \pi_1$ and $\pi_3 = b \pi_1$, where $a, b = 0.8, 0.9, 1.0$. (We denote the vectors of the π_{i2} and the π_{i3} by π_1 and π_2 respectively.) Thus, there are 9 simulated examples.

Then, for each (a, b) we generated counts for a multinomial probability mass function with probabilities $p_{i1}^{(t)} \pi_{i1}, p_{i2}^{(t)} \pi_{12}, p_{i3}^{(t)} \pi_{13}, p_{i1}^{(t)} (1 - \pi_{i1}), p_{i2}^{(t)} (1 - \pi_{12}), p_{i3}^{(t)} (1 - \pi_{13})$. We denote these cell counts by $y_{i1}, y_{i2}, y_{i3}, z_{i1}, z_{i2}, z_{i3}$ and the number of respondents is $r_i = \sum_{j=1}^{3} y_{ij}$. Then, we fit the nonignorable nonresponse model to the above data using the MH sampler, and we obtained M = 1,000 values $(p_{ij}^{(h)}, \pi_{ij}^{(h)}), h = 1, ..., M$. For each value, we computed $\tilde{\delta}^{(h)} = \sum_{i=1}^{c} n_i \delta_i^{(h)} / \sum_{i=1}^{c} n_i$ where $\delta_i^{(h)} = \sum_{j=1}^{3} p_{ij}^{(h)} \pi_{ij}^{(h)}$.

In Table 11 we report posterior means, standard deviations, numerical standard errors (using the batch means method) and 95% credible interval for the probability of responding for each choice of (a, b). We also computed $\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} \lfloor \mathbf{y}, \mathbf{r})$ by counting the number of $\tilde{\delta}^{(h)}$ that are as large as $\tilde{\delta}^{(t)}$. An extremely large or small value of this latter quantity suggests model failure.

We plotted the estimates of the posterior densities of $\tilde{\delta}$ by choices of a and b which we obtained by using normal kernel density estimator with an optimal window width from an output analysis of the MH algorithm. The densities are an unimodal, peaked and almost symmetric. By increasing (a, b) from (0.8,0.8) to (1.0, 1.0), the mode of the posterior densities increase.

Table 11
Characteristics of the probability of responding

		π_3		
π_2	stat	$0.8 * \pi_1$	$0.9 * \pi_1$	$1.0 * \pi_1$
$0.8 * \pi_1$	true	0.690	0.719	0.748
	avg	0.712	0.739	0.764
	std	0.016	0.015	0.014
	nse	0.0030	0.0031	0.0029
	CI	(0.678, 0.742)	(0.708, 0.767)	(0.734, 0.750)
	prob	0.082	0.095	0.135
$0.9 * \pi_1$	true	0.706	0.735	0.764
	avg	0.710	0.742	0.776
	std	0.017	0.016	0.014
	nse	0.0030	0.0031	0.0031
	CI	(0.673, .0.742)	(0.712, 0.769)	(0.745, 0.802)
	prob	0.377	0.303	0.210
$1.0 * \pi_1$	true	0.722	0.751	0.780
	avg	0.726	0.758	0.784
	std	0.017	0.015	0.015
	nse	0.0036	0.0036	0.0026
	CI	(0.693, 0.757)	(0.725, 0.784)	(0.750, 0.809)
	prob	0.399	0.318	0.380

Note: avg = posterior mean; std = standard deviation; nse = numerical standard error; CI = 95% credible interval; prob=$\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} \mid y, r)$; the 34 values of π_1 range from .73 to .83.

In Table 11 we show that all the credible intervals contain the true values and the posterior means are close to the true value with the least discrepancy for the near ignorable nonresponse cases. The standard deviations are very similar across the nine simulated examples. Also, the numerical standard errors (nse) are small and similar for all nine simulated examples. The estimates of $\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} | \mathbf{y}, \mathbf{r})$ range from 0.30 to 0.40, except for the most nonignorable nonresponse cases in which (a, b) = (.8, .8) and (.8, .9). Thus, the model does perform reasonably well.

6. CONCLUSION

We have described a Bayesian methodology that can be used to analyze multinomial data for small areas when there is nonignorable nonresponse. A hierarchical model is used, and we have shown that it performs reasonably well. In fact, we have extended the method of Stasny (1991) in two directions: (a) we have considered multinomial data with more than two cells (binomial) and (b) we have done a full Bayesian analysis. Both (a) and (b) have been implemented for small areas

The Markov chain Monte Carlo method permits an assessment of the complex structure of the multinomial nonresponse estimation. Our empirical analysis and simulation study indicate good performance of the model for these data. Thus, the method of ratio estimation currently used in NHANES III may be replaced by our Bayesian method as the nonrespondents' characteristics might differ from those of the respondents. In fact, an application of our model to the NHANES III data shows that in each county there are substantial differences in the proportions of individuals at the three BMI levels by age and sex. This can be seen in Table 1 when the observed counts are summed over the counties. But, we have obtained inference (including measure of precision) for each county by age, race and sex.

Our methodology can be extended in three ways. First, it is feasible to use a model that incorporates an extent of nonignorability, rather than just the dichotomy of ignorable nonresponse and nonignorable nonresponse. Second, one can use other prior distributions (*e.g.*, Dirichlet process prior) to model heterogeneity in the clustering of the areas rather than assuming homogeneity of the areas as we have done. Third, one can use a fourth stage in our model to accommodate clustering within households as well as clustering within areas (counties) in NHANES III. These tasks are very difficult.

ACKNOWLEDGEMENT

This work was done at the National Center For Health Statistics while Balgobin Nandram was the first ASA/NCHS Research Fellow and Geunshik Han was on sabbatical leave from Hanshin University, Korea.

APPENDIX 1

Metropolis-Hastings Samplers

For the ignorable nonresponse model, (μ_1, τ_1) and (μ_{21}, τ_{21}) are independent a posteriori with

$$p(\mu_1, \tau_1 \mid \mathbf{y}, \mathbf{r}) \alpha p(\mu_1, \tau_1) \prod_{i=1}^{c} \left\{ \frac{D(\mathbf{y}_i + n_i - r_i + \mu_1 \tau_1)}{D(\mu_1 \tau_1)} \right\} \quad (A.1)$$

and

$$p(\mu_{21}, \tau_{21} \mid \mathbf{y}, \mathbf{r}) \alpha p(\mu_{21}, \tau_{21})$$
$$\prod_{i=1}^{c} \left\{ \frac{B(r_i + \mu_{21} \tau_{21}, r_i - y_i + (1 - \mu_{21}) \tau_{21})}{B(\mu_{21} \tau_{21}, (1 - \mu_{21}) \tau_{21})} \right\} \quad (A.2)$$

where $p(\mu_1, \tau_1)$ and $p(\mu_{21}, \tau_{21})$ are the prior distributions. Samples can be obtained from each of (A.1) and (A.2) using the MH algorithm of Nandram (1998).

For the nonignorable nonresponse model, it is convenient to condition on $\mathbf{z}$ to obtain

$$p(\mu_3, \tau_3 \mid \mathbf{z}, \mathbf{y}, \mathbf{r}) \alpha p(\mu_3, \tau_3) \prod_{i=1}^{c} \left\{ \frac{D(\mathbf{y}_i + \mathbf{z}_i + \mu_3 \tau_3)}{D(\mu_3 \tau_3)} \right\} \quad (A.3)$$

$$p(\mu_{4j}, \tau_{4j} \mid \mathbf{z}, \mathbf{y}, \mathbf{r}) \alpha p(\mu_{4j}, \tau_{4j})$$
$$\prod_{i=1}^{c} \left\{ \frac{B(y_{ij} + \mu_{4j} \tau_{4j}, z_{ij} + (1 - \mu_{4j}) \tau_{4j})}{B(\mu_{4j} \tau_{4j}, (1 - \mu_{4j}) \tau_{4j})} \right\}, \quad (A.4)$$

where $p(\mu_3, \tau_3), p(\mu_{4j}, \tau_{4j}), j = 1, ..., , J$ are the prior distributions. Given $\mathbf{z}$, (A.3) and (A.4) are independent with

$$p(z_{i1} = t_{i1}, ..., z_{iJ} = t_{iJ} \mid \mathbf{y}, \mathbf{r}, \mu_4, \tau_4, \mu_{3j}, \tau_{3j}, j = 1, ..., J) =$$
$$w_{it_{i1}t_{i2}...t_{iJ}} / \sum_{t_{i1}=0}^{n_i - r_i} \cdots \sum_{t_{iJ}=0}^{n_i - r_i} w_{it_{i1}t_{i2}...t_{iJ}}, \quad (A.5)$$

for $t_{ij} = 0, 1, ..., n_i - r_i, \sum_{j=1}^{J} t_{ij} = n_i - r_i$,

$$w_{it_{i1}t_{i2}...t_{iJ}} = \binom{n_i - r_i}{t_{i1}, ..., t_{iJ}} D(\mathbf{y}_i + \mathbf{t}_i + \mu_3 \tau_3)$$
$$\prod_{j=1}^{J} B(y_{ij} + \mu_{4j} \tau_{4j}, t_{ij} + (1 - \mu_{4j}) \tau_{4j}).$$

We ran the MH sampler by drawing a random deviate from each of (A.3), (A.4), and (A.5). It is easy to draw a random deviate from (A.5). Samples were obtained from each of (A.3), (A.4) and (A.5) using the MH algorithm of Nandram (1998).

APPENDIX 2

Nonlinear least squares estimates

Let

$$v_{ijl} = \log \left\{ \sum_{s=1}^{j} q_{isl} \Big/ \left(1 - \sum_{s=1}^{j} q_{isl} \right) \right\}, j = 1, ..., J - 1 = J'.$$

These v_{ijl} are obtained for each iterate from the Metropolis-Hastings sampler. To solve the nonlinear least squares problem we minimized

$$\sum_{i=1}^{c} \sum_{j=1}^{J'} \sum_{l=1}^{8} \left\{ v_{ijl} - e^{\varphi_i} \left(\theta_j - (\mu_i + \alpha_l) \right) \right\}^2 \quad (A.1)$$

subject to the constraints $\sum_{i=1}^{c} \mu_i = 0, \sum_{j=1}^{J'} \theta_j = 0, \sum_{l=1}^{8} \alpha_k = 0$, and letting $e^{\varphi_i} = \psi_i^{-1}, \sum_{i=1}^{c} 1\text{n} \;\; \psi_i = 0$.

Taking partial derivatives to find the least squares estimate, we have

$$\hat{\varphi}_i = \log \left\{ \frac{\sum_{j=1}^{J'} \sum_{l=1}^{8} v_{ijl} \left(\hat{\theta}_j - \hat{\mu}_i - \hat{\alpha} \right)}{\sum_{j=1}^{J'} \sum_{l=1}^{8} \left(\hat{\theta}_j - \hat{\mu}_i - \hat{\alpha}_l \right)^2} \right\} = \log \psi_i^{-1} \quad (A.2)$$

where

$$\hat{\theta}_j = \left[\sum_{i=1}^{c} e^{2\hat{\varphi}_i} \left\{ \frac{1}{8} \sum_{l=1}^{8} \left(e^{-\hat{\varphi}_i} v_{ijl} + \hat{\mu}_i + \hat{\alpha}_l \right) \right\} \right] \Big/ \sum_{i=1}^{c} e^{2\hat{\varphi}_i}, \quad \text{(A.3)}$$

$$\hat{\mu}_i = \left(\frac{1}{8J'} \right) \sum_{l=1}^{8} \sum_{j=1}^{J'} \left\{ \hat{\theta}_j - \left(\hat{\alpha}_l + e^{-\hat{\varphi}_i} v_{ijl} \right) \right\} \quad \text{(A.4)}$$

and

$$\hat{\alpha}_l = \sum_{i=1}^{c} \frac{1}{J'} \sum_{j=1}^{J'} e^{2\hat{\varphi}_i} \left\{ \hat{\theta}_j - (\hat{\mu}_i + e^{-\hat{\varphi}_i} v_{ijl}) \right\} \Big/ \sum_{i=1}^{c} e^{2\hat{\varphi}_i}. \quad \text{(A.5)}$$

With these settings we draw the q_{ijl} from a MH algorithm, and the nonlinear least squares problem is solved using an iterative method to get values of φ_i, θ_j, μ_i and α_l. Let

$$v_{ijl}^{(h)} = \log \left\{ \sum_{s=1}^{j} q_{isl}^{(h)} \Big/ \left(1 - \sum_{s=1}^{j} q_{isl}^{(h)} \right) \right\},$$

where $q_{isl}^{(h)}$ denotes the value of q_{isl} at the h^{th} iterate of the MH algorithm. Then we minimize (A.l) subject to the above constraints at the h^{th} iterate to obtain $\varphi_i^{(h)}$, $\theta_j^{(h)}$, $\mu_i^{(h)}$ and $\alpha_l^{(h)}$. These iterates provide an estimate of the posterior distributions of φ_i, θ_j, μ_i and α_l. Convergence occurred for our application in less then 10 iterations.

REFERENCES

ALBERT, J.H., and GUPTA, A.K. (1985). Bayesian methods for binomial data with applications to a nonresponse problem. *Journal of the American statistical Association.* 80, 167-174.

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.

BASU, D., and PEREIRA, C.A. (1982). On the Bayesian analysis of categorical data: The Problem of nonresponse. *Journal of Statistical Planning and Inference.* 6, 345-362.

CRAWFORD, S. L., JOHNSON, W.G. and LAIRD, N.M. (1993). Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey (with discussions). In *Case Studies in Bayesian Statistics* (Eds. C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Sinpurwalla). New York: Springer-Verlag, 78-117.

DE HEER, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics.* 15, 129-142.

DEELY, J.J., and LINDLEY, D.V. (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association.* 76, 833-841.

FORSTER, J.J., and SMITH, 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.

GROVES, R.M., and COUPER, M.P. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons, Inc.

HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement.* 5, 475-492.

KUCZMARSKI, R.J., CARROL, M.D., FLEGAL, K.M. and TROIANO, R. P. (1997). Varying body mass index cutoff points to describe overweight prevalence among U. S. adults: NHANES III (1988 to 1994). *Obesity Research.* 5, 542-548.

KAUFMAN, G.M., and KING, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association.* 68, 670-678.

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

MALEC, D., DAVIS, W. and CAO, X. (1999). Model-based small area estimates of over-weight prevalence using sample selection adjustment. *Statistics in Medicine.* 18, 3189-3200.

MOHADJER, L., BELL, B. and WAKSBERG, J. (1994). National Health and Nutrition Examination Survey III-accounting for item nonresponse bias. *National Center for Health Statistics.*

NANDRAM, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation.* 61, 97-126.

NATIONAL CENTER FOR HEALTH STATISTICS (1992). Third National Health and Nutrition Examination Survey. *Vital and Health Statistics Series.* 2, 113.

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey. *Vital and Health Statistics, Series* 1, 32.

OLSON, R.L. (1980). A least squares correction for selectivity bias. *Econometrica.* 48, 1815-1820

PARK, T. (1998). An approach to categorical data nonignorable nonresponse. *Biometrics.* 54, 1579-1590.

PARK, T., and BROWN, M.B. (1994). Models for categorical data with nonignorable non-response. *Journal of the American Statistical Association.* 89, 44-52.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons, Inc.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika.* 63, 581-590.

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

SCHAFER, J.L., EZZATI-RICE, T.M., JOHNSON, W., KHARE, M., LITTLE, R.J.A. and RUBIN, D.B. (1996). The NHANES III multiple imputation project. Survey Research Methods, *Proceedings of the American Statistical Association.* 28-37.

STASNY, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association.* 86, 296-303.

STASNY, E.A., KADANE, J.B. and FRITSCH, K.S. (1998). On the fairness of death penalty jurors: A comparison of Bayesian models with different levels of hierarchy and various missing-data mechanisms. *Journal of the American Statistical Association.* 93, 464-477.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 157-168
Statistics Canada

Assessing the Bias Associated with Alternative Contact Strategies in Telephone Time-Use Surveys

JAY STEWART[1]

ABSTRACT

In most telephone time-use surveys, respondents are called on one day and asked to report on their activities during the previous day. Given that most respondents are not available on their initial calling day, this feature of telephone time-use surveys introduces the possibility that the probability of interviewing the respondent about a given reference day is correlated with the activities on that reference day. Furthermore, noncontact bias is a more important consideration for time-use surveys than for other surveys, because time-use surveys cannot accept proxy responses. Therefore, it is essential that telephone time-use surveys have a strategy for making subsequent attempts to contact respondents. A contact strategy specifies the contact schedule and the field period. Previous literature has identified two schedules for making subsequent attempts: a convenient-day schedule and a designated-day schedule. Most of these articles recommend the designated-day schedule, but there is little evidence to support this viewpoint. In this paper, we use computer simulations to examine the bias associated with the convenient-day schedule and three variations of the designated-day schedule. The results support using a designated-day schedule, and validate the recommendations of the previous literature. The convenient-day schedule introduces systematic bias: time spent in activities done away from home tends to be overestimated. More importantly, estimates generated using the convenient-day schedule are sensitive to the variance of the contact probability. In contrast a designated-day-with-postponement schedule generates very little bias, and is robust to a wide range of assumptions about the pattern of activities across days of the week.

KEY WORDS: Telephone time-use surveys; Contact strategies; Bias; Computer simulations.

1. INTRODUCTION

Telephone time-use surveys present a unique data collection challenge because respondents are called on one day and asked to report on their activities during the previous day. The challenge arises because most respondents – about 75% (Kalton 1985) – are not contacted on their original calling day, necessitating additional contact attempts. In most surveys, it does not matter when these additional attempts are made, because respondents are being asked to report about a fixed reference period. And in most surveys recall does not suffer too much if respondents are contacted several days after the initial calling day. But in time-use surveys, respondents' ability to recall their activities on a given day falls off dramatically after a day or so, which means that the respondent must be assigned a new reference day if no contact is made on the initial calling day. As we will see below, this scenario introduces the possibility that the probability of interviewing the respondent about a given reference dåy is correlated with the activities on that reference day. Therefore it is essential that these surveys have a strategy for making subsequent attempts to contact respondents that does not introduce bias.

Contact Strategies

A contact strategy is comprised of a contact schedule and a field period. The contact schedule specifies which days of the week that contact attempts will be made, and the field period specifies the maximum number of weeks attempts will be made.

Contact schedules fall into two main categories: designated-day schedules and convenient-day schedules. Both types of schedule randomly assign each respondent to an initial calling day. If the respondent is contacted on the initial calling day, the interviewer attempts to collect information about the reference day, which is the day before the calling day. It is for subsequent contact attempts that these schedules differ.

Under a designated-day schedule, there are two approaches to making subsequent contact attempts. The interviewer could call the respondent on a later date, and ask the respondent to report activities for the original reference day. This approach maintains the original reference day, but extends the recall period. Harvey (1993) recommends allowing a recall period of no more than two days. The second approach is to postpone the interview and assign the respondent to a new reference day. Kalton (1985) recommends postponing the interview by exactly one week, so that the new reference day is the same day of the week as the original reference day.

These approaches are not mutually exclusive. For example, Statistics Canada's designated-day schedule allows interviewers to call respondents up to two days after the reference day (Statistics Canada 1999), and to postpone the interview by one week if the respondent cannot be reached after the second day of attempts. The interview can be postponed no more than three times (Statistics Canada). To illustrate, if the initial reference day is Monday the 1st, the respondent is called on Tuesday the 2nd and, if

[1] Jay Stewart, Office of Employment Research and Program Development, Bureau of Labor Statistics, 2 Massachusetts Avenue, NE Room 4945.

necessary, on Wednesday the 3rd. If no interview is obtained on either of these days, the respondent is called on Tuesday the 9th and, if necessary, on Wednesday the 10th, and asked to report on activities done on Monday the 8th. This process continues until the respondent is interviewed, refuses, or until four weeks pass.

The convenient-day schedule does not maintain the designated reference day. If no contact is made, the interviewer calls on the next day and each subsequent day until the respondent is contacted. Once contact is made, the interviewer attempts to complete the interview or, if the respondent is unwilling to complete the interview at that time, reschedule it to a day that is convenient for the respondent. The reference day is always the day prior to the interview. It is worth noting that because respondents are not likely to schedule interviews on busy days, allowing them to choose their interview day is really no different than the interviewer proposing consecutive days (or calling on consecutive days) until the respondent accepts. Hence, one may think of the convenient-day schedule as being functionally identical to an every-day contact attempt schedule.

A variant of the convenient-day schedule described above was used in the 1992-1994 Environmental Protection Agency (EPA) Time Diary Study conducted by the University of Maryland (see Triplett 1995). Respondents were not assigned to an initial calling day. Instead, they were assigned to either the weekday or the weekend sample. For example, those who were assigned to the weekend sample could be called on Sunday (to report about Saturday) or Monday (to report about Sunday). Interviewers were instructed to make at least 20 call attempts before finalizing the case as noncompleted.

Most methodological papers argue in favor of using a designated-day schedule (Kinsley and O'Donnell 1983; Kalton 1985; Lyberg 1989; Harvey 1993; and Harvey 1999). For example, Lyberg (1989) argues that the convenient-day schedule may introduce bias because "the respondent may choose a day when he/she is not busy, a day he/she is not engaged in socially unacceptable behavior, a day he/she thinks is representative, *etc.*" Kinsley and O'Donnell (1983) argue that the convenient-day schedule could exaggerate the number of events taking place outside the home, because the respondent is more likely to be interviewed on a day that immediately follows a day that he or she was out of the house.

Two of these studies directly compare the designated-day and convenient-day schedules (Kinsley and O'Donnell 1983; Lyberg 1989). In Kinsley and O'Donnell (1983), the experimental design divided the sample into two groups. They found that the two schedules produced similar response rates, and that the demographic composition was similar for both samples. They also found that the estimated time spent away from home was much higher under the convenient-day schedule than under the designated-day schedule. But it is impossible to determine whether the convenient-day schedule overestimates time spent away from home or if the designated-day schedule underestimates time spent away from home, because the truth is not known. In Lyberg (1989), two diaries were collected from each respondent. One was collected using a designated-day schedule and the other was collected using a convenient-day schedule. However, the convenient-day diaries were conducted by an interviewer, while the designated-day diaries were self-administered several days after the convenient-day interview. So it is impossible to determine whether any differences were due to differences in contact schedules or whether they were due to mode effects.

Two studies (Lyberg 1989; Laaksonen and Pääkkönen 1992) investigate the effect of postponement on response rates. Both studies found that postponement increases response rates. Laaksonen and Pääkkönen (1992) also found that it was difficult to evaluate whether postponement introduces bias. Their results showed that respondents who postponed their interview spent less time on housekeeping and maintenance, and more time on shopping and errands. However, it is unclear whether these differences are the result of bias introduced by postponement, unobserved heterogeneity that is correlated with the postponement probability, or simply random noise. In any case, they argued that the differences were small, so that any bias was small.

One advantage of the convenient-day schedule is that it is possible to make many contact attempts in a short period of time. In contrast, the designated-day schedule – as proposed – permits only one contact attempt per week. So it is natural to ask: Would it be reasonable to modify the designated-day schedule to allow some form of day-of-week substitution? For example, if the respondent cannot be reached on Tuesday to report about Monday, would it be acceptable to contact the respondent on, say, Thursday and ask him or her to report about Wednesday? This modified schedule would allow for more contact attempts without having to extend the field period.

Because this type of substitution makes sense only if the substitute days are fairly similar to the original days, the first step was to determine which days, if any, were similar to one another. In earlier work, Stewart (2000) showed that Monday through Thursday are very similar to each other, Fridays are slightly different from the other weekdays, and Saturday and Sunday are very different from the weekdays and from each other. Hence, it would be reasonable to allow day-of-week substitution at least for Monday through Thursday.

Activity Bias and Noncontact Bias

When selecting a contact strategy, we need to be concerned with two types of bias: activity bias and noncontact bias. Activity bias occurs when the probability of contacting and interviewing a potential respondent on a particular day is correlated with his or her activities on that day. Note that here and throughout the paper, the term contact probability refers to the probability of a productive contact (one that results in an interview). In order to isolate the effects of using alternative contact strategies, it is assumed that respondents always agree to an interview when contacted. Noncontact bias occurs when differences in contact probabilities across individuals are caused by differences in activities across individuals. Two simple numerical examples will illustrate these biases.

Example 1 – Activity Bias: Suppose that potential respondents' days fall into two categories: hard-to-contact (HTC) days and easy-to-contact (ETC) days. Further suppose that interviewers never contact respondents on HTC days (*i.e.*, that $P_H = 0$, where P_H is the contact probability on an HTC day), and that they always contact respondents on ETC days (*i.e.*, that $P_E = 1$, where P_E is the contact probability on an ETC day). Finally, suppose that the probability that any day is an ETC day is 0.5, so that on average half of each potential respondent's days are ETC and half are HTC. Note that all potential respondents are identical in the sense that the probability that any given day is an ETC day is 0.5 for all potential respondents. For simplicity, I assume that the activities of a given day can be summarized by an "activity index," I_J, where $I_J = 1 - P_J$ ($J = H, E$). The activity index represents time spent in activities that are negatively correlated with the contact probability. Thus, HTC days are days in which more time is spent in activities that are done away from home (working, shopping, active leisure, *etc.*), while ETC days are days in which more time is spent in activities that are done at home (housework, passive leisure, *etc.*). The true average activity index for the population of potential respondents is 0.5 ($= 0.5 \times 1 + 0.5 \times 0$).

If a convenient-day contact schedule is used and there is no limit on the number of call-backs, then HTC days are oversampled. To see why this occurs, it is instructive to work through the two possible contact sequences. If the initial contact attempt occurs on an ETC day, then the respondent is contacted and asked about the previous day (the diary day). Because HTC and ETC days are equally likely, on average half of these diary days will be HTC and the other half will be ETC. Therefore, the average activity index for the diary days of these respondents is equal to 0.5, which is the same as the population average. If, on the other hand, the initial contact day is an HTC day, then no interview takes place and the respondent is called on the following day. Contact attempts continue every day until the respondent is reached (on an ETC day). The average activity index for the diary days of these respondents is equal to one, because the respondent is always interviewed on an ETC day that immediately follows an HTC day. So if a given day is HTC (*i.e.*, the respondent does a lot of activities away from home), then it is more likely that that day will be selected as the reference day. Hence, the probability of interviewing the respondent on a given reference day is correlated with the activities on that reference day. Since half of the initial contact attempts are made on HTC days and half are made on ETC days, the average activity index for the final sample is equal to 0.75 ($= 0.5 \times 0.5 + 0.5 \times 1$).

Example 2 – Noncontact Bias: Now suppose that potential respondents differ with respect to their contact probabilities, and that the contact probabilities for each individual do not vary from day to day. Suppose also that half of all potential respondents are HTC, with $P_H = 0.25$, and that the other half are ETC, with $P_E = 0.75$. If we attempt to contact each potential respondent four times, given these probabilities, virtually all (99.6%) ETC potential respondents are contacted. In contrast, only 68.4% of HTC potential respondents are contacted. The overall contact rate is 84% (99.6% × 0.50 + 68.4% × 0.50), but the final sample is not representative: 59.3% of the sample are ETC and only 40.7 % are HTC. Therefore, estimates based on this sample will tend to underestimate the time spent in activities done by HTC people, and overestimate the time spent in activities done by ETC people.

The biases described above are not limited to time-use surveys. Although most surveys take steps to minimize noncontact bias, less attention has been devoted to activity bias. For example, in addition to their main focus on collecting event history information on employment, the National Longitudinal Surveys also include a few questions about labor force activities (employment and hours) during the week prior to the interview. Because these interviews tend to be scheduled at the convenience of the respondent, the respondent's activities during the reference week will be correlated with the probability of interviewing the respondent about that reference week. The intuition behind this correlation is exactly the same as that in Example 1. This correlation introduces bias into hours-worked estimates, although the direction of the bias is indeterminate. Hours worked per week tend to be overestimated for respondents who were unable to schedule an interview because of a heavy work schedule, and tend to be underestimated for respondents who were away on vacation. Activity bias is also an issue for travel surveys. Time spent away from home will tend to be overestimated if respondents are asked about, say, the four weeks prior to the interview. Asking respondents about a fixed reference period can eliminate this bias.

It is worth noting that noncontact bias is a more important consideration for time-use surveys than for other surveys, because, unlike most other surveys, time-use surveys cannot accept proxy responses. If proxy responses could be accepted then data on HTC individuals could be collected from proxies, who may be easier to contact. This would weaken the correlation between the individual's activities and the probability of collecting data about that individual.

The rest of the paper is organized as follows. In section 2, four contact strategies are introduced, and simple simulations are used to assess the activity bias associated with each strategy. In section 3, the simulations are augmented with data from the May 1997 Work Schedule Supplement to the Current Population Survey and the 1992-94 University of Maryland Time Diary Study, and how the bias varies by specific activity is examined. In addition, the overall bias is decomposed to assess the relative contribution of activity bias and noncontact bias. Section 4 summarizes these results and makes recommendations.

2. CONTACT STRATEGIES, CORRELATED ACTIVITIES, AND ACTIVITY BIAS

In this section, the activity biases associated with the convenient-day schedule and each of the three variants of the designated-day schedule are compared. These schedules are defined as follows:

1. Convenient day (CD): Attempt to contact potential respondents every day following the initial contact attempt until contact is made or until the field period ends.
2. Designated day (DD): Attempt to contact potential respondents only once (no subsequent attempts).
3. Designated day with postponement (DDP): Attempt to contact potential respondents on the same day of the week as the initial attempt until contact is made or until the field period ends (as recommended by Kalton 1985).
4. Designated day with postponement and substitution (DDPS): Attempt to contact potential respondents every other day following the initial contact attempt until contact is made or until the field period ends.

The DDPS schedule assumes alternating Tuesday/Thursday and Wednesday/Friday contact days. Whether the first week is Tuesday/Thursday or Wednesday/Friday depends on the start day, which is randomly assigned.

As seen in Example 1, it is straightforward to show that a convenient-day schedule can introduce activity bias into time-use estimates when the base contact probability is the same each day (0.5) except for random noise (+0.5 with probability ½ or -0.5 with probability ½). Even though Stewart (2000) shows that Monday through Thursday are very similar on average, it is likely that the contact probabilities for some individuals vary systematically by day each week. For example, some individuals may be hard to contact on Monday, Wednesday, and Friday of each week. This systematic variation makes it considerably more complicated to determine whether sample estimates are biased, and to determine the direction and extent of that bias. One could model contact strategies and analytically solve for the bias under different assumptions about the pattern of contact probabilities. However, this is a cumbersome process, because each assumption about the pattern of contact probabilities across days would require a separate solution. In contrast, computer simulations are an ideal way to assess the bias associated with alternative contact strategies under different assumptions about the pattern of contact probabilities. The computer program is simpler and produces more intuitive results than the analytical solution. And it is easy to modify the program to allow for different patterns. In section 3, realism is added to the simulations by incorporating real time-use data – something that would be impossible to do when taking an analytical approach.

Simulations

The simulation strategy was very straightforward. First, four weeks worth of "data" for each of 10,000 potential respondents was created. In order to focus on contact strategies, the sampling procedures are ignored and it is assumed that the sample of potential respondents is representative of the population. The simulations are designed to compare the four contact schedules above, so it is assumed that the "week" is five days long. Eligible diary days were restricted to Monday through Thursday, because, as noted above, these days are the most similar to each other. The next step was to simulate attempts to contact these respondents using the four contact schedules described above. Finally, the estimates generated using each schedule were compared to the true sample values.

To simplify the simulations I abstracted from specific activities, as in the examples above, and characterized each day using an activity index, I_J, $(J = H, E)$ that ranges from 0 to 1. The activity index is given by $I_J = 1 - P_J$ where P_J is the probability of contacting and interviewing the respondent. To simulate the variation in activities across days, the contact probability on a given day is:

$$P_J = \bar{P}_J + \varepsilon,$$

where $\bar{P}_J$ is the average contact probability on an HTC (J = H) or an ETC (J = E) day, and $\varepsilon \sim U(-\hat{\varepsilon}, \hat{\varepsilon})$. I assume that $\bar{P}_H < \bar{P}_E$, which means that, on average, respondents are less likely to be contacted on HTC days than on ETC days. To insure that contact probabilities lie in the [0,1] interval, I set $\hat{\varepsilon}$ so that $\hat{\varepsilon} < \min(\bar{P}_H, 1 - \bar{P}_E)$.

There are many assumptions one can make regarding the pattern of activities across days. The simplest case is where all days are identical except for random noise. But as noted above, it is possible that potential respondents are systematically harder to contact on some days than others. To cover a wide range of activity patterns, the simulations were performed under the following eight assumptions about the pattern of HTC and ETC days in each of the four weeks:

1. Actual values of the activity index are distributed as U(0,1), so that the average value is 0.5.
2. The first two days of every week are HTC and the last three days are ETC (HHEEE).
3. The first three days of every week are HTC and the last two days are ETC (HHHEE).

4. The first four days of every week are HTC and the last day is ETC (HHHHE).
5. The first day of every week is ETC and the last four are HTC (EHHHH).
6. The first two days of every week are ETC and the last three are HTC (EEHHH).
7. The first three days of every week are ETC and the last two are HTC (EEEHH).
8. For half the sample Monday, Wednesday, and Friday are HTC and Tuesday and Thursday are ETC (HEHEH). For the other half of the sample the reverse is true (EHEHE).

In pattern 1, the base probability of contacting the respondent is the same, so that all of the variation in probabilities is due to the random term. In patterns 2-7, HTC days are grouped together either at the beginning of the week or at the end of the week. And in pattern 8, the base probabilities alternate between HTC and ETC days. To focus on activity bias, separate simulations were performed for each of the 8 patterns described above. Thus, within a simulation all individuals have the same pattern of base probabilities.

Table 1 shows the results from a representative subset of the 153 simulations performed. The first four columns show the average contact probability on HTC and ETC days, the value of $\hat{\varepsilon}$, and the true average activity index. The remaining columns contain estimates of the bias associated with the four contact schedules. The bias was computed as the difference between the estimated amount of time spent in each activity and the true amount of time spent in each activity, and then the difference was expressed as a percentage of the true value. Entries with an asterisk indicate that the bias is statistically different from the zero at the 5% level.

Table 1
Activity Bias Associated with Each Contact Strategy Under Alternative Assumptions About the Correlation of Activities Across Days

	Average Contact Probability				Estimated Bias (Expressed as a percent of the true activity index)			
Activity Pattern	Hard-to-contact days	Easy-to-contact days	$\hat{\varepsilon}$	True Average Activity Index	CD	DD	DDP	DDPS
Identical Base Probabilities								
	0.50		0.10	0.500	0.7*	-0.1	0.0	0.1
	0.50		0.30	0.500	5.3*	-0.3	0.1	0.2
	0.50		0.50	0.500	15.1*	-0.9	0.4	0.7
Grouped Base Probabilities								
HHEEE	0.75	0.25	0.05	0.500	0.7	-10.7*	-4.7*	-13.8*
	0.75	0.25	0.25	0.500	5.2*	-10.9*	-4.8*	-13.9*
	0.60	0.40	0.05	0.500	-0.1	-2.2*	-0.7*	-2.8*
	0.60	0.40	0.20	0.500	2.5*	-2.6*	-0.7*	-2.5*
HHHEE	0.75	0.25	0.05	0.625	-2.7*	-9.7*	-4.0*	-12.7*
	0.75	0.25	0.25	0.625	0.8	-10.3*	-4.1*	-12.8*
	0.60	0.40	0.05	0.550	-0.4*	-1.8*	-0.6*	-2.5*
	0.60	0.40	0.20	0.550	1.9*	-2.4*	-0.5	-2.2*
HHHHE	0.75	0.25	0.05	0.750	0.1	-0.1	0.1	0.0
	0.75	0.25	0.25	0.750	2.3*	-0.5	0.2	0.2
	0.60	0.40	0.05	0.600	0.1*	0.0	0.0	0.0
	0.60	0.40	0.20	0.600	1.9*	-0.3	0.2	0.2
EHHHH	0.75	0.25	0.05	0.625	1.7*	1.0	1.4*	0.7
	0.75	0.25	0.25	0.625	4.2*	-0.3	1.2*	0.7
	0.60	0.40	0.05	0.550	1.1*	0.3	0.5*	0.3
	0.60	0.40	0.20	0.550	2.9*	0.0	0.6*	0.4
EEHHH	0.75	0.25	0.05	0.500	-18.2*	-17.1*	-4.3*	-21.7*
	0.75	0.25	0.25	0.500	-15.9*	-17.9*	-4.5*	-20.9*
	0.60	0.40	0.05	0.500	-2.0*	-2.2*	-0.4	-2.6*
	0.60	0.40	0.20	0.500	-0.4	-2.4*	-0.3	-2.6*
EEEHH	0.75	0.25	0.05	0.375	-16.6*	-17.6*	-5.5*	-20.3*
	0.75	0.25	0.25	0.375	-11.4*	-17.6*	-5.6*	-19.6*
	0.60	0.40	0.05	0.450	-2.0*	-2.3*	-0.4	-2.5*
	0.60	0.40	0.20	0.450	0.0	-2.5*	-0.5	-2.5*
Alternating Base Probabilities								
HEHEH/EHEHE	0.75	0.25	0.05	0.500	31.5*	26.4*	9.6*	28.5*
	0.75	0.25	0.25	0.500	34.7*	26.5*	9.7*	29.4*
	0.60	0.40	0.05	0.500	5.6*	4.5*	1.3*	5.1*
	0.60	0.40	0.20	0.500	7.8*	4.3*	1.2*	5.1*

Note: Asterisks indicate that the estimated average activity index is statistically different from the true value at the 5% level.

Pattern 1 – Identical Base Probabilities with Random Noise

This pattern is essentially the same as in the numerical example above. The main result is that all of the contact schedules generate unbiased estimates for the average activity index, except the CD schedule. As expected, the CD schedule overestimates the average activity index. More importantly, when using the CD schedule, the estimated average activity index – and hence the bias when activities are uncorrelated across days – is positively correlated with the variance of ε. As the variance increases from 0.003 ($\hat{\varepsilon} = 0.1$) to 0.083 ($\hat{\varepsilon} = 0.5$), the bias increases from less than 1% to 15%. One can see the intuition behind this result by noting that a large negative realization of ε on a particular day makes it less likely that the respondent will be contacted on that day, and hence, more likely that that day will become the diary day. None of the other contact schedules are sensitive to the variance of ε.

Patterns 2-7 – Grouped Base Probabilities

The results are mixed when HTC days are grouped at either the beginning or the end of the week. In the simulations where $\bar{P}_E - \bar{P}_H$ is relatively small (0.2), all of the contact schedules perform reasonably well. The absolute value of the bias is less than 3% in all cases. However, when $\bar{P}_E - \bar{P}_H$ is relatively large (0.5), there are significant differences in the bias associated with each contact schedule. The DDP schedule performs the best overall. The bias exceeds 5% (in absolute value) only in pattern 7 (EEEHH), for which the bias is – 5.5%. In contrast, when using the DD and DDPS schedules, the bias is in the 10 – 14% range in patterns 2 (HHEEE), 3 (HHHEE), and in the 16-20% range in patterns 6 (EEHHH), and 7 (EEEHH). The differences between the DD and DDPS schedules and the DDP schedule for these patterns are significant, both statistically and in practical terms. In patterns 4 (HHHHE) and 5 (EHHHH) the DDP schedule performs slightly worse than the DD and DDPS schedules, but the bias is so small (less than 1.5%) that the difference is of no practical significance. The CD schedule fares somewhat better than the DD and DDPS schedules. The bias is less than 5%, except in patterns 6 and 7 where the bias is in the 11 – 18% range. As in pattern 1 above, the estimated average activity index increases with the variance of ε under the CD schedule, but not under any of the other schedules. And as can be seen from Table 1, in patterns where the bias is negative (patterns 6 and 7), an increase in the variance of ε decreases the bias.

Pattern 8 – Alternating Base Probabilities

All of the contact schedules generate biased estimates, because ETC days are undersampled. As above, all of the schedules perform reasonably well when $\bar{P}_E - \bar{P}_H$ is relatively small. The bias is in the 5-8% range for all schedules except DDP, for which the bias is about 1 %. However, when $\bar{P}_E - \bar{P}_H$ is large, all of the contact schedules generate significant bias. The bias of about 10% for the DDP schedule is higher than for the other patterns but it is smaller than the 25-35% bias for the other schedules. Again, these differences are significant statistically, and they are significant in practical terms.

The reason that the DDPS schedule generates a large activity bias is that contact attempts are made on two HTC days and then on two ETC days (or the reverse). This pattern results in contacting respondents on a relatively large fraction of ETC days, and hence, diary days will be disproportionately HTC days. Not surprisingly, if the DDPS schedule is modified so the respondent is contacted on the same two days each week, there is virtually no bias.

It is clear from these simulations that the activity bias associated with each contact schedule depends on the pattern of activities across days, the contact probabilities on HTC and ETC days, and the variance of those probabilities. However, it is also clear that the DDP schedule outperforms the other schedules regardless of the pattern assumed. If each pattern is viewed as a different type of respondent, then the overall bias (which includes both activity and noncontact bias) depends on the relative frequency of each type in the population. Information on the incidence of each type would allow one to measure the overall bias, and, for each strategy, decompose the overall bias it into the portion due to activity bias, and the portion due to noncontact bias. This is investigated in the next section.

3. AUGMENTED SIMULATIONS

If one is willing to make some additional assumptions, it is possible to augment the simulations using data from other sources. The first assumption is that individuals' work schedules are a reasonable proxy for the patterns of HTC and ETC days, so that work days correspond to HTC days and nonwork days correspond to ETC days. The second assumption is that it is possible to replicate an individual's week by taking one day from each of five individuals.

Data from the May 1997 Work Schedule Supplement to the Current Population Survey (CPS) were used to obtain information about individuals' work schedules. Note that because of the need to know the prevalence of each type of schedule for the entire population, nonworkers were also included. Table 2 shows the patterns of work (W) days and nonwork (N) days from the May 1997 CPS. Approximately 88% of all individuals fall into two patterns. Forty-eight percent work all five weekdays, and 39% do not work any weekdays. Another 4% work four weekdays and have either Friday or Monday off. The remaining individuals do not exhibit any discernible pattern. To simplify the simulations, it was assumed that individuals either worked all 5 weekdays (workers) or that they did not work any weekdays (nonworkers).

Table 2
Distribution of Work Schedules

Activity Pattern						
M	Tu	W	Th	F	Percent	Cumulative Percent
–	–	–	–	–	39.40	39.40
W	W	W	W	W	48.11	87.51
W	W	W	W	–	2.63	90.14
–	W	W	W	W	1.63	91.77
W	W	W	–	–	0.81	92.58
W	W	–	–	–	0.26	92.84
–	–	–	W	W	0.37	93.21
–	–	W	W	W	0.68	93.89
W	–	W	–	W	0.49	94.38
–	W	–	W	–	0.25	94.63
–	–	–	–	W	0.51	95.14
W	–	–	–	–	0.25	95.39
W	W	–	W	W	0.73	96.12
W	–	–		W	0.36	96.48
W	–	–	W	W	0.70	97.18
Other patterns					2.82	100.00
Total					100.00	

Note: A "W" indicates a workday, and a "-" indicates a nonwork day. Author's tabulations from the May 1997 Work Schedule Supplement to the CPS. Observations were weighted using supplement weights. The sample size is 89,746 observations.

To generate information on individual activities, data from the 1992-94 EPA Time Diary Study, conducted by the University of Maryland were used. This dataset contains time-diaries for a sample of 7,408 adults (see Triplett 1995). Because each individual was interviewed only once, there is only one observation per person. The following repeated sampling method was used to construct 8 weeks worth of data for a sample of 18,974 "individuals." The diary data were divided into workdays and nonwork days. A diary day was considered a workday if the individual did any paid work during the day. Workdays were assigned to workers and nonwork days were assigned to nonworkers. Mondays were drawn from Monday observations, Tuesdays were drawn from Tuesday observations, *etc.* No observation was used more than once for a given individual, but the same observation could be used for more than one individual. The final sample proportions look fairly similar to the proportions from the CPS. Fifty-eight percent of individuals in the final sample were workers and 42% were nonworkers, which is reasonably close to the ratio of workers to nonworkers (1.38 *vs.* 1.23) in the CPS.

To compute the contact probabilities, it was necessary to make a third assumption. Following Pothoff, Manton, and Woodbury (1993), the contact probability was assumed to be equal to the number of minutes spent in activities done at home (excluding sleeping) divided by the time spent in all activities other than sleep. This process for generating contact probabilities has two important properties: (1) the contact probability for a given day is related to the activities done on that day, and (2) one group of potential respondents (workers) has a lower average probability of a productive contact (0.36 *vs.* 0.72).

Tables 3a and 3b summarize the bias estimates from the augmented simulations. Table 3a shows the bias estimates assuming a 4-week field period, and Table 3b shows the same estimates assuming an 8-week field period. Each of the first four columns contains estimates of the bias associated with the four contact strategies. The entries for each strategy and each 1-digit activity include estimates of the activity bias for workers and nonworkers, and an estimate of the overall bias. The overall bias includes noncontact bias, so it is possible that the overall bias is larger (or smaller) than the activity bias for either group. The bias was computed as in the previous simulations, strategy and as before, an asterisk indicates that the bias is significantly different from the zero at the 5% level. The fifth column shows the true time spent in each activity by group and overall.

Comparing Tables 3a and 3b, we can see that the main difference is that, except for the DD strategy for which the field period is irrelevant, the overall bias is smaller when the field period is 8 weeks. This smaller overall bias is due mainly to the increased number of contact attempts, which disproportionately increases the probability that workers are contacted and makes the sample more representative (see Table 4). In contrast, estimates of the activity bias associated with the various contact strategies are not sensitive to the length of the contact period. The rest of this discussion will focus on the results in Table 3b.

The DD strategy generated virtually no activity bias. There were a few activities – Active Leisure, Entertainment/Socializing, Organizational Activities, Education/Training, and Active Child Care for workers, and Active Child Care for nonworkers – for which the activity bias was rather large, but none of these bias estimates are statistically significant. The overall bias for the DD strategy is quite large for most activities, which, as will be seen below, is primarily due to noncontact bias.

Comparing the other three strategies, one can see two patterns emerge. First, activity bias is significantly smaller (and generally not statistically significant) when using the DDP strategy or the DDPS strategy than when using the CD strategy. Second, the bias in the CD estimates follows the expected pattern. The bias tends to be positive for activities that are done away from home (Active Leisure, Entertainment/Socializing, Organizational Activities, Education/Training, Purchasing Goods/Services, and Paid Work), and negative for activities done at home (Passive Leisure, Personal Care, Active Child Care, and Housework). This pattern is consistent with research cited in the introduction that finds that reported time spent away from home is greater under a convenience-day strategy than under a designated-day strategy. More important, it is now clear that this finding is due to bias in convenient-day strategies rather than bias in designated-day strategies.

Table 3a
Estimated Bias – Augmented Simulations (4 Week Field Period)

Activity/Emp. Status					Time Spent in
Employment Status	CD	DD	DDP	DDPS	Activity (Truth)
Passive Leisure					
Nonworkers	-8.44*	0.12	-1.54	-1.03	314.72
Workers	-5.40*	1.07	0.43	0.82	152.04
Overall	-8.62*	13.56*	2.53*	0.38	220.70
Active Leisure					
Nonworkers	9.80*	-2.75	0.99	-0.66	65.94
Workers	-0.07	-7.34	-4.69	1.91	26.89
Overall	4.03*	11.75*	3.31	1.08	43.37
Entertainment/Socializing					
Nonworkers	19.41*	-2.01	-0.25	-1.20	67.30
Workers	8.63*	7.14	5.21	3.72	27.87
Overall	13.11*	15.78*	5.64*	1.37	44.51
Organizational Activities					
Nonworkers	19.58*	-0.98	9.00	3.84	19.25
Workers	13.77*	6.95	7.17	7.48	8.72
Overall	15.24*	15.26*	12.37*	5.99	13.16
Education/Training					
Nonworkers	32.77*	-0.42	12.54*	8.92*	43.60
Workers	-1.17	7.63	0.57	1.59	13.16
Overall	19.17*	22.02*	15.39*	8.00*	26.01
Personal Care					
Nonworkers	-0.50	-0.29	-0.49	-0.44	663.04
Workers	-0.52*	0.01	-0.06	-0.13	580.71
Overall	-0.79*	2.20*	0.34	-0.15	615.46
Purchasing Goods/Services					
Nonworkers	12.62*	1.35	0.11	-1.28	72.98
Workers	-4.05	4.62	-3.62	-5.43*	23.28
Overall	4.67*	22.36*	4.25*	-1.49	44.25
Active Child Care					
Nonworkers	-7.89*	5.11	-1.06	-0.54	24.13
Workers	-7.69*	-6.05	-4.09	-0.92	12.64
Overall	-9.09*	14.21*	0.77	-0.09	17.49
Housework					
Nonworkers	-8.88*	1.71	0.33	2.27	169.04
Workers	-10.55*	0.85	-2.03	-0.14	57.92
Overall	-11.49*	20.77*	4.53*	2.52*	104.82
Paid Work					
Nonworkers	—	—	—	—	—
Workers	2.95*	-0.77	0.25	-0.27	536.77
Overall	6.74*	31.44*	-7.74*	-1.87*	310.22

Note: Asterisks indicate that the bias in the estimated time spent in the activity is significantly different from zero at the 5% level.

Noncontact Bias

In general, the contact rate increases and the sample becomes more representative as the number of contact attempts increases (see Table 4). The contact rate is the lowest under the DD strategy (40%), and the sample is the least representative. Under both the DDP and the DDPS schedules, the contact rate increases and the sample becomes more representative as the field period increases from 4 to 8 weeks. Using a DDPS schedule with an 8-week field period (16 contact attempts) results in a contact rate of 80% and a representative sample. Not surprisingly, the sample generated by the DDP schedule with an 8 week field period is virtually identical to the one generated by the DDPS schedule with a 4 week field period.

Activity Bias *vs.* Noncontact Bias

To get a clearer picture of the contribution of each type of bias to the overall bias, the overall bias was decomposed into the portion due to activity bias, the portion due to noncontact bias, and the portion due to the interaction between the two biases. The overall bias for activity *a* and group *g* (workers or nonworkers) is given by:

$$F_g X_{ag} - F_g^* X_{ag}^* =$$

$$\underbrace{F_g^*(X_{ag} - X_{ag}^*)}_{\text{Activity}} + \underbrace{X_{ag}^*(F_g - F_g^*)}_{\text{Noncontact}} + \underbrace{(F_g - F_g^*)(X_{ag} - X_{ag}^*)}_{\text{Interaction}}$$

Table 3b
Estimated Bias – Augmented Simulations (8 Week Field Period)

Activity/Emp. Status					Time Spent in
Employment Status	CD	DD	DDP	DDPS	Activity (Truth)
Passive Leisure					
Nonworkers	-8.63*	-0.09	-1.62	-1.21	315.38
Workers	-5.24*	1.28	0.39	1.10	151.72
Overall	-8.72*	-13.51*	-0.35	-0.31	220.79
Active Leisure					
Nonworkers	10.62*	-2.03	1.76	0.06	65.46
Workers	0.00	-7.29	-3.50	2.21	26.87
Overall	4.49*	12.30*	0.50	0.82	43.16
Entertainment/Socializing					
Nonworkers	19.77*	-1.72	-0.15	-0.91	67.10
Workers	8.09*	6.64	5.52	2.76	28.00
Overall	13.06*	15.80*	2.47	0.40	44.50
Organizational Activities					
Nonworkers	18.92*	-1.53	8.59	3.25	19.36
Workers	14.03*	7.00	3.18	7.25	8.72
Overall	14.89*	14.88*	7.14*	4.76	13.21
Education/Training					
Nonworkers	33.56*	0.18	12.91*	9.55*	43.34
Workers	-0.72	8.24	0.77	2.01	13.09
Overall	19.73*	22.74*	10.29*	7.32*	25.86
Personal Care					
Nonworkers	-0.50	-0.29	-0.48	-0.44	663.03
Workers	-0.55*	0.00	-0.08	-0.16	580.81
Overall	-0.82*	2.20*	-0.17	-0.29	615.51
Purchasing Goods/Services					
Nonworkers	12.64*	1.36	-0.09	-1.28	72.97
Workers	-4.41	4.23	-3.66	-5.45*	23.36
Overall	4.48*	22.23*	-0.42	-2.58	44.30
Active Child Care					
Nonworkers	-7.67*	5.36	-1.04	-0.31	24.07
Workers	-8.02*	-6.18	-4.98	-1.65	12.66
Overall	-9.14*	14.30*	-2.23	-0.89	17.48
Housework					
Nonworkers	-9.02*	1.55	0.20	2.10	169.30
Workers	-10.55*	0.80	-2.15	-0.20	57.95
Overall	-11.64*	20.63*	0.17	1.34	104.94
Paid Work					
Nonworkers	—	—	—	—	—
Workers	2.96*	-0.78	0.30	-0.26	536.82
Overall	6.86*	-31.44*	-0.86	-0.22	310.25

Note: Asterisks indicate that the bias in the estimated time spent in the activity is significantly different from zero at the 5% level.

Table 4
Contact Rate Summary – Augmented Simulations

Field Period		CD	DD	DDP	DDPS	Truth
4 weeks	Contact Rate	89.68	40.35	71.79	78.39	
	Percent Nonworkers	40.08	60.07	46.82	43.14	42.21
	Percent Workers	59.92	39.93	53.18	56.86	57.79
8 weeks	Contact Rate	89.79	40.35	78.87	80.17	
	Percent Nonworkers	40.02	60.07	42.88	42.19	42.21
	Percent Workers	59.98	39.93	57.12	57.81	57.79

where F_g is the fraction of the sample in group *g*, and X_{ag} is the time spent in activity *a* by group *g*, and asterisks indicate the true values. The total bias for activity *a* is obtained by summing this expression over workers and nonworkers, and is given by:

$$\begin{aligned}\sum_{g=W,N}(F_gX_{ag}-F_g^*X_{ag}^*) &= \sum_{g=W,N}F_g^*(X_{ag}-X_{ag}^*)\\ &+\sum_{g=W,N}X_{ag}^*(F_g-F_g^*)\\ &+\sum_{g=w,N}(F_g-F_g^*)(X_{ag}-X_{ag}^*),\end{aligned}$$

there are several things to take from these decompositions (shown in Table 5). First, under the CD schedule, all of the overall bias is due to activity bias. The large number of contact attempts virtually guarantees a representative sample, so that increasing the field period from 4 to 8 weeks does not make much difference. In contrast, noncontact bias accounts for all of the bias under the DD schedule. Under both the DDP schedule and the DDPS schedule there is virtually no activity bias, and noncontact bias decreases dramatically as the field period is increased from 4 to 8 weeks. Not surprisingly, the noncontact bias for the DDP schedule with an 8-week field period is about the same as the noncontact bias under the DDPS schedule with a 4-week field period. In these simulations, the sample becomes fully representative when the field period is long enough to allow 16 contact attempts. Finally, the small magnitude of the interaction terms reflects the fact that activity and noncontact biases associated with each contact strategy are negatively correlated.

Table 5
Bias Decomposition – Augmented Simulations

	4 – week field period				8 – week field period			
	Total Bias	Activity Bias	Noncontact Bias	Interaction	Total Bias	Activity Bias	Noncontact Bias	Interaction
Passive Leisure								
CD	-8.62	-7.23	-1.57	0.18	-8.72	-7.29	-1.62	0.19
DD	13.56	0.50	13.16	-0.10	13.51	0.46	13.24	-0.18
DDP	2.53	-0.75	3.40	-0.11	-0.35	-0.83	0.50	-0.02
DDPS	0.38	-0.29	0.69	-0.02	-0.31	-0.30	-0.01	0.00
Active Leisure								
CD	4.03	6.27	-1.92	-0.32	4.49	6.80	-1.96	-0.35
DD	11.75	-4.40	16.08	0.06	12.30	-3.92	15.97	0.26
DDP	3.31	-1.05	4.15	0.20	0.50	-0.13	0.60	0.03
DDPS	1.08	0.26	0.84	-0.02	0.82	0.83	-0.02	0.00
Entertainment/Socializing								
CD	13.11	15.51	-1.89	-0.51	13.06	15.53	-1.92	-0.54
DD	15.78	1.30	15.82	-1.34	15.80	1.32	15.69	-1.21
DDP	5.64	1.72	4.08	-0.17	2.47	1.91	0.59	-0.02
DDPS	1.37	0.58	0.82	-0.04	0.40	0.42	-0.02	0.00
Organizational Activities								
CD	15.24	17.36	-1.70	-0.42	14.89	17.06	-1.76	-0.40
DD	15.26	2.05	14.28	-1.08	14.88	1.72	14.39	-1.23
DDP	12.37	8.30	3.69	0.39	7.14	6.53	0.54	0.07
DDPS	5.99	5.24	0.74	0.01	4.76	4.77	-0.02	0.00
Education & Training								
CD	19.17	22.84	-2.49	-1.18	19.73	23.53	-2.56	-1.24
DD	22.02	1.94	20.90	-0.82	22.74	2.54	20.90	-0.69
DDP	15.39	9.04	5.40	0.96	10.29	9.36	0.78	0.14
DDPS	8.00	6.78	1.09	0.13	7.32	7.35	-0.02	0.00
Personal Care								
CD	-0.79	-0.51	-0.28	0.00	-0.82	-0.53	-0.29	0.00
DD	2.20	-0.13	2.39	-0.06	2.20	-0.13	2.39	-0.06
DDP	0.34	-0.26	0.62	-0.02	-0.17	-0.26	0.09	0.00
DDPS	-0.15	-0.27	0.12	0.00	-0.29	-0.29	0.00	0.00
Purchasing Goods/Services								
CD	4.67	7.55	-2.39	-0.49	4.48	7.44	-2.45	-0.51
DD	22.36	2.34	20.06	-0.04	22.23	2.23	20.00	0.00
DDP	4.25	-1.02	5.18	0.10	-0.42	-1.18	0.75	0.01
DDPS	-1.49	-2.54	1.04	0.01	-2.58	-2.55	-0.02	0.00
Active Child Care								
CD	-9.09	-7.81	-1.40	0.11	-9.14	-7.82	-1.43	0.10
DD	14.21	0.45	11.72	2.04	14.30	0.53	11.66	2.12
DDP	0.77	-2.32	3.03	0.07	-2.23	-2.69	0.44	0.01
DDPS	-0.09	-0.69	0.61	0.00	-0.89	-0.87	-0.01	0.00
Housework								
CD	-11.49	-9.42	-2.26	0.18	-11.64	-9.51	-2.32	0.19
DD	20.77	1.43	18.93	0.41	20.63	1.31	18.95	0.37
DDP	4.53	-0.43	4.89	0.08	0.17	-0.55	0.71	0.01
DDPS	2.52	1.50	0.99	0.03	1.34	1.36	-0.02	0.00
Paid Work								
CD	6.74	2.95	3.69	0.11	6.86	2.96	3.79	0.11
DD	-31.43	-0.77	-30.90	0.24	-31.44	-0.78	-30.90	0.24
DDP	-7.74	0.25	-7.98	-0.02	-0.86	0.30	-1.16	0.00
DDPS	-1.87	-0.27	-1.61	0.00	-0.22	-0.26	0.03	0.00

4. SUMMARY AND RECOMMENDATIONS

Telephone time-use surveys have unique characteristics that make data collection more challenging. Unlike most other surveys, time-use surveys cannot accept proxy responses, so it is more likely that the probability of contacting a potential respondent is correlated with his or her activities. And because telephone time-use surveys ask respondents to report on their activities during the previous day, it is possible that the probability of interviewing the respondent about a given reference day will be correlated with the activities on that reference day. This paper shows how these characteristics can generate noncontact bias and activity bias. Two sets of computer simulations showed that the extent of these biases depends on the survey's strategy for contacting potential respondents.

In the first set of simulations, it was shown that the extent of the bias associated with any given contact schedule depends on the pattern of easy-to-contact (ETC) and hard-to-contact (HTC) days. The designated-day-with-postponement (DDP) schedule outperformed the other contact schedules for all of the activity patterns examined. These simulations also showed that estimates generated using a convenient-day (CD) schedule are sensitive to the within-person variance of the contact probability. Estimates of the time spent in activities that are positively correlated with the contact probability (for example, activities done at home) decrease as the variance increases. In contrast, estimates generated by other contact schedules are not sensitive to the within-person variance of the contact probability.

Given the results of the simple simulations, it is clear that the overall bias for the different contact strategies depends on the relative frequency of each pattern in the population. Direct data on these patterns do not exist, so the first set of simulations was augmented using CPS data on work schedules and actual time-use data from the 1992-94 EPA Time Diary Study. The results from the augmented simulations confirm those from the simple simulations, and show how the bias can affect estimates of time spent in specific activities. As expected, the CD contact strategy introduces systematic activity bias into time-use estimates. The time spent in activities done at home is underestimated, while time spent in activities done away from home is overestimated. There is no systematic activity bias in the samples generated by the DDP and DDPS strategies. The simulations also show that increasing the number of contact attempts reduces noncontact bias.

These results clearly show that the choice of contact strategy matters and point to two recommendations.

First, time-use surveys should use the DDP schedule. The DDP schedule generates less activity bias than the other contact schedules under all of the activity patterns tested. The DDPS schedule performed nearly as well in the more common activity patterns. But given that contact rates and field costs are a function of the number of contact attempts, the DDPS offers no cost advantage over the DDP schedule. Hence, there is no reason to choose the DDPS schedule over the DDP schedule.

Second, time-use surveys need to take steps to minimize noncontact bias. Because noncontact bias is largely a function of the number of contact attempts, an obvious way to minimize noncontact bias would be to increase the number of contact attempts. No further elaboration will be made on this point, because other authors have looked at this issue in depth. For example, Bauman, Lavradas and Merkle (1993) show that age and employment status are related to the number of callbacks and that additional callbacks generate a more representative sample, and Botman, Massey and Kalsbeek (1989) propose a method for determining the optimal number of callbacks. Another alternative would be to try to increase the probability of contacting potential respondents. This could be done by determining when they are likely to be home and calling at those times, or by allowing them to call on their designated interview day. Paying incentives is another way to make potential respondents become "more available." A less costly approach to minimizing noncontact bias would be to adjust sample weights. Pothoff *et al.* (1993) show that, when the variable being measured is correlated (across individuals) with the contact probability, weighting based on the number of callbacks is practical and effective. In the end, the correct mix of these approaches will depend on the constraints facing the survey manager.

ACKNOWLEDGEMENTS

I thank John Eltinge, Mike Horrigan, Anne Polivka, Jim Spletzer, and Clyde Tucker for their comments and suggestions. The views expressed here are mine, and do not necessarily reflect those of the Bureau of Labor Statistics.

REFERENCES

BAUMAN, S.L., LAVRADAS, P.J. and MERKLE, D.M. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 1070-1075.

BOTMAN, S.L., MASSEY, J.D. and KALSBEEK, W.D. (1989). Cost-efficiency and the number of allowable callbacks in the national health interview survey. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 434-439.

HARVEY, A. (1993). Guidelines for time diary data collection. *Social Indicators Research.* 30, 197-228.

HARVEY, A. (1999). Guidelines for time use data collection and analysis. *Time Use Research in the Social Sciences,* (Eds. W.E. Pentland, A.S. Harvey, P. Lawton and M.A. McColl). New York: Kluwer Academic/Plenum Publishers, 19-45.

KALTON, G. (1985). Sample design issues in time diary studies. *Time, Goods, and Well-Being,* (Eds. F.T. Juster and F.P. Stafford). Ann Arbor: University of Michigan, Institute of Social Research, 333-351.

KINSLEY, B., and O'DONNELL, T. (1983). Marking time: methodology report of the Canadian time use pilot study–1981. *Explorations in Time Use* (vol. 1), Ottawa: Department of Communications, Employment and Immigration.

LAAKSONEN, S., and PÄÄKKÖNEN, H. (1992). Some methodological aspects on the use of time budget data. *Housework Time in Bulgaria and Finland*, (Eds. L. Kirjavainen, B. Anachkova, S. Laaksonen, I. Niemi, H. Pääkkönen and Z. Staikov). 86-104.

LYBERG, I. (1989). Sampling, nonresponse, and measurement issues in the 1984-85 Swedish time budget survey. *Proceedings of the Fifth Annual Research Conference*: Department of Commerce, Bureau of the Census, 210-238.

POTHOFF, R.F., MANTON, K.G. and WOODBURY, M.A. (1993). Correcting for nonavailability in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*. 88, 424, 1197-1207.

STATISTICS CANADA (1999). *Overview of the Time Use of Canadians in 1998*, General Social Survey, Catalogue No. 12F0080XIE; Ottawa, Canada.

STEWART, J. (2000). Alternative indexes for comparing activity profiles. Paper presented at the 2000 International Association for Time-Use Research Conference, Belo Horizonte, Brazil.

TRIPLETT, T. (1995). Data Collection Methods for Estimating Exposure to Pollutants Through Human Activity Pattern Data: A National Micro-behavioral Approach. mimeo, Survey Research Center, University of Maryland.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 169-181
Statistics Canada

Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples

ROBERT M. BELL and DANIEL F. MCCAFFREY[1]

ABSTRACT

Linearization (or Taylor series) methods are widely used to estimate standard errors for the coefficients of linear regression models fit to multi-stage samples. When the number of primary sampling units (PSUs) is large, linearization can produce accurate standard errors under quite general conditions. However, when the number of PSUs is small or a coefficient depends primarily on data from a small number of PSUs, linearization estimators can have large negative bias. In this paper, we characterize features of the design matrix that produce large bias in linearization standard errors for linear regression coefficients. We then propose a new method, bias reduced linearization (BRL), based on residuals adjusted to better approximate the covariance of the true errors. When the errors are i.i.d., the BRL estimator is unbiased for the variance. Furthermore, a simulation study shows that BRL can greatly reduce the bias even if the errors are not i.i.d. We also propose using a Satterthwaite approximation to determine the degrees of freedom of the reference distribution for tests and confidence intervals about linear combinations of coefficients based on the BRL estimator. We demonstrate that the jackknife estimator also tends to be biased in situations where linearization is biased. However, the jackknife's bias tends to be positive. Our bias reduced linearization estimator can be viewed as a compromise between the traditional linearization and jackknife estimators.

KEY WORDS: Complex samples; Linearization; Jackknife; Satterthwaite approximation; Degrees of Freedom.

1. INTRODUCTION

Regression analysis of multi-stage samples has become very common in recent years (for example, Ellickson and McGuigan 2000; Shapiro, Morton, McCafrrey, Senterfitt, Fleishman, Perlman, Athey, Keesey, Goldman, Berry and Bozzette 1999; Goldstein 1991; Landis, Lepkowski, Ekland and Stehouver 1982). Although hierarchical models (Bryk and Raudenbush 1992; Gelman, Carlin, Stern and Rubin 1995, Chapter 13) allow analysis of both fixed and random effects, many analysts prefer the simplicity of standard regression models when random effects are not of direct interest. Standard regression estimators produce unbiased parameter estimates that can be efficient, but the default standard error estimators do not account for the sample design, resulting in inconsistent standard errors (Kish 1965; Skinner 1989a). Various methods produce consistent standard error estimates applicable when the number of primary sampling units (PSUs) is sufficiently large. These include sample reuse methods such as the jackknife, bootstrap and balance repeated replication as well as linearization (or Taylor series) methods.

Linearization (Skinner 1989b) is a nonparametric method for estimating the standard errors of design-based statistics such as means and ratios as well as coefficients from linear and nonlinear regression models. By nonparametric, we mean that linearization does not rest on any assumptions about the within-PSU error structure, such as an assumption of constant intra-cluster correlation. When the number of PSUs can be considered large, linearization produces consistent standard errors in the presence of multiple features of complex sample designs-stratification, multi-stage sampling, and sampling weights-as well as heteroskedastic errors (Fuller 1975). Because of these desirable properties and its increased availability in software such as SUDAAN, Stata, and SAS Version 8.0 (Shah, Barnwell, and Bieler 1997; StataCorp. 1999; SAS Institute, Inc. 1999), linearization has become a common method for estimating standard errors and confidence intervals and for conducting statistical tests on data from complex sample designs (for example, Ellickson and McGuigan 2000; Shapiro *et al.* 1999; Rust and Rao 1996). Linearization has also been proposed for estimating standard errors from Generalized Estimating Equations (GEE) fit to multi-stage data (Zeger and Liang 1986).

However, the linearization method has limitations. When the number of primary sampling units is small, standard error estimates can be severely biased low, they can have large coefficients of variation, and the standard degrees of freedom may be far too liberal (Kott 1994; Murray, Hannan, Wolfinger, Baker and Dwyer 1998). Consequently, standard linearization inference for coefficients based mainly on data from a small number of PSUs may produce confidence intervals that are too narrow and tests with Type I error rates that are substantially higher than their nominal values. Sample reuse methods like the jackknife have similar limitations.

In this paper, we characterize the design factors (*i.e.*, the distribution of explanatory variables within and between PSUs) that produce large bias in linearization and jackknife

[1] Robert M. Bell, Statistics Research Department, AT&T Labs-Research, Room C211, 180 Park Ave., Florham Park, NJ 07932; Daniel F. McCaffrey, Statistics Group, RAND, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516.

standard errors for linear regression coefficients and demonstrate that the problem can persist even when the number of PSUs is quite large. We then propose an alternative to the standard linearization estimator that is unbiased for independent, identically distributed (i.i.d.) errors and tends to greatly reduce bias otherwise. We also present approximate degrees of freedom for use with tests and confidence intervals based on our variance estimator. Simulation results show improved small sample properties of our alternative estimator and test compared with those of more traditional methods. Finally, we present an example of our methods using data from a national experiment evaluating care for depression.

2. BIAS OF THE LINEARIZATION METHOD

For simplicity, we restrict consideration in the body of this paper to unweighted linear regression for two-stage nonstratified samples. Extensions to weighted estimators and stratified samples are presented in McCaffrey, Bell and Botts (2001) and discussed further in section 8.

Let n equal the number of PSUs and m_i equal the number of final sampling units from the i-th PSU, for $i = 1, ..., n$. The overall sample size is $M = \sum_i m_i$. We assume that $y_{ij} = \beta' x_{ij} + \varepsilon_{ij}$, where ε has mean 0 and covariance matrix $\mathbf{V}$, and where y_{ij}, x_{ij}, and ε_{ij} all refer to the j-th observation from the i-th PSU. We drop the standard OLS assumption of i.i.d. errors, assuming only that errors from distinct PSUs are uncorrelated. Specifically, we assume that $\mathbf{V}$ is block diagonal, with $m_i \times m_i$ blocks $\mathbf{V}_i$ for $i = 1, ..., n$. In addition to the notation of this model, throughout the paper, we let $\mathbf{I}$ denote an $M \times M$ identity matrix and $\mathbf{I}_i$ equal an $m_i \times m_i$ identity matrix.

Let $\hat{\beta}$ denote the estimated coefficients of the linear regression model. To simplify presentation, we generally discuss a linear combination of the regression coefficients, $l' \hat{\beta}$, for an arbitrary column vector l. For the special case where one element of $l = 1$ and the rest are 0, $l' \hat{\beta}$ equals a single estimated coefficient. If errors are uncorrelated across PSUs, the variance of $l' \hat{\beta}$, is

$$\mathrm{Var}(l'\hat{\beta}) = l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l, \quad (1)$$

where $\mathbf{X}$ and $\mathbf{X}_i$ are the design matrices for the entire sample and for PSU i, respectively.

The standard linearization estimator of the variance of $l' \hat{\beta}$ is given by:

$$v_L = l'(\mathbf{X}'\mathbf{X})^{-1} \left(c \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{r}_i \mathbf{r}_i' \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \quad (2)$$

where $\mathbf{r}_i$ is the vector of residuals for the i-th PSU. Comparison of (1) and (2) shows that linearization simply involves estimating $\mathbf{V}_i$ by a constant c times the outer product of the residuals. The constant c is typically set equal to $n/(n-1)$, the value used by SUDAAN and the Stata svy procedures (Shah, Barnwell, and Bieler 1997; StataCorp. 1999). For GEE procedures, Zeger and Liang (1986) set $c = 1$.

Under fairly general conditions, nv_L converges in probability to the variance of the asymptotic distribution of $\sqrt{n}\,(l'\hat{\beta} - l'\beta)$ and the relative bias of v_L is O(1/n) as the number of PSUs gets large (Fuller 1975; Kott 1994). To demonstrate convergence for the bias of v_L, Kott (1994) assumes that the number of observations from every PSU is bounded and that elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are bounded by B/n for a constant B. These assumptions effectively ensure that the influence of any PSU on the final estimate diminishes as the number of PSUs grows. Convergence of the bias of v_L holds for heteroskedastic data from stratified samples with unequal sampling weights and arbitrary correlation structure within PSUs. Unfortunately, consistency does not guarantee good properties for small to moderate numbers of PSUs.

Theorem 1. When $\mathbf{V} = \sigma^2 \mathbf{I}$ and $c = n/(n-1), E(v_L) \leq \mathrm{Var}(l'\hat{\beta})$ with equality if and only if $l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{X}_i$ is constant across i.

Proof. Without loss of generality, we assume that $\sigma^2 = 1$ so that $\mathbf{V} = \mathbf{I}$. The residual vector $\mathbf{r}$ can be written as $(\mathbf{I} - \mathbf{H})\varepsilon$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat or projection matrix for $\mathbf{X}$. Thus, we have that $\mathbf{r}_i = (\mathbf{I} - \mathbf{H})_i \varepsilon$, where $(\mathbf{I} - \mathbf{H})_i$ contains the m_i rows of $(\mathbf{I} - \mathbf{H})$ for the i-th PSU. Consequently,

$$\begin{aligned} E(v_L) &= \left(\frac{n}{n-1} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad \left(\sum_{i=1}^{n} \mathbf{X}_i' (\mathbf{I} - \mathbf{H})_i E(\varepsilon\varepsilon') (\mathbf{I} - \mathbf{H})_i' \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l. \\ &= \left(\frac{n}{n-1} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad \sum_{i=1}^{n} \left(\mathbf{X}_i' \mathbf{X}_i - \mathbf{X}_i' \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i' \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \quad (3) \end{aligned}$$

because $E(\varepsilon\varepsilon') = I$ and $(\mathbf{I} - \mathbf{H})_i (\mathbf{I} - \mathbf{H})_i' = (\mathbf{I}_i - \mathbf{H}_{ii})$ for $\mathbf{H}_{ii} = \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i'$. Let $\mathbf{D}_i = \mathbf{X}_i' \mathbf{X}_i - (1/n)(\mathbf{X}'\mathbf{X})$. Note that $\sum_i \mathbf{D}_i = \sum_i \mathbf{X}_i' \mathbf{X}_i - \mathbf{X}'\mathbf{X} = 0$. Thus,

$$E(v_L) = \left(\frac{n}{n-1}\right) l'(\mathbf{X}'\mathbf{X})^{-1}$$

$$\sum_{i=1}^{n} \left(\mathbf{X}_i'\mathbf{X}_i - [(1/n)\mathbf{X}'\mathbf{X} + \mathbf{D}_i](\mathbf{X}'\mathbf{X})^{-1}[(1/n)\mathbf{X}'\mathbf{X} + \mathbf{D}_i]\right)$$

$$(\mathbf{X}'\mathbf{X})^{-1} l$$

$$= \left(\frac{n}{n-1}\right) l'(\mathbf{X}'\mathbf{X})^{-1}$$

$$\left(\mathbf{X}'\mathbf{X} - (1/n)\mathbf{X}'\mathbf{X} - \sum_{i=1}^{n} \mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_i\right)(\mathbf{X}'\mathbf{X})^{-1} l$$

$$= l'(\mathbf{X}'\mathbf{X})^{-1} l - \left(\frac{n}{n-1}\right) l'(\mathbf{X}'\mathbf{X})^{-1}$$

$$\left(\sum_{i=1}^{n} \mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_i\right)(\mathbf{X}'\mathbf{X})^{-1} l$$

$$= \mathrm{Var}(l'\hat{\beta}) - \left(\frac{n}{n-1}\right)\left(\sum_{i=1}^{n} a_i' (\mathbf{X}'\mathbf{X})^{-1} a_i\right) \tag{4}$$

for $a_i = \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1} l = [\mathbf{X}_i' \mathbf{X}_i - (1/n)(\mathbf{X}'\mathbf{X})] (\mathbf{X}'\mathbf{X})^{-1} l$. Because $(\mathbf{X}'\mathbf{X})^{-1}$ is positive definite, $E(v_L) \le \mathrm{Var}(l'\hat{\beta})$ with equality if and only if $a_i \equiv 0$, or equivalently, $\mathbf{X}_i'\mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ is constant across the i.

Replication methods do not necessarily avoid the problem of bias for regression variance estimators. A jackknife estimator for multi-stage samples can be derived from the set of pseudo values $\{\beta_{[i]}\}$, estimates of β from data that exclude the i-th PSU:

$$v_{JK} = [(n-1)/n] \sum\nolimits_i l'\left(\tilde{\beta}_{[i]} - \hat{\beta}\right)\left(\tilde{\beta}_{[i]} - \hat{\beta}\right)' l \tag{5}$$

(Cochran 1977; Rust and Rao 1996). If $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i, then

$$v_{JK} = [(n-1)/n] l' (\mathbf{X}'\mathbf{X})^{-1} \sum\nolimits_i \mathbf{X}_i' (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$$

$$\mathbf{r}_i \mathbf{r}_i' (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l, \tag{6}$$

which follows from the updating formula $(\mathbf{X}'\mathbf{X} - \mathbf{X}_i'\mathbf{X}_i)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i' (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1}$ (Cook and Weisberg 1982; Bell and McCaffrey 2002, page 34). Some authors (Efron and Tibshirani 1993) suggest an alternative jackknife estimator with $\hat{\beta}$ replaced by the mean of the $\tilde{\beta}_{[i]}$'s in (5). These two methods provide very similar estimates in our simulations, so we discuss only the version based on (5) in what follows.

Theorem 2. When $\mathbf{V} = \sigma^2 \mathbf{I}$ and $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i, then $E(v_{JK}) \ge \mathrm{Var}(l'\hat{\beta})$ with equality if and only if $l' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i' \mathbf{X}_i$ is constant across i (proof in appendix).

The following example shows that the conditions for linearization and the jackknife estimators to be unbiased are very restrictive even for simple linear regression.

Example 1. Consider simple linear regression. We have that

$$\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1} l = \frac{m_i}{Ms^2}\begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & s_i^2 + \bar{x}_i^2 \end{bmatrix}\begin{bmatrix} s^2 + x^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} l$$

where s^2 and $\{s_i^2\}$ are ML estimates for the overall and within-PSU variances of x, with divisors M and $\{m_i\}$, respectively. So we have

$$\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1} l =$$

$$\frac{m_i}{Ms^2}\begin{bmatrix} s^2 + \bar{x}^2 - \bar{x}_i\,\bar{x} & \bar{x}_i - \bar{x} \\ (s^2 + \bar{x}^2)\bar{x}_i - (s_i^2 + \bar{x}_i^2)\bar{x} & s_i^2 + \bar{x}_i^2 - \bar{x}_i\,\bar{x} \end{bmatrix} l.$$

To have v_L and v_{JK} unbiased for the slope, *i.e.*, for $l' = (0,1)$, we must have that $m_i(\bar{x}_i - \bar{x})$ and $m_i(s_i^2 + \bar{x}_i^2 - \bar{x}_i\,\bar{x})$ are both constant across i. The former implies that $\bar{x}_i \equiv \bar{x}$, and together they imply that $m_i s_i^2 = \sum_j (x_{ij} - \bar{x})^2$ is constant. Note that m_i need not be constant. These two conditions are not sufficient to guarantee unbiasedness for $l' = (0,1)$, however. Additional algebra shows that the bias in the linearization estimator for the variance of the slope equals

$$-\frac{n}{(n-1)M^3 s^4}\left\{\sum_{i+1}^{n} [m_i(\bar{x}_i - \bar{x})]^2 + \sum_{i=1}^{n}\left[\sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 - \bar{m}s^2\right]^2\right\}.$$

Consequently, the bias includes a part that is proportional to the weighted variance of the PSU means of x and another that is proportional to the variance of the within-PSU sums of squares.

The example shows that when the errors are i.i.d., v_L is unbiased only under very restrictive conditions. When $\mathbf{V} \ne \mathbf{I}$, Theorems 1 and 2 do not hold, and the bias in v_L can even be positive (see Example 2 of Bell and McCaffrey 2002).

In general, v_L tends to have negative bias. The estimator is the sum over PSUs of squares of linear combinations of residuals, $c^{1/2} l' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i' \mathbf{r}_i$. These sums of squares tend to be too small for two reasons: residuals are generally smaller than true errors due to overfitting, and residuals tend to have lower intra-cluster correlation than the errors. The factor $c = n/(n-1)$ corrects completely for these problems only in very restricted circumstances like the conditions in Theorem 1.

The bias of the linearization estimator (or the jackknife) increases with the between-PSU variance of the explanatory variables. Consequently, explanatory variables that are (nearly) constant within PSUs tend to exhibit the largest bias. When there are several such explanatory variables, there can be substantial underestimation of intra-cluster correlations, leading to large bias in estimated variances for all the corresponding coefficients. Even greater bias potential appears to occur when certain PSUs account for most of the variability in the covariates and have disproportionate impact on the determination of $l'\hat{\beta}$.

3. THE BIAS REDUCED LINEARIZATION METHOD

Phillip Kott has proposed two methods for reducing the bias in linearization. Kott (1994) suggested correcting the bias in v_L by using the residuals and the design matrix to estimate the negative of the bias of v_L by $\hat{R}$ ($\hat{R} > 0$, typically) and setting $v_{K94} = v_L/(1 - \hat{R}/v_L)$. Kott suggested the estimator v_{K94} rather than the more obvious $(v_L + \hat{R})$ as *ad hoc* compensation for the relative bias in $\hat{R}$ as an estimator of the true negative bias, R.

In his 1996 paper, Kott suggests calculating the ratio of $\mathrm{Var}(l'\hat{\beta})$ to $E(v_L)$ under the assumption that $\mathbf{V} = \mathbf{I}$ and adjusting v_L by the ratio. If $\mathbf{V} = \mathbf{I}$ then the resulting estimator v_{K96} will be unbiased.

In the context of generalized estimating equations, Mancl and DeRouen (2001) take a different approach to correcting the bias in the linearization estimator. They suggest adjusting the residuals from each PSU to reduce the bias in $\mathbf{r}_i\mathbf{r}_i'$ as an estimator of $\mathbf{V}_i$. For the unweighted linear model given in section 2, they approximate $E(\mathbf{r}_i\mathbf{r}_i')$ by $(\mathbf{I}_i - \mathbf{H}_{ii})\mathbf{V}_i(\mathbf{I}_i - \mathbf{H}_{ii})$ and suggest replacing $\mathbf{r}_i$ in $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}\mathbf{r}_i$ equation (2). Thus, for unweighted linear models the Mancl and DeRouen estimator equals $n/(n-1)v_{JK}$ and the properties on this estimator follow from the properties of the jackknife estimator.

We present an alternative approach that we first proposed in 1997 (McCaffrey and Bell 1997). The method is also based on replacing $\mathbf{r}_i$ in equation (2) with adjusted residuals of the form $\mathbf{r}_i^* = \mathbf{A}_i\mathbf{r}_i$ intended to act more like the true errors ε_i. Like Kott (1996), we derive an estimator that eliminates the bias of v_L when $\mathbf{V}$ equals $\mathbf{U}$, a specified block-diagonal covariance matrix, and reduces the bias for other $\mathbf{V}$. Like Mancl and DeRouen (2001) we adjust the residuals from each PSU. However, using $\mathbf{U}$ we derive an alternative approximation to the $E(\mathbf{r}_i\mathbf{r}_i')$ and our resulting estimator is not proportional to the jackknife but rather can be seen as a compromise between the linearization and jackknife estimators. Our approach is also a generalization of the method of MacKinnon and White (1985), who adjust individual residuals to produce a heteroskedastically-consistent variance estimator (in the sense of White 1980) that is unbiased when the errors are independent and homoskedastic.

Theorem 3. For a specified block-diagonal covariance matrix $\mathbf{U}$, consider the class of estimators $v_{L^{\cdot}} = l'(\mathbf{X}'\mathbf{X})^{-1}$ $(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{A}_i'\mathbf{X}_i)$ $(\mathbf{X}'\mathbf{X})^{-1}l$, where $\mathbf{A}_i$, satisfies $\mathbf{A}_i[(\mathbf{I}-\mathbf{H})_i\mathbf{U}(\mathbf{I}-\mathbf{H})_i']\mathbf{A}_i' = \mathbf{U}_i$ for $i = 1, \ldots, n$. If $\mathbf{V} = k\mathbf{U}$ for some scalar k, then $E(v_{L^{\cdot}}) = \mathrm{Var}(l'\hat{\beta})$.

Proof. The expected value of $v_{L^{\cdot}}$ is given by

$$E(v_{L^{\cdot}})$$

$$= l'(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{A}_i(\mathbf{I}-\mathbf{H})_i(k\mathbf{U})((\mathbf{I}-\mathbf{H})_i'\mathbf{A}_i'\mathbf{X}_i\right)$$

$$(\mathbf{X}'\mathbf{X})^{-1}l$$

$$= l'(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'(k\mathbf{U}_i)\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1}l = \mathrm{Var}(l'\hat{\beta}).$$

Without external evidence to the contrary, an analyst is likely to use a working covariance matrix of the form $\mathbf{U} = \sigma^2\mathbf{I}$, which simplifies the condition on $\mathbf{A}_i$ to $\mathbf{A}_i(\mathbf{I}_i - \mathbf{H}_{ii})\mathbf{A}_i' = \mathbf{I}_i$ or

$$\mathbf{A}_i'\mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}. \tag{7}$$

We set $\mathbf{U} = \mathbf{I}$ in what follows.

A solution to equation (7) exists for PSU i whenever $(\mathbf{I}_i - \mathbf{H}_{ii})$ is full rank, which is true if all the eigenvalues of $\mathbf{H}_{ii}$ are strictly less than 1 (the eigenvalues of $\mathbf{H}_{ii}$ are always between 0 and 1). An eigenvalue of $\mathbf{H}_{ii}$ may equal 1 – *e.g.*, when the model includes a dichotomous explanatory variable that is one if and only if an observation falls in the i-th PSU.

For $m_i > 1, \mathbf{A}_i$ is not unique. If $\mathbf{A}_i$ satisfies $\mathbf{A}_i'\mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, then so does $\mathbf{O}\mathbf{A}_i$, for any $m_i \times m_i$ orthogonal matrix $\mathbf{O}$. If $\mathbf{V} = \sigma^2\mathbf{I}$, the choice of $\mathbf{A}_i$ is unimportant because any solution to (7) will produce an unbiased variance estimator. However, the resulting estimators are biased when $\mathbf{V} \neq \sigma^2\mathbf{I}$, and the bias can vary greatly with the choice of $\mathbf{A}_i$. Heuristically, it makes sense to choose the solution $\mathbf{A}_i$ "closest" to the identity matrix, so as to "mix" the residuals as little as possible. Two promising candidates are the Cholesky decomposition of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, which has all 0's below the diagonal, and the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$. Let $\mathbf{P}$ be an orthogonal matrix whose columns are the eigenvectors of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ and $\mathbf{\Lambda}$ be a diagonal matrix containing the corresponding eigenvalues of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, so that $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$. Then for $\mathbf{\Lambda}^{1/2}$ equal to the elementwise square root of $\mathbf{\Lambda}$, $\mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}'$ is symmetric and solves (7). In contrast, multiplying either of these two solutions by a random orthogonal matrix could greatly distort the residuals.

Among the class of adjusted residuals of the form $\mathbf{A}_i\mathbf{r}_i$ where $\mathbf{A}_i$ satisfies (7), those based on the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, $\mathbf{r}_i^* = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}'\mathbf{r}_i$, are "best" in the sense of Theil (1971) – *i.e.*, they minimize the expected sum of the squared differences between the estimated and true i.i.d. errors (see pages 36-37 of Bell and McCaffrey 2002 for details). When there is intra-cluster correlation, simulation results in section 6 suggest that the bias of v_L^* based on the symmetric square root is greatly reduced compared with that of the traditional linearization estimator, v_L. For these reasons, we consider only the symmetric root in the remainder of the paper and refer to the estimator using this root as the biased reduced linearization estimator, v_{BRL}.

As Kott (1994) proved for v_L, if the number of units in every PSU is bounded and the elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are bounded by B/n for some constant B (*i.e.*, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = O(1/n)$), then the bias in v_{BRL} is $O(n^{-2})$ and the relative bias is $O(1/n)$ (Bell and McCaffrey 2002, page 15).

4. VARIANCE OF THE ESTIMATORS AND TESTING

We note that v_L, v_{BRL}, and v_{JK} can all be written in the form

$$v^* = cl'(\mathbf{X}'\mathbf{X})^{-1}\textstyle\sum_i \mathbf{X}_i{}'\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i{}'\mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l,$$

where: $c = n/(n-1)$, 1, or $(n-1)/n$, respectively, and $\mathbf{A}_i = \mathbf{I}_i, (\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}$, or $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, respectively. This formulation of the estimators shows that v_{BRL} can be viewed as a compromise between v_L and v_{JK}, chosen to offset their opposing biases.

Theorem 4. Let the error terms be distributed as multivariate normal with mean **0** and nonsingular covariance matrix **V**. Then for any variance estimator of the form

$$v^* = cl'(\mathbf{X}'\mathbf{X})^{-1}\textstyle\sum_i \mathbf{X}_i{}'\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i{}'\mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l,$$

v^* equals the weighted sum of independent χ_1^2 random variables where the weights are the eigenvalues of the $n \times n$ matrix for $\mathbf{G} = \{\mathbf{g}_i'\mathbf{V}\mathbf{g}_j\}$, for $\mathbf{g}_i = c^{1/2}(\mathbf{I}-\mathbf{H})_i'$ $\mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ (proof in appendix).

We can write v_L as a quadratic form $\mathbf{y}'\mathbf{G}^*\mathbf{y}$, where the M-by-M matrix $\mathbf{G}^* = \sum_{i=1}^n \mathbf{g}_i\mathbf{g}_i'$, so that v_L is a weighted sum of independent chi-square random variables with weights equal to the eigenvalues of $\mathbf{G}^*\mathbf{V}$. The proof consists of showing that the nonzero eigenvalues of $\mathbf{G}^*\mathbf{V}$ equal the nonzero eigenvalues of **G**.

The mean and variance of v^* are simple functions of the eigenvalues of **G**, namely $E(v^*) = \sum_{i=1}^n \lambda_i E(u_i^2) = \sum_{i=1}^n \lambda_i$ and $\mathrm{Var}(v^*) = \sum_{i=1}^n \lambda_i^2 \mathrm{Var}(u_i^2) = \sum_{i=1}^n 2\lambda_i^2$. If $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ for $i = 1,\ldots,n$ are constant, conditions for v_L and v_{JK} to be unbiased, then Theorem 4 implies that av_L, av_{JK}, and av_{BRL} are all distributed χ_{n-1}^2 for $a = (n-1)/\mathrm{Var}(l'\hat{\beta})$ (Bell and McCaffrey 2002, pages 41-42). However, in general, the $\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ will not be constant and the squared coefficient of variation will exceed $2/(n-1)$, the corresponding statistic for a χ_{n-1}^2 random variable.

This excess variability is of particular concern when considering reference distributions for testing the null hypothesis that $l'\beta = 0$, with test statistics of the form $t = l'\hat{\beta}/\sqrt{v^*}$. For v_L, Shah, Holt and Folsom (1977) suggested comparing t to a reference t-distribution with $n-1$ degrees of freedom, which is now the default in Stata (Stata Corp. 1999), SUDAAN (Shah, Barnwell and Bieler 1997) and SAS (SAS Institute 1999). The choice of $n-1$ degrees of freedom is motivated by the fact that v_L can be written as the sum of squares of n random variables $c^{1/2}l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{r}_i$. However, because the variance of $(n-1)v_L/E(v_L)$ tends to be greater than $2(n-1)$, tests that use a t-distribution with $n-1$ degrees of freedom would tend to have Type I error rates that exceed the nominal value, even if v_L were unbiased.

Satterthwaite (1946) suggested approximating the distribution of a linear combination of χ_1^2 variables by χ_f^2 (up to a constant) where the first two moments of the linear combination match those of χ_f^2. We would approximate v_L, v_{BRL} or v_{JK} by a χ_f^2 where $f = 2/cv^2 = (\sum_{i=1}^n \lambda_i)^2/\sum_{i=1}^n \lambda_i^2$ and the λ_i are the eigenvalues of the corresponding matrix **G**. Tests based on reference t-distributions with f degrees of freedom would be expected to provide better Type I error rates than tests based on $n-1$ degrees of freedom. Rust and Rao (1996) also suggest using a Satterthwaite approximation to estimate the degrees of freedom for the jackknife estimator. They present results for the estimator of a mean, while Theorem 4 extends this approach to testing linear combinations of regression coefficients. Kott (1994, 1996) suggests using the Satterthwaite approximation to estimate the degrees of freedom for tests based on his alternatives to linearization.

The coefficient of variation for any of the nonparametric variance estimators can be very large for certain designs. High variability occurs under the same conditions that v_L and v_{JK} are most biased – when residuals from only a few PSUs effectively determine the final variance estimate. This variability of the estimators is an inherent cost of using nonparametric techniques.

Because the Satterthwaite degrees of freedom f requires specifying the unknown matrix **V**, we have investigated two methods for setting **V**. The first treats **V** as block-diagonal and estimates each block with the outer-product of the residuals for the PSU. Because preliminary simulation results showed that degrees of freedom based on this empirical estimate of **V** produced tests that were extremely conservative, we do not present any simulation results for this method. Kott (1994) also found that estimating **V** for use in the formula for estimated degrees of freedom proved unsatisfactory. Instead, we used a second method that sets **V** identically equal to the identity matrix – *i.e.*, it assumes independent, homoskedastic errors for purposes of determining degrees of freedom.

The distribution of v_{BRL} (and the other variance estimators) tends to be less skewed and have less mass in the lower tail than the distribution of a χ_f^2 where f equals the Satterthwaite degrees of freedom. Hence, reference t-distributions based on the Satterthwaite approximation tend to overestimate tail probabilities. For example, when data from a couple of PSUs nearly determine the value of a coefficient, the Satterthwaite degrees of freedom can be less than two, incorrectly implying a chi-square density that is infinite at zero. Consequently, the probability of very large t-statistics may not be as large as the Satterthwaite approximation would imply, especially when the Satterthwaite degrees of freedom are less than 4 or 5.

5. SIMULATION METHODS

We use a Monte Carlo simulation to study the properties of alternative variance estimators and tests for a balanced two-stage cluster sample with $n = 20$ PSUs and a constant $m = 10$ observations in each PSU. All simulation replications use a common design matrix **X** with four explanatory variables chosen to represent a range of difficulty for nonparametric variance estimators. The first two explanatory variables, x_1 and x_2, are dichotomous (0 or 1) and constant within PSU. The variable x_1 is 1 in half the clusters: 1, 3,...,19, while x_2 is 1 in just three clusters: 9, 10, and 11. Both x_3 and x_4 were generated from standard normal distributions. They differ in that x_3 was generated from a multivariate normal with intra-cluster correlation of 0.5 within PSU, while x_4 was generated from independent normal distributions. Observed intra-cluster correlations are 1.00, 1.00, 0.62 and -0.04, respectively. Observed correlations among the explanatory variables are all very small with the exception of $\mathrm{Corr}(x_1, x_2) = 0.14$, $\mathrm{Corr}(x_1, x_3) = 0.25$ and $\mathrm{Corr}(x_1, x_4) = -0.11$. The estimated regression coefficients are linear combinations of the dependent variable with multipliers given by the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which are shown in Figure 1. For the first three coefficients, and to a lesser extent $\hat{\beta}_3$, observations from the same PSU tend to have similar multipliers. Of more importance, $\hat{\beta}_2, \hat{\beta}_0$, and $\hat{\beta}_3$ are determined primarily by results in a small number of PSUs with relatively large multipliers (in absolute value). For example, Figure 1 shows that the multipliers for $\hat{\beta}_3$ are large for the second PSU, which has a mean that is over two standard deviations from the average PSU mean. In general, variance in the PSU means gives some PSUs greater weight for estimating $\hat{\beta}_3$.

The dependent variable was generated from the equation $y_{ij} = \beta' x_{ij} + \varepsilon_{ij}$, where $\beta = 0$ and the ε_i's are standard multivariate normal random variables with intra-cluster correlation ρ. We use three alternative values of $\rho = 0, 1/9$, and 1/3, corresponding to design effects for the sample mean of DEFF = 1, 2, and 4, respectively (DEFF$= 1 + (m-1)\rho$). Monte Carlo results are based on 100,000 replications of **y** for our fixed **X**.

We evaluated the ordinary least squares (OLS) variance estimator, $s^2 l'(\mathbf{X}'\mathbf{X})^{-1} l$, and five nonparametric variance estimators: the standard linearization estimator given in equation (2) with $c = n/(n-1)$; the jackknife estimator given in (5); bias reduced linearization; and Kott's two adjustments to linearization. BRL and the Kott adjustments are all based on working intra-cluster correlations of $\rho = 0$.

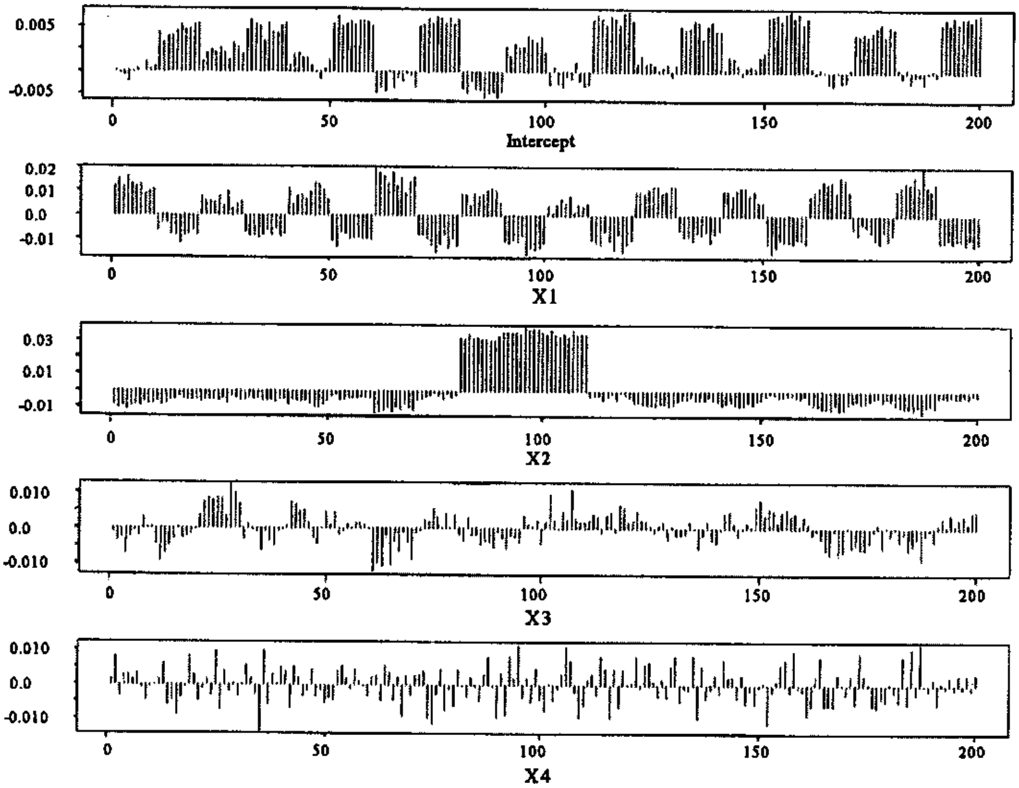

Figure 1. Values of the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for the design matrix used in simulations

We estimated Type I error rates for eight alternative test procedures based on 100,000 replications from the null hypothesis where each $\hat{\beta}_k = 0$, for $k = 0$ to 4. Each procedure compares a "t-statistic" against a reference t-distribution. For the t's based on linearization, the jackknife, and BRL, we use critical values from t-distributions with both $(n-1) = 19$ degrees of freedom and the corresponding Satterthwaite approximation. For Kott's methods, we use his proposed degrees of freedom. All computations were implemented in SAS.

6. SIMULATION RESULTS

Table 1 shows the bias of several variance estimators for the five regression coefficients (including the intercept) for $\rho = 0, 1/9$, and 1/3. Except for Kott (1994), all values are exact based on the X matrix described above. Because Kott (1994) cannot be written as a linear functional, its bias is estimated from the Monte Carlo simulations, and the standard error of the bias is shown in parentheses.

Table 1
Bias of Variance Estimators (as a Percentage of the True Variance)

Estimator	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
			$\rho = 0$		
OLS	0.0	0.0	0.0	0.0	0.0
Linearization	-9.6	-13.2	-32.5	-13.3	-1.8
Jackknife	11.7	17.2	51.2	17.6	2.1
Kott (1994)	4.0	2.5	-1.0	2.2	4.7
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	0.0	0.0	0.0	0.0	0.0
BRL	0.0	0.0	0.0	0.0	0.0
			$\rho = 1/9$		
OLS	-50.2	-49.7	-50.7	-37.7	4.1
Linearization	-10.3	-14.2	-33.2	-17.1	-2.5
Jackknife	11.0	16.4	50.1	19.8	3.2
Kott (1994)	3.9	2.7	-0.8	1.5	4.6
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-0.8	-1.2	-1.0	-4.4	-0.7
BRL	-0.7	-1.0	-0.8	-1.2	0.1
			$\rho = 1/3$		
OLS	-75.8	-75.5	-76.2	-65.3	13.8
Linearization	-10.7	-14.8	-33.5	-19.9	-4.1
Jackknife	10.7	15.9	49.5	21.4	5.9
Kott (1994)	3.6	2.4	-0.6	1.4	4.4
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-1.2	-1.9	-1.5	-7.7	-2.3
BRL	-1.0	-1.5	-1.3	-2.1	0.4

Note: All values are exact except for Kott (1994), which is based on 100,000 simulation replications.

The OLS variances are unbiased for $\rho = 0$, but they are badly biased for $\rho = 1/9$ and 1/3. As discussed in Wu, Holt and Holmes (1988), the OLS variances are too small by roughly a factor of $1/[1 + \rho(m-1)ICC_x]$, where ICC_x denotes the intra-cluster correlation for an x variable. Hence, for PSU-level variables (including the intercept), the OLS variances are too small by roughly a factor of 1/DEFF. Similarly, the bias is smaller, but still substantial for x_3, the individual-level variable with large intra-cluster correlation. The positive bias for the OLS variance of $\hat{\beta}_4$ results from the slight negative intra-cluster correlation for x_4.

Linearization and the jackknife each suffer from large biases, relatively independent of ρ, but the biases point in opposite directions. For each estimator, the magnitude of the bias varies greatly among the coefficients. The largest biases (in absolute value) occur for $\hat{\beta}_2$, which depends mainly on the data from three PSUs. The next greatest biases occur for $\hat{\beta}_3$, followed closely by $\hat{\beta}_1$ and $\hat{\beta}_0$.

Except for $\hat{\beta}_4$, Kott (1994) has much smaller magnitude bias than linearization. However, the method tends to overcompensate, often resulting in notable positive bias. An exception is $\hat{\beta}_2$, for which Kott's estimator remains biased low.

By design, Kott (1996) and BRL eliminate the bias for $\rho = 0$. Consequently, choice among these alternatives should rest mainly on how well they hold down bias for $\mathbf{V} \neq \mathbf{I}$. Both methods reduce the magnitude of bias dramatically relative to linearization for $\rho = 1/9$ and 1/3. Although differences between the two methods are often small, BRL does uniformly better, with its worst bias being -2.1 percent. While Kott (1996) is practically indistinguishable from BRL for the PSU-level variables, it performs substantially worse for $\hat{\beta}_3$ and $\hat{\beta}_4$.

The linearization, jackknife, BRL and Kott estimators are highly correlated with similar coefficients of variation. For any given regression coefficient, the correlation among the variance estimators always exceeded 0.969, with most exceeding 0.99 (not shown). The smallest correlations tended to be between the jackknife and other estimators. The coefficients of variation (also not shown) were largest for Kott (1994) and tended to be smallest for linearization and Kott (1996) (except for the intercept). For the intercept, the jackknife had the smallest coefficient of variation. The relative variance of the BRL estimator was similar to that of the alternative nonparametric methods. Its coefficient of variation was between 1 and 6 percent larger than that of the linearization estimator but about 5 to 10 percent smaller than that of Kott (1994). Thus, the five nonparametric variance estimators tend to differ from each other mainly by constant factors, and Table 1 summarizes the main difference among these variance estimators.

Table 2 shows the Satterthwaite degrees of freedom for each of the five coefficients for the linearization, jackknife, BRL and Kott variance estimators. For all estimators the degrees of freedom were calculated assuming $\mathbf{V} = \mathbf{I}$ and consequently depend only on the design matrix and not on the values of $\mathbf{y}$. The approximations are similar for linearization and BRL although the linearization degrees of freedom tend to be slightly larger reflecting the fact that for this design matrix the relative variances of the BRL estimators are marginally larger than those for linearization. Kott's approximation derives the coefficient of variation for a linearization-type estimator based on the true errors rather than the residuals. As a result, Kott's approximate degrees of freedom, which are larger than those for linearization or BRL, tend to overstate the precision of his estimator (see Kott 1994, section 6). Across all four estimators, the approximations are smallest for $\hat{\beta}_2$.

Table 2
Degrees-of-Freedom for Selected Estimators

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Satterthwaite (LIN)	9.02	14.45	3.30	11.56	16.65
Satterthwaite (Jackknife)	9.52	13.30	2.62	9.06	16.23
Satterhwaite (BRL)	9.24	14.08	2.90	10.26	16.45
Kott's method	10.33	16.41	4.32	11.36	17.44

Table 3 shows that Type I error rates for the standard linearization method with $(n-1)$ degrees of freedom consistently exceed 5 percent for all three values of ρ. Type I errors are most common for $\hat{\beta}_2$, where they reach as high as 16 percent, but they also occur much too frequently for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_3$, ranging from 7.0 to 8.8 percent. The magnitude of this problem correlates closely with the size of the bias of the linearization estimator (see Table 1). Type I error rates are much lower, 5.7 to 6.4 percent, for tests based on the Satterthwaite degrees of freedom. Thus using the alternative degrees of freedom improved the Type I error rates by about 30 to 88 percent.

There is a less consistent pattern for the Type I error probabilities for the jackknife. The jackknife with $(n-1)$ degrees of freedom tends to be conservative for $\hat{\beta}_1$ and $\hat{\beta}_3$, in accord with the positive bias in the jackknife variance. In contrast, the probability of Type I error is much too large for $\hat{\beta}_2$, and a bit too large in two of three cases for the intercept $\hat{\beta}_0$. The apparent explanation is that the choice of $(n-1)$ as the degrees of freedom for the reference t-distribution sometimes counteracts the bias in the jackknife variance. This conclusion is supported by the very low Type I error rates for the jackknife with Satterthwaite degrees of freedom; smaller degrees of freedom combined with large positive biases result in very conservative tests.

BRL with $(n-1)$ degrees of freedom improves substantially on linearization with the same degrees of freedom. Because BRL is unbiased when $\rho = 0$, comparing the fifth row of the table against the first demonstrates the reduction in Type I errors that results from removing the bias of linearization. Excluding $\hat{\beta}_4$, BRL reduces Type I error rates by about 45 to 88 percent. However, BRL with $(n-1)$ degrees of freedom remains consistently liberal, especially for $\hat{\beta}_2$. Comparison of rows 2 and 5 of each section shows the relative impact of bias reduction and the Satterthwaite adjustment. For $\hat{\beta}_0$ and $\hat{\beta}_2$, degrees of freedom are more important, while bias matters more for $\hat{\beta}_1$ and $\hat{\beta}_3$. Performance for BRL with the Satterthwaite approximation is very good, except for $\hat{\beta}_2$, where the Type I error falls to about 3 percent.

Table 3
Type I Error Rates for Tests of the Null Hypothesis that $\beta = 0$

Estimator	Df	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
				$\rho = 0$		
Linearization	n-1	7.54	7.00	15.99	7.35	5.38
Linearization	Satt	5.75	6.45	6.33	6.28	5.18
Jackknife	n-1	5.01	3.92	7.58	4.52	5.02
Jackknife	Satt	3.80	3.43	1.41	3.26	4.77
Kott (1994)	Kott	4.87	5.03	7.13	5.21	4.67
Kott (1996)	Kott	5.11	5.08	4.85	4.76	5.07
BRL	n-1	6.28	5.37	11.25	5.90	5.21
BRL	Satt	4.73	4.86	3.12	4.72	5.00
				$\rho = 1/9$		
Linearization	n-1	7.81	7.14	16.19	8.18	5.34
Linearization	Satt	6.03	6.60	6.43	7.05	5.14
Jackknife	n-1	5.31	4.06	7.63	4.49	4.77
Jackknife	Satt	4.11	3.61	1.48	3.24	4.51
Kott (1994)	Kott	5.07	5.03	7.00	5.51	4.56
Kott (1996)	Kott	5.42	5.28	5.14	5.32	5.01
BRL	n-1	6.52	5.50	11.27	6.23	5.08
BRL	Satt	5.04	5.00	3.19	4.93	4.84
				$\rho = 1/3$		
Linearization	n-1	8.10	7.28	16.39	8.79	5.66
Linearization	Satt	6.30	6.78	6.62	7.53	5.44
Jackknife	n-1	5.45	4.11	7.76	4.56	4.67
Jackknife	Satt	4.13	3.61	1.51	3.35	4.46
Kott (1994)	Kott	5.14	5.06	7.02	5.80	4.84
Kott (1996)	Kott	5.59	5.44	5.14	5.88	5.31
BRL	n-1	6.76	5.63	11.55	6.45	5.19
BRL	Satt	5.18	5.14	3.30	5.26	4.98

Note: Entries with a true value of 5.00 percent have standard errors of 0.07 percent.

Tests based on Kott's 1994 estimator with his proposed degrees of freedom perform very well for the coefficients where the variance estimator is biased upward. It appears that the upward bias in the variance estimator is offset by the upward bias in the approximate degrees of freedom. Kott's variance estimator is slightly negatively biased for $\hat{\beta}_2$ and therefore the upward bias in the degrees of freedom compounds the bias in the estimator resulting in a Type I error rate of about 7 percent for all three values of ρ.

Tests based on Kott's 1996 estimator also perform well. For almost all the coefficients and all values of ρ the Type I error rate is close to 5 percent. The exception is the test for $\hat{\beta}_3$ when $p = 1/3$, which has an error rate of 5.88 percent as a result of the moderate bias in the variance estimator.

7. EXAMPLE FROM THE PARTNERS IN CARE EXPERIMENT

We illustrate the methods in this paper using data from Partners in Care, a longitudinal experiment assessing the effect of "quality improvement" programs on care for depression in managed care organizations (MCOs) (Wells *et al.* 2000). The experiment followed 1356 patients who screened positive for depression in 1996-1997 in 43 clinics of seven MCOs. Clinics were assigned at random to one of three experimental cells: usual care, a quality improvement program supplemented by resources for medication follow-up, or a quality improvement program supplemented by resources for access to psychotherapists. Clinics were assigned at random after forming 27 clinic sets–three for each of nine blocks (six MCOs constituted single blocks, and one MCO was divided into three blocks based on ethnic mix of the clinics). Within blocks of more than three clinics, clinic sets were combined to match as closely as possible on anticipated sample size and patient characteristics. See Wells *et al.* (2000) for additional details.

We present results from an OLS regression on the mental health summary score from the SF-12 (Ware, Kosinski and Keller 1995) for 1048 patients at 6-month follow-up. Scores were standardized to have mean 50 and standard deviation 10 in a general population, with higher scores indicating better health. As in Wells *et al.* (2000), the explanatory variable of primary interest is an intervention indicator that estimates the combined effect of medication or therapy versus care as usual. The first two columns of Table 4 show OLS coefficients and standard errors for the intervention effect and all the covariates used by, but not reported in, Wells *et al.* (2000). Our regression differs from theirs because we do not weight for nonresponse or impute for missing values of the outcome variable, but the results for the intervention effect agree reasonably closely.

Table 4
Comparison of OLS, Linearization, and BRL Inference for Partner-in-Care

Explanatory Variable	$\hat{\beta}_j$	SE_{OLS}	$\frac{SE_{LIN}}{SE_{OLS}}$	$\frac{SE_{BRL}}{SE_{OLS}}$	DF_{BRL}	P-value OLS	LIN	BRL
PSU-Level								
Intercept	28.795	3.409	1.03	1.06	23.7	0.000	0.000	0.000
Intervention	1.724	0.746	0.73	0.84	15.4	0.021	0.003	0.015
Block 1	1.386	1.867	0.63	0.80	2.7	0.458	0.244	0.426
Block 2	-0.031	1.576	0.88	1.07	3.6	0.984	0.982	0.986
Block 3	-1.042	1.230	0.53	0.61	3.9	0.397	0.117	0.241
Block 4	0.038	1.231	0.62	0.73	4.5	0.976	0.961	0.968
Block 5	-3.707	1.503	0.66	0.78	4.7	0.014	0.001	0.027
Block 6	-0.025	1.562	1.15	1.32	4.9	0.987	0.989	0.991
Block 7	-2.784	1.644	0.84	0.97	7.0	0.090	0.051	0.126
Block 8	0.822	1.233	0.93	1.03	12.0	0.505	0.476	0.527
Demographic								
Black	0.972	1.448	0.74	0.79	7.6	0.502	0.369	0.419
Hispanic	0.202	1.004	0.73	0.75	24.3	0.841	0.785	0.791
Other nonwhite	-1.033	1.409	0.77	0.80	21.6	0.463	0.349	0.369
Female	-0.502	0.803	1.09	1.12	23.1	0.532	0.571	0.581
Log of net worth + $1,000	0.015	0.215	0.87	0.89	23.6	0.943	0.936	0.937
Less than high school	-1.690	1.217	1.00	1.04	25.3	0.165	0.173	0.192
Some college	-1.140	0.879	0.77	0.78	26.0	0.195	0.097	0.108
College graduate	-0.703	1.047	0.78	0.79	21.1	0.502	0.393	0.404
Age	0.059	0.032	0.91	0.93	26.5	0.064	0.047	0.056
Married	0.541	0.748	1.05	1.07	28.5	0.470	0.496	0.504
Baseline Health								
1 chronic condition (of 19)	-0.973	1.039	0.92	0.94	23.7	0.349	0.313	0.327
2 chronic conditions	0.198	1.116	0.87	0.90	23.0	0.859	0.840	0.846
3+ chronic conditions	-0.201	1.132	0.90	0.91	24.0	0.859	0.844	0.847
Depression and dysthymia	-5.305	1.335	0.93	0.95	25.8	0.000	0.000	0.000
Depression or dysthymia	-3.882	0.982	1.12	1.15	23.7	0.000	0.001	0.002
Prior depression only	-2.396	1.109	1.02	1.05	21.2	0.031	0.040	0.052
Mental component of SF-12	0.287	0.036	1.11	1.14	26.6	0.000	0.000	0.000
Physical comp of SF-12	0.079	0.036	0.88	0.89	24.6	0.029	0.017	0.022
Anxiety disorder	-2.438	0.749	1.20	1.23	26.3	0.001	0.010	0.014

Because patients from the same clinics could have similar outcomes, OLS standard errors could easily be too low-especially for PSU–level variables like Intervention. Columns 3 and 4 of Table 4 show the ratios of linearization and BRL standard errors to the OLS standard errors. We use clinic as the PSU because there is very little reason to expect correlations of errors across clinics after controlling for block.

Using the method of Wu, Holt and Holmes (1988), we estimate the intra-clinic correlation of the errors as -0.0026, easily consistent with a true value of 0. Nonetheless, there is no reason to expect any of the correct standard errors to fall much below those obtained from OLS. Column 3 of Table 4 shows that the linearization standard errors frequently fall far below those obtained from OLS – especially for the PSU–level explanatory variables at the top of the table. Similarly, linearization with a reference t_{n-1} often produces much smaller P–values than does OLS. BRL improves over linearization. BRL standard errors are always larger and sometimes substantially larger than the linearization standard errors. For example, the BRL estimates for PSU–level explanatory variables are on average 15 percent larger than the linearization estimates. On the other hand, BRL standard errors for PSU–level variables are still often smaller than the OLS estimates. Thus, even though BRL estimators should be nearly unbiased, the variability in the estimators results in estimates for some coefficients that are small. The variability is also reflected in degrees of freedom that are very small for the block indicators and, while larger for patient level variables, are still considerably less than 42, the number of clusters minus one. The degrees of freedom are especially small, 7.6, for the indictor variable Black (equal to one if the patient was African American and zero otherwise). Plots analogous to Figure 1 show that Black was concentrated in three clusters. The Black indicator equals zero for all the patients in 24 of 43 clusters, and 48 of the 78 African Americans in the sample were found in just three clusters. As discussed in sections 2 and 4, the concentration of Black into a small number of clusters results in high variance for both estimators and large bias in the linearization estimator, both of which can be seen in Table 4.

8. DISCUSSION

Although linearization is a valuable tool that provides consistent standard errors and valid inference as the number of PSUs grows large in multi-stage samples, users should recognize problems with the method. Estimated variances of linear regression coefficients (including domain means) tend to be biased low – especially for coefficients (or linear combinations of coefficients) that depend largely on data from a small number of PSUs. Depending on the design, large biases can persist even when the total number of PSUs is quite large. The standard jackknife for multi-stage samples tends to have at least as large bias in the opposite direction. Similarly, using a reference t distribution with degrees of freedom equal to one less than the number of PSUs may greatly understate the uncertainty in the estimated variance. Because the two problems (bias and overstated degrees of freedom) tend to occur in tandem for linearization, confidence intervals and statistical tests based on that method may be far too liberal.

Bias reduced linearization (BRL) produces unbiased variance estimates in the event that errors are homoskedastic and uncorrelated, and it tends to greatly reduce bias for other covariance structures investigated in our simulations. In our simulations, BRL consistently exhibited smaller biases than linearization by 90 percent or more and tended to improve substantially on Kott's 1994 adjusted linearization method. Results for BRL were comparable to those for Kott's 1996 method.

When BRL was used with the estimated Satterthwaite degrees of freedom, statistical inference improved greatly in comparison with the standard use of linearization. Bias reduction and Satterthwaite degrees of freedom seemed to contribute about equally to the improved performance. Although Satterthwaite's approximation may overcompensate, leading to conservative inference in certain situations, the problem does not seem noteworthy until the Satterthwaite degrees of freedom drop below 5 (based, in part, on simulations not reported in this paper). In such cases, analysts might choose to estimate critical values using simulations based on Theorem 4.

It is important to note some limitations of our simulation results. First, we only report results for four distinct explanatory variables plus an intercept. We choose those variables to span a wide variety of situations. Although some might describe x_2 as extreme or pathological, it is not outside the range of situations that we have seen in our own consulting work. Variables like x_2 can results from group-randomized trials (see section 7) or observational data where only a few PSUs exhibit a particular trait or from use of a series of dummy variables to represent levels of a categorical variable. Second, we present results only for $n = 20$ PSUs. To the extent that **X** remains similar as n increases (*e.g.*, by replication), Equation (4) implies that the bias declines in proportion to $1/(n-1)$. Also, the results observed for $n = 20$ could occur for much larger n if the bulk of the variation in **X** is contributed by a few PSUs, and the determination of $l'\hat{\beta}$ depends similarly on a small number of PSUs. Finally, to reduce the number of factors affecting the results, we simplified the design in several ways: constant PSU sizes, no weights or strata, and little multicollinearity. We suspect that relaxing any of those constraints would actually tend to make standard linearization and the jackknife perform worse. We do not believe that the choice of $m = 10$ for the PSU size had much impact either way on our findings.

Although we believe that our proposed methods will prove valuable to analysts of multi-stage samples, these methods will not completely solve the inference problem for unweighted linear regression. Both authors have frequently observed the disturbing situation where standard linearization methods produced shorter confidence intervals than methods that ignore the design. Certainly, the bias of v_L and improper use of $n - 1$ degrees of freedom contribute to the frequency of this phenomenon, but our methods would not eliminate its occurrence (see section 7). Linearization, like sample reuse methods, necessarily produces estimators with high variance for some or possibly all coefficients in certain designs. When confronted with situations like the coefficients for our x_2, where the Satterthwaite degrees of freedom fall near 3 or lower, analysts should seriously consider whether they can afford the large variability, and corresponding loss of power, that comes with nonparametric variance estimators. Parametric alternatives like hierarchical linear models or inference based on estimating a common intra-class correlation across all the PSUs (Wu, Holt, and Holmes 1988) should produce more stable results.

Although this paper has focused on unweighted linear regression for samples without stratification, we have no reason to expect that the bias and degrees-of-freedom problems of linearization would be lessened by stratification or for either weighted least squares or generalized linear models (GLMs). As shown in McCaffrey, Bell and Botts (2001) the BRL method extends immediately to weighted linear regression by using $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ in the main condition of Theorem 3. Because solutions to GLMs, such as logistic regression, are equivalent to the final steps of iteratively reweighted least squares (McCullagh and Nelder 1989), the obvious choice for these models is to use BRL based on the final weights and to set $\mathbf{U} = \mathbf{W}^{-1}$. Nevertheless, Theorem 3 does not extend to GLMs because the weights are estimated from the data, and we have not investigated the properties of BRL in this context.

Korn and Graubard (1995) suggest $v_L^{1/2}$ as a standard error estimator for stratified samples in situations where the stratification is non-informative. The same reasoning applies to $v_{\mathrm{BRL}}^{1/2}$. Fuller (1975) proposed an alternative design consistent standard error estimator for stratified samples. Bell and McCaffrey (2002, pages 32-33) show that by adjusting the vector of residuals for each stratum, BRL can reduce or remove the model bias that can exist in Fuller's estimator.

ACKNOWLEDGEMENTS

We thank the referees and associate editor for valuable comments on an earlier draft. This work is supported in part by NSF Grant 0001763.

APPENDIX

Proofs of Theorems 2 and 4

Proof of Theorem 2

Following the first steps of the proof of Theorem 1, equation (6) implies that

$$E(v_{JK}) =$$

$$\left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i' (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1} l.$$

The existence of $(\mathbf{I}_i = \mathbf{H}_{ii})^{-1}$ implies that the eigenvalues of $\mathbf{H}_{ii}$ are strictly less than 1, so that $(\mathbf{I}_i = \mathbf{H}_{ii})^{-1}$ can be written as $\sum_{j=0}^{\infty} \mathbf{H}_{ii}^j$. Consequently, letting $\mathbf{D} = (1/n)(\mathbf{X}'\mathbf{X})$ and $\mathbf{D}_i = (\mathbf{X}_i'\mathbf{X}_i) - \mathbf{D}$, we have

$$E(v_{JK})$$

$$= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{n}\left(\sum_{k=1}^{\infty}\left[(\mathbf{D}+\mathbf{D}_i)(\mathbf{X}'\mathbf{X})^{-1}\right]^k l\right)$$

$$= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{n}\left(\sum_{k=1}^{\infty}\sum_{r=0}^{k}\binom{k}{r}\frac{1}{n^{k-r}}\left[\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}\right]^r l\right)$$

$$= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{n}\left(\sum_{r=0}^{\infty}\sum_{\substack{s=0\\ r+s>0}}^{\infty}\binom{r+s}{r}\frac{1}{n^{s}}\left[\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}\right]^r l\right)$$

The term for $r = 0$ equals $l'(\mathbf{X}'\mathbf{X})^{-1} l = \mathrm{Var}(l'\hat{\beta})$. The term for $r = 1$ equals 0. By the binomial theorem,

$$\sum_{s=0}^{\infty}\binom{r+s}{r}\frac{1}{n^s} = \left(\frac{n}{n-1}\right)^{r+1},$$

so that the remaining terms can be paired, for $r = 2, 4, 6, \ldots$, to give

$$\left(\frac{n}{n-1}\right)^r l' \sum_{i=1}^{n}\left\{\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_i\right]^{r/2}\right.$$

$$\left.\left[(\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}\right][\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}]^{r/2}\right\} l$$

The middle factor in the summation can be written as ,

$$\left(\frac{n-2}{n-1}\right)(\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1}\right)(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_i'\mathbf{X}_i)(\mathbf{X}'\mathbf{X})^{-1},$$

which is positive definite, so that the whole expression must be positive. Consequently, we have shown that $E(v_{JK}) \geq \mathrm{Var}(l'\hat{\beta})$ with equality if and only if $l'(\mathbf{X}'\mathbf{X})^{-1\prime}\mathbf{D}_i = 0$, which is true if and only if $l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{X}_i$ is constant across i.

Proof of Theorem 4

$$v^* = c\sum_{i=1}^{n} l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{A}_i(\mathbf{I}-\mathbf{H})_i\varepsilon\varepsilon'(\mathbf{I}-\mathbf{H})_i'$$

$$\mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$$

$$= \varepsilon'\sum_{i=1}^{n}\mathbf{g}_i\mathbf{g}_i'\varepsilon.$$

Let **P** equal the matrix of eigenvectors and $\mathbf{\Lambda}$ denote the diagonal matrix with elements $\lambda_1, \ldots, \lambda_M$ equal to the eigenvalues of $\mathbf{V}^{1/2}\sum_{i=1}^{n}\mathbf{g}_i\mathbf{g}_i'\mathbf{V}^{1/2} = \mathbf{B}'\mathbf{B}$ where $\mathbf{B}' = \mathbf{V}^{1/2}[\mathbf{g}_1\mathbf{g}_2\ldots\mathbf{g}_n]$. Let $\mathbf{u} = \mathbf{P}'\mathbf{V}^{-1/2}\mathbf{y}$ where $\mathbf{V}^{1/2}\mathbf{V}\mathbf{V}^{-1/2} = \mathbf{I}$ defines $\mathbf{V}^{1/2}$, then the elements of **u** are independent normal variables with variance 1 and

$$v^* = \mathbf{u}'\mathbf{\Lambda}\mathbf{u} = \sum_{i=1}^{M}\lambda_i u_i^2.$$

Let λ_i be any nonzero eigenvalue of $\mathbf{B}'\mathbf{B}$, then there exits a nonzero vector **z** such that $\mathbf{B}'\mathbf{B}\mathbf{z} = \lambda_i\mathbf{z}$ and $\mathbf{B}\mathbf{B}'\mathbf{B}\mathbf{z} = \lambda_i\mathbf{B}\mathbf{z}$. Because $\mathbf{B}\mathbf{z} \neq \mathbf{0}$, λ_i is an eigenvalue of $\mathbf{B}\mathbf{B}'$. Similarly, any nonzero eigenvalue of $\mathbf{B}\mathbf{B}'$ is also an eigenvalue of $\mathbf{B}'\mathbf{B}$. Therefore, the nonzero eigenvalues of $\mathbf{B}'\mathbf{B}$ equal the nonzero eigenvalues of $\mathbf{B}\mathbf{B}' = \{\mathbf{g}_i'\mathbf{V}\mathbf{g}_j\}$.

REFERENCES

BELL, R.M., and MCCAFFREY, D.F. (2002). *Bias Reduction in Linearization Standard Errors for Linear Regression with Multi-Stage Samples.* AT&T Labs-Research, Florham Park, NJ, TD-4S9H9T, www.research.att.com/~rbell.

BRYK, A.S., and RAUDENBUSH, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newberry Park, CA: Sage.

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

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

EFRON, B., and TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

ELLICKSON, P.L., and MCGUIGAN, K.A. (2000). Early predictors of adolescent violence. *American Journal of Public Health.* 90, 566-572.

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā* C, 37, 117-32.

GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

GOLDSTEIN, H. (1991). Multilevel Modeling of Survey Data. *The Statistician.* 40, 235-244.

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

KORN, E.L., and GRAUBARD, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A, General.* 158, 263-295.

KOTT, P.S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology.* 20, 159-64.

KOTT, P.S. (1996). Linear regression in the face of specification error: model-based exploration of randomization-based techniques. *Proceedings of the Survey Methods Section,* Statistical Society of Canada. 39-47.

LANDIS, J.R., LEPKOWSKI, J.M., EKLAND, S.A. and STEHOUWER, S.A. (1982). A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and Health Statistics,* Series 2, 92, Washington, D.C: US Government Printing Office.

MACKINNON, J.G., and WHITE, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics.* 29, 305-325.

MANCL, L.A., and DEROUEN, T.A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics.* 57, 126-134.

MCCAFFREY, D.F., and BELL, R.M. (1997). Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units. Paper presented at the Joint Statistical Meetings, Anaheim CA.

MCCAFFREY, D.F., BELL, R.M. and BOTTS, C.H. (2001). Generalizations of bias reduced linearization. *Proceeding of the Survey Research Methods Section,* American Statistical Association.

MCCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models.* Second Edition, London: Chapman and Hall.

MURRAY, D. M., HANNAN, P. J., WOLFINGER, R. D., BAKER, W.L. and DWYER, J.H. (1998). Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine.* 17, 1581-1600.

RUST, K.F., and RAO, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research.* 5, 283-310.

SAS INSTITUTE INC. (1999). *SAS/STAT* User's Guide, Version* 8. Cary, NC: Author.

SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics.* 2, 110-114.

SEARLE, S.R., CASELLA, G. and MCCULLOCH, C.E. (1992). *Variance Components.* New York: John Wiley & Sons Inc.

SHAH, B.V., BARNWELL, B.G. and BIELER, G.S. (1997). *SUDAAN User' Manual, Release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

SHAH, B.V., HOLT, M. M. and FOLSOM, R.E. (1977). Inference About Regression Models from Survey Data. *Bulletin of the*

International Statistical Institute. 41, 43-57.

SHAPIRO, M.F., MORTON, S.C., MCCAFFREY, D.F., SENTERFITT, J.W., FLEISHMAN, J.A., PERLMAN, J.F., ATHEY, L.A., KEESEY, J.W., GOLDMAN, D.P., BERRY, S. H. and BOZZETTE, S.A. (1999). Variations in the care of hiv-infected adults in the United States; results from the hiv cost and services utilization study. *Journal of the American Medical Association*. 281, 2305-2315.

SKINNER, C.J. (1989a). Introduction to Part A. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, and T.M.F. Smith). New York: John Wiley & Sons Inc. 23-57.

SKINNER, C.J. (1989b). Domain means, regression and multivariate analyses. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons Inc. 59-88.

STATACORP. (1999). *Stata Statistical Software: Release 6.0*. College Station, TX: Author.

THEIL, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons Inc.

WARE, J.E., JR., KOSINSKI, M. and KELLER, S.D. (1995). *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales*. Boston, Mass: The Health Institute, New England Medical Center.

WELLS, K.B., SHERBOURNE, C., SCHOENBAUM, M., DUAN, N., MEREDITH, L., UNUTZER, J., MIRANDA, J., CARNEY, M. and RUBENSTEIN, L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association*. 283, 212-220.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 48, 817-838.

WU, C.J.F., HOLT, D. and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*. 83, 150-9.

ZEGER, S.L., and LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 42, 121-130.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 183-190
Statistics Canada

Design Effects of Sampling Frames in Establishments Survey

MONROE G. SIRKEN[1]

ABSTRACT

When stand-alone sampling frames that list all establishments and their measures of size are available, establishment surveys typically use the Hansen-Hurwitz (HH) pps estimator to estimate the volume of transactions that establishments have with populations. This paper proposes the network sampling (NS) version of the HH estimator as a potential competitor of the HH estimator. The NS estimator depends on the population survey-generated establishment frame that lists households and their selection probabilities in a population sample survey, and the number of transactions, if any, of each household with each establishment. A statistical model is developed in this paper to compare the efficiencies of the HH and NS estimators in single-stage and two-stage establishment sample surveys assuming the stand-alone sampling frame and the population survey-generated frame are flawless in coverage and size measures.

KEY WORDS: Stand-alone establishment frames; Population survey-generated establishment frames; Hansen-Hurwitz estimator; Network sampling estimator.

1. INTRODUCTION

Listings of establishments that have transactions with households in population sample surveys serve as sampling frames of establishment surveys whenever the transactions reported by households in the population surveys are matched with the records of their establishments. For example, the listings of establishments that have transactions with households in the National Medical Expenditure Panel Survey (MEPS), a national population sample survey, serve as sampling frames for medical provider surveys that supplement and verify the medical expenditures of the transactions reported by MEPS household respondents (Cohen 1998). However, listings of establishments that have transactions with households in population sample surveys rarely serve as frames of establishment surveys that collect information about the transactions that establishments have with all households. The Current Price Index (CPI) produced by the Bureau of Labor Statistics is a notable and rare exception of a Federal establishment survey that depends on a population survey-generated sampling frame. The CPI Pricing Survey, a national retail establishment survey, that collects prices for a basket of consumer goods purchased by all customers, uses as its sampling frame the listings of retail establishments that have transactions with households in the CPI Continuing Point of Purchase Survey. (Leaver and Valliant 1995).

After reviewing plans of the National Center for Health Statistics (NCHS) to restructure its family of independent national surveys of health providers (hospitals, physicians, clinics, *etc.*), a Panel of the Committee on National Statistics proposed (Wunderlich 1992) using listings of health care providers reported by households in the National Health Interview Survey (NHIS), an ongoing national household sample survey (Massey, Moore, Parsons and Tadros 1991) as the sampling frames for national surveys of health care providers. The Committee thought that, especially in the current environment of rapid changes in listings of health care providers due to rapid changes in the nation's health care delivery system, the NHIS-generated health care provider frames would be more accurate and easier and less expensive to construct and maintain than the free-standing health care provider frames currently in use. Soon after the Panel report was issued, NCHS initiated a research project on population survey-generated sampling frames that is briefly summarized below.

Initially, the research focused almost exclusively on the statistical properties of NHIS-generated frames of health care providers. Judkins, Berk, Edwards, Mohr, Stewart and Waksberg (1995) studied the quality of the free-standing health provider frames currently in use or of potential use, and discussed the kinds of medical providers for which NHIS-generated frames would seem to have the greatest potential. Subsequently, Judkins, Marker, Waksberg, Botman and Massey (1999) made rough comparisons of the efficiencies of dental surveys using the NHIS-generated sampling frame and using the free-standing frame, and concluded that NHIS-generated health care provider frames deserve serious consideration whenever reasonably complete free-standing health care provider frames with reasonably good size measures are unavailable.

In recent years, the research has focused on the statistical properties of estimators that depend on population-generated sampling frames and has become more theoretically focused than formerly. The conceptual difficulties initially encountered in developing unbiased estimators for the population survey-generated frame because the same establishments have transactions with multiple households were overcome by applying network sampling theory. (Sirken

[1] Monroe G. Sirken, Senior Research Scientist, National Center for Health Statistics, U.S.A.

1997; Thompson 1992). Sirken, Shimizu and Judkins (1995) developed the network sampling version of the HH estimator, referred to in this paper as the NS estimator, and Sirken and Shimizu (1999) developed the network sampling version of the Horwitz-Thompson (HT) estimator. This paper develops a statistical error model that compares the efficiencies of the NS estimator that depends on the population survey-generated frame, and the HH estimator that depends on the free-standing frame. The error model assumes both frames are flawless in establishment coverage and size measures and have equivalent construction and maintenance costs. Though the model assumes a srs design for the population survey that generates population survey-generated sampling frame, the model can be applied to other kinds of population survey designs that are not considered in this paper.

This paper is organized as follows. Notation follows in section 2. Section 3.1 and section 3.2 respectively present the pps self-weighted HH estimator and variance of the two-stage establishment sample survey that depends on the free-standing sampling frame, and the NS estimator and variance of a two-stage establishment survey that depends on the population survey-generated frame. The error model is developed in sections 4.1- 4.4. The difference between two-stage HH and NS variances of equivalent expected sample sizes is developed in section 4.1. In section 4.2, the first stage variance component of the two-stage NS estimator is split into variance components representing effects of households with and without transactions, and section 4.3 shows the design effects of the NS estimator in single stage sampling. Second stage variance components of the NS and HH estimators are compared in section 4.4. In the concluding section 5, the error model's major findings comparing efficiencies of HH and NS estimators in single-stage and two-stage establishment surveys are briefly summarized, and limitations of the model are briefly discussed. The appendix presents the proof of a statistical statement appearing in section 4.2.

2. NOTATION

Let N_j = the number of households having transactions with establishment $j(j = 1, 2, ..., R), N_o$ = the number of households not having transactions with any establishments, and N^* = the number of distinct households having transactions with R establishments. Then, $N = N^* + N_o$ = the total number of households.

Let M_{ij} = the number of transactions of establishment $j(j = 1, 2, ..., R)$ with household $i(i = 1, 2, ..., N)$, where $M_{ij} \geq 0$ when establishment j has transactions with household i, and $M_{ij} = 0$ when establishment j and household i do not have transactions. Then, $M_j = \sum_{i=1}^{N} M_{ij}$ = the number of transactions of establishment j with N households, and $M = \sum_{j=1}^{R} M_j$ = the number of transactions of M establishments with N households, and $\bar{M} = M/N$ the average number of transactions per household.

Let X_{jk} denote the value of the x-variate for transaction $k(k = 1, ..., M_j)$ of establishment $j(j = 1, 2, ..., R)$. Then, $X_j = \sum_{k=1}^{M_j} X_{jk}$ = the sum of the x-variate over the M_j transactions of establishment j, and $X = \sum_{j=1}^{R} X_j$ = sum of the x-variate over the M transactions of R establishments. Let $\bar{X}_j = X_j/M_j$ = the average value of the x-variate over the M_j transactions of establishment j, and $\bar{X} = X/M$ = the average value of the x-variate over M transactions.

3. ESTIMATORS AND VARIANCES

3.1 The HH Estimator and Variance

Consider a two-stage self weighted establishment sample survey using a free-standing establishment sampling frame that lists all R establishments and their measures of size, $M_j (j = 1, 2, ..., R)$. Establishments are the primary sampling units (psu's), and transactions are the secondary sampling units. A sample of r establishments is selected with pps with replacement from the free-standing frame, and a sample of size $t_{\text{HH}} < \min(M_1, ..., M_j, ..., M_R)$ transactions each, where t_{HH} is a positive integer, is independently selected by simple random sampling without replacement for each sample establishment $j(j = 1, 2, ..., r)$.

The unbiased self-weighted pps HH estimator of X is

$$X'_{\text{HH}} = \frac{M}{r} \sum_{j=1}^{r} \bar{X}_j' \tag{1}$$

where $\bar{X}_j' = \sum_{k=1}^{t_{\text{HH}}} X_{ij} / t_{\text{HH}}$ is the unbiased estimate of $\bar{X}_j = X_j/M_j (j = 1, 2, ..., R)$. Because establishments are selected with replacement, the HH estimator counts $\bar{X}_j$ as many times as establishment j is selected in the sample.

The variance of the X'_{HH} is (Thompson 1992)

$$\text{Var}(X'_{\text{HH}}) = \frac{M^2}{r} \sigma^2_{\text{HH}1} + \frac{M}{r t_{\text{HH}}} \sum_{j=1}^{R} (M_j - t_{\text{HH}}) \sigma_j^2 \tag{2}$$

where the first and second terms respectively on the right side of (2) are the first and second stage variance components, and

$$\sigma^2_{\text{HH}1} = \frac{1}{M} \sum_{j=1}^{R} M_j (\bar{X}_j - X/M)^2 \tag{3}$$

is the between establishment population variance, and

$$\sigma_j^2 = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (X_{jk} - X_j/M_j)^2 \tag{4}$$

is the within establishment population variance of establishment j.

3.2 The NS Estimator and Variance

Consider a two-stage establishment sample survey that depends on a population survey-generated frame. The frame lists n sample households $H_i'(i = 1, 2, ..., n)$ that were enumerated in a population sample survey. For each listed household H_i', the frame provides π_i, its selection probability in the household survey, and M_{ij}, the number of its transactions with each distinct establishment $j(j = 1, 2, ..., R)$. (The M_{ij}'s are reported by household respondents in the population sample survey).

Each of the n listed households in the population survey-generated frame represents a cluster of establishments ranging in size from 0 to R establishments with whom the household has transactions. The n clusters of establishments are the primary sampling units, and the $M_j\,(j = 1, 2, ..., r)$ transactions of the r sampled establishments are secondary sampling units. The transaction sample for establishment j $j(j = 1, 2, ..., R)$ is selected as follows: a srs sample of size $t_{\text{NS}}\,M_{ij} < \text{Min}(M_1, M_2, ..., M_r)$ transactions is independently selected without replacement for each sample household $H_i'\ (i = 1, 2, ..., n)$, where t_{NS} is a positive integer. The transaction sample size of establishment $j(j = 1, 2, ..., R)$ is equal to $t_{\text{NS}} \sum_{i=1}^{n} M_{ij}$, and the total transaction sample size is equal to τt_{NS}, where $\tau = \sum_{i=1}^{n} \sum_{j \in A_i} M_{ij} =$ the sum of the transactions over n sample households is a random variable.

The NS estimator of X is

$$X_{\text{NS}}' = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j \in A_i} M_{ij}\, \bar{X}_j'\ (i)$$

where A_i is the cluster of distinct establishments that have transactions with sample household H_i, and

$$\bar{X}_j'\ (i)\ = \sum_{k=1}^{t_{\text{NS}} M_{ij}} X_{jk} / (t_{NS}\, M_{ij})$$

is an unbiased estimate $\bar{X}_j$ for a sample of $t_{\text{NS}}\,M_{ij}$ transactions of establishment j. Because households are selected with replacement, the NS estimator counts the quantity $\sum_{j \in A_i} M_{ij} \bar{X}_j'(i)$ every time household $H_i\,(i = 1, 2, ..., n)$ is selected in the sample, and because the same establishment has transactions with multiple households, the NS estimator counts the quantity $M_{ij}\, \bar{X}_j'(i)$ every time a sample household $i\ (i = 1, 2, ..., n)$ contains establishment j.

Assuming a srs design in the population survey, $\pi_i = n/N$, and the network sampling estimator is

$$X_{\text{NS}}' = \frac{N}{n} \sum_{i=1}^{n} \sum_{j \in A_i} M_{ij}\, \bar{X}_j'(i). \tag{5}$$

The NS estimator is an unbiased estimator of X.

$$\begin{aligned} E(X_{\text{NS}}') &= \sum_{i=1}^{n} E \sum_{j \in A_i} M_{ij} \bar{X}_j'(i) = \sum_{i=1}^{N} \sum_{j \in A_i} M_{ij}\, \bar{X}_j' \\ &= \sum_{i=1}^{R} M_j\, \bar{X}_j = \sum_{J=1}^{R} X_j = X. \end{aligned}$$

The NS estimator in (5) is self-weighted because we have assumed that the n households are selected by srs. It would be a self-weighted estimator whenever the sample design of the population sample survey that generates the establishment sampling frame is self-weighted. When $N = N^* = M$, implying that N^* households each has a single transaction, and $N_0 = N - N^*$ households are without transactions, and when $n = r$ and $t_{\text{NS}} = t_{\text{HH}}$, the HH and NS estimators are equivalent.

$$\begin{aligned} X_{\text{NS}}' &= \frac{N}{n} \sum_{i=1}^{M} \sum_{j \in A_i} M_{ij}\, \bar{X}_j'(i) = \frac{N}{n} \sum_{i=1}^{N} \sum_{j \in A_i} \bar{X}_j' \\ &= \frac{M}{r} \sum_{j=1}^{R} \bar{X}_j' = X_{\text{HH}}. \end{aligned} \tag{6}$$

The variance of the NS estimator (5), under srs sampling with replacement of n households and independent selections of $t_{\text{NS}}\,M_{ij}$ transaction by srs without replacement for each establishment j linked to household H_i, is (Sirken *et al.* 1995)

$$\text{Var}(X_{\text{NS}}') = \frac{N^2}{n} \sigma_{\text{NS}\,1}^2 + \frac{N}{n t_{\text{NS}}} \sum_{i=1}^{N} \sum_{j=1}^{R} M_{ij}\, \frac{M_j - t_{\text{NS}}\, M_{ij}}{M_j}\, \sigma_j^2 \tag{7}$$

where the first and second terms respectively on the right side of (7) are the first and second stage variance components, and

$$\sigma_{\text{NS}\,1}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j \in A_i} M_{ij}\, \bar{X}_j' - X/N \right)^2 \tag{8}$$

is the population variance between households, and σ_j^2, the population variance within establishment j as defined in (4). An unbiased estimate of NS variance is

$$\text{Var}(X_{\text{NS}}') = \frac{N^2}{n(n-1)} \sum_{i=1}^{n} \left[\sum_{j \in A_i} M_{ij}\, \bar{X}_j'(i) - \bar{X}' \right]^2 \tag{9}$$

where $\bar{X}' = X'/N$.

4. THE ERROR MODEL

4.1 HH and NS Variances of Equivalent Expected Sample Size

Subtracting (2) from (7), the difference between the variances of the HH and NS estimators of X is

$$\mathrm{Var}\,(X'_{\mathrm{NS}}) - \mathrm{Var}(X'_{\mathrm{HH}}) = \left[\frac{N^2}{n}\sigma^2_{\mathrm{NS}1} - \frac{M^2}{r}\sigma^2_{\mathrm{HH}1}\right]$$

$$+ \left[\frac{N}{nt_{\mathrm{NS}}}\sum_{i=1}^{N}\sum_{j=1}^{R} M_{ij}\frac{M_j - t_{\mathrm{NS}}M_{ij}}{M_j}\sigma_j^2\right.$$

$$\left. - \frac{M}{rt_{\mathrm{HH}}}\sum_{j=1}^{R}(M_j - t_{\mathrm{HH}})\sigma_j^2\right] \qquad (10)$$

where the first and second set of bracketed terms respectively on the right side of (10) represent the differences between the primary and secondary variance components of the HH and NS estimators of X.

Let $m_{\mathrm{NS}} = \tau t_{\mathrm{NS}}$ = the size of the transaction sample in the establishment survey using the population survey-generated frame, where t_{NS}, a positive integer, is the size of the transaction sample selected per transaction of the n sample households, and $\tau = \sum_{i=1}^{n}\sum_{j\epsilon A_i} M_{ij}$ = sum of the transactions of n sample households.

Clearly, τ is a random variable and its expected value conditional over all samples of n households is $E(\tau|n) = n\bar{M}$ where $\bar{M} = M/N$ = average household transaction size. It follows that $E(m_{\mathrm{NS}}\,|\,n) = t_{\mathrm{NS}}\,E(\tau|n) = n\bar{M}t_{\mathrm{NS}}$ is the expected transaction sample size of the NS estimator conditional over all samples of n households.

Let $m_{\mathrm{HH}} = rt_{\mathrm{HH}}$ = the size of the transaction sample in the establishment survey using the stand-alone frame, where r = the establishment sample size, and t_{HH} = the transaction sample size per selected establishment. Let $r = E(\tau|n) = n\bar{M}$ and let $t_{\mathrm{HH}} = t_{\mathrm{NS}} = t$, and it follows the expected transaction sample sizes of the NS and HH estimators conditional over all samples of n households are equivalent, namely, $E(m_{\mathrm{HH}}|n) = tE(\tau|n) = nt\bar{M} = E(m_{\mathrm{NS}}|n)$.

Calibrating the establishment and transaction sample sizes in this manner assures that HH and the NS establishment surveys are conducted under roughly the same fiscal constraints if per establishment and per transaction field costs are about the same in both surveys. It is noteworthy, however, that this cost equation does not take into account the differences in costs between constructing and maintaining stand-alone establishment frames and population survey-generated establishment frames.

Substituting $r = n\bar{M}$, $t_{\mathrm{HH}} = t_{\mathrm{NS}} = t$, and $\mathrm{M} = \mathrm{N}\,\bar{M}$ in formula (9), the difference between the NS and HH variances of equivalent expected establishment and transaction sample size conditional over all samples of n households is

$$\mathrm{Var}\,(X'_{\mathrm{NS}}) - \mathrm{Var}(X'_{\mathrm{HH}}) = \frac{N^2}{n}[\sigma^2_{\mathrm{NS}1} - \bar{M}\sigma^2_{\mathrm{HH}1}]$$

$$- \frac{N}{nt}\sum_{j=1}^{R}\sigma_j^2\,[(M_j - t) - \sum_{i=1}^{N}\frac{M_{ij}(M_j - M_{ij})}{M_j}] \qquad (11)$$

The first term and second terms respectively on the right side of (11) represent the difference between the first stage and second stage variance components of the NS and HH estimators of equivalent expected sample sizes conditional over all samples of n households.

4.2 Decomposition of the Single Stage NS Population Variance

Typically, some households do not have transactions with any establishments, and the percentage varies by type of establishment. For example, medical care utilization by families in the United States varies greatly by type of health care provider (Dicker and Sunshine 1987). During a 12 month period, 70 percent of families were not admitted to hospitals, 7 percent did not have ambulatory physician visits, and 28 percent did not have dental visits.

Let

$P = \dfrac{N^*}{N}$ = fraction of N households with one or more transactions, and

$P_0 = 1 - P\dfrac{N_0}{N}$ = fraction of N households without any transactions.

We demonstrate in the Appendix that the single stage population variance of the NS estimator of X, when expressed as a function of P, decomposes into 2 parts

$$\sigma^2_{\mathrm{NS}1}(P) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\epsilon A_i} M_{ij}\,\bar{X}_j - \frac{X}{N}\right)^2$$

$$= P\,\sigma^2_{\mathrm{NS}1^*} + \sigma^2(P)\,E^2_{\mathrm{NS}1^*}\quad 0<P\leq 1 \qquad (12)$$

where

$$\sigma^2_{\mathrm{NS}1} = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\epsilon A_i} M_{ij}\,\bar{X}_j - \frac{X}{N^*}\right)^2 \qquad (13)$$

is the single stage population variance of the x-variate over the truncated population of N^* households with one or more transactions,

$$E^2_{\mathrm{NS1}^*} = \left(\frac{X}{N^*}\right)^2 = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\,\bar{X}_j\right)^2 - \sigma^2_{\mathrm{NS1}^*} \quad (14)$$

is the expected value squared of the x-variate over the truncated population of N^* households and

$$\sigma^2\,(P) = P\,(1-P) \quad (15)$$

is the variance of the binomial variable P. For fixed M, the function $\sigma^2_{NS1}(\mathrm{P}|M)$ is maximum when

$$P = P_{\max} = \frac{1}{2}\left[(\sigma^2_{\mathrm{NS1}^*}/E^2_{\mathrm{NS1}^*}) + 1\right] \le 1.$$

If $\sigma^2_{\mathrm{NS1}^*} \ge E^2_{\mathrm{NS1}^*}$, $P_{\max} = 1$ and if $\sigma^2_{\mathrm{NS1}} < E^2_{\mathrm{NS1}^*}$, $1/2 < P_{\max} < 1$.

When $P = 1$, $\sigma^2(P = 1) = 0$ and therefore $\sigma^2_{\mathrm{NS1}}(P=1) = \sigma^2_{\mathrm{NS1}^*}$. If $P = \bar{M} = (M/N) = 1$, implying that each of N households has a single transaction,

$$\sigma^2_{\mathrm{NS1}}\,(P = \bar{M} = 1) = \sigma^2_{\mathrm{NS1}^*}\,(N^* = M) = \sigma^2_{\mathrm{HH1}} \quad (16)$$

because

$$\sigma^2_{\mathrm{NS1}^*}\,(N^* = M) = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\,\bar{X}_j - \frac{X}{N^*}\right)^2$$

$$= \frac{1}{M}\sum_{j=1}^{R} M_j\left(\bar{X}_j - \frac{X}{M}\right)^2 = \sigma^2_{\mathrm{HH1}}, \quad (17)$$

and, $\sigma^2\,(P = 1) = 0$. In other words when $P = \bar{M} = 1$, implying each of the N households has a single transaction, the variance of the NS1 estimator which would then depend on a srs of transactions with replacement is equivalent to the variance of the HH1 estimator that depends on a pps cluster sample of equivalent sample size selected with replacement.

4.3 Design Effects in Single Stage Sampling

Let

$$X'_{\mathrm{NS1}} = \frac{N}{n}\sum_{i=1}^{N}\sum_{j\in A_i} M_{ij}\,\bar{X}_j = \text{the unbiased NS estimator}$$

of X in single stage sampling, and

$$X'_{\mathrm{HH1}} = \frac{M}{R_{\mathrm{HH}}}\sum_{j=1}^{r_{\mathrm{HH}}} \bar{X}_j = \text{the unbiased HH estimator of } X$$

in single stage sampling.

Define the single stage sampling total design effect of the NS1 estimator as the ratio of the variances of the NS1 and HH1 estimators of equivalent sample size conditional over all samples of n households.

$$\lambda\,(P) = \frac{\mathrm{Var}(X'_{\mathrm{NS1}})}{\mathrm{Var}(X'_{\mathrm{HH1}})} = \frac{\sigma^2_{\mathrm{NS1}}\,(P)}{\bar{M}\;\sigma^2_{\mathrm{HH1}}} \quad (18)$$

where $\lambda\,(\mathrm{P}) < 1$ indicates that the NS1 estimator is more efficient than the HH1 estimator, and $\lambda(P) > 1$ indicates that the HH1 estimator is more efficient than the NS1 estimator.

We noted in (12) and (15) that $\sigma^2_{\mathrm{NS1}}(\mathrm{P}) = \mathrm{P}\,\sigma^2_{\mathrm{NS1}^*} + \mathrm{P}(1-P)(X/N^*)^2$, and in (16) that $\sigma^2_{\mathrm{HH1}} = \sigma^2_{\mathrm{NS1}^*}(N^* = M)$. Making these substitutions in (18), the total design effect becomes

$$\lambda\,(P) = \mathrm{deft}^2_{\mathrm{NS1}} + (1-P)\;\; Z_{\mathrm{NS1}}, \quad 0 < P \le 1 \quad (19)$$

where

$$Z_{\mathrm{NS1}} = \frac{P(X/N^*)^2}{\bar{M}\,\sigma^2_{\mathrm{NS1}^*}\,(N^* = M)} \quad (20)$$

is the effect due to the N_o households without transactions, and

$$\mathrm{deft}^2_{\mathrm{NS1}} = \left[\frac{\mathrm{P}\,\sigma^2_{\mathrm{NS1}^*}}{\bar{M}\,\sigma^2_{\mathrm{HH1}}}\right] = \left[\frac{\mathrm{P}\,\sigma^{2^*}_{\mathrm{NS1}}}{\bar{M}\;\sigma^2_{\mathrm{NS1}^*}\;(N^* = M)}\right] \quad (21)$$

is effect due to the N^* households with transactions. In other words, $deft^2_{\mathrm{NS1}}$ is the design effect of *network sampling* a population of N^* household clusters containing one or more transactions, with equal probability and replacement, compared to *network sampling* a population of M transactions, of equivalent expected sample size, by srs and replacement. [The reader is referred to Kish (1982) for the definition of $deft^2$].

The total design effect in (19) depends on $deft^2_{\mathrm{NS}}$ and $\mathrm{Z}_{\mathrm{NS1}}$ and, P, and the values of these parameters, as well as relationships between them, are likely to vary considerably between surveys, and between variables and population domains in the same surveys. Though, in theory, the NS1 estimator could be more efficient than HH1 estimator, in reality that outcome seems highly unlikely because cluster sampling is typically less efficient than srs. A necessary condition for the NS1 estimator to be as efficient or more efficient than the HH1 estimator is that $deft^2_{\mathrm{NS1}} \le 1 - (1-\mathrm{P})\mathrm{Z}_{\mathrm{NS1}}$, and this condition is unlikely to be met particularly if P is small, and if the within household transaction clustering is mostly due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments.

4.4 Comparing Efficiencies in Two-stage Sampling

In two stage sampling, the difference between the HH and NS second stage variance components for equivalent expected sample size of $nt\bar{M}$ transactions conditional over all samples of n households, the second term on the right side of equation (11), reduces to

$$\frac{N}{nt}\sum_{j=1}^{R} \sigma_j^2 \left\{(M_j - t) - \sum_{j=1}^{N} \frac{M_{ij}(M_j - tM_{ij})}{M_j}\right\}$$

$$= \frac{N}{n}\sum_{j=1}^{R} \frac{\rho_j}{M_j}\sigma_j^2 \tag{22}$$

where $\rho_j / M_j = 1/M_j \sum_{i=1}^{N} M_{ij}(M_{ij} - 1)$ is the difference between the HH and NS second stage finite population corrections for establishment j. If none of the N households have multiple transactions with establishment j, the HH and NS second stage variances of establishment j are equivalent and $\rho_j = 0$. Otherwise, $\rho_j > 0$ and second stage variance for establishment j is larger for the HH than the NS estimator. The value of ρ_j is maximum when establishment j has M_j transactions with a single household.

The second stage variance components of the HH and NS estimators are equivalent $\sum_{j=1}^{R} \rho_j = 0$, when, that is, none of the H households have multiple transactions with any of the R establishments. Of course, second stage variances are equivalent if transactions are selected with replacement or the within establishment variances, $\sigma_j^2 = 0 (j = 1, 2, ..., R)$. Except for these contingencies, however, the second stage variance is always larger for the HH estimator than for the NS estimator, and the magnitude of the difference depends on the extent of within household clustering of transactions with the same establishments, and the magnitudes of the within establishment variances.

If none of the N^* households have multiple transactions with the same establishments, the difference between the variances of the HH and NS estimators are equivalent in single stage and two stage establishment sample surveys. Otherwise, the difference between HH and NS variances is less in two stage than in single stage establishment sample surveys because whenever households have multiple transactions with the same establishments the second stage variance is greater for the HH estimator than for the NS estimator.

5. SUMMARY AND CONCLUDING REMARKS

The error model presented in this paper compares efficiencies of two estimators of the volume of transactions between establishments and populations in single-stage and two-stage establishment sample surveys. The Hansen-Hurwitz (HH) estimator depends on a stand-alone sampling frame that lists every establishment and the volume of its transactions with all households during a specified calendar period. The network sampling (NS) estimator depends on a population survey-generated frame that lists the households and their selection probabilities in a population sample survey, and for each household, lists the number of its transactions with each distinct establishment during the specified calendar period.

Also, the NS and HH estimators depend on different establishment survey sample designs. In single-stage sampling, the HH estimator depends on a design in which establishments are the selection units and they are selected with pps with replacement, and the NS estimator depends on a design in which households are the selection units and they are selected with their selection probabilities in the population survey, which the error model assumes is srs with replacement. In two-stage sampling, transactions are the second stage sampling units of the HH and NS estimators. The HH estimator depends on fixed-size transaction samples that are selected by srs independently without replacement. The NS estimator depends on transaction sample sizes that are proportional to the number of transactions of each household with each establishment, and are selected independently by srs without replacement.

The NS and HH estimators are equally efficient, if and only if, every household in the entire population has one and only one transaction. Otherwise, neither the NS or the HH estimator is necessarily more efficient than the other. Nevertheless, it seems likely that the HH estimator will be more efficient than the NS estimator in single-stage establishment survey sampling, and perhaps substantially more efficient especially when large fractions of households do not have any transactions, and/or when the within household clustering of transactions among households with transactions is principally due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments. In two-stage sampling, the outcome is not as transparent as in single stage sampling because the second stage variance component is larger for the HH estimator than the NS estimator by an amount that depends on the extensiveness of within household clustering of transactions with the same establishments.

Arguably the foremost limitation of the error model presented in this paper is the presumption that the stand-alone and population survey-generated sampling frames are flawless in coverage and size measures. However, comparative costs of constructing and maintaining good quality stand-alone and population-generated establishment sampling frames are likely to vary greatly from survey to survey. Though the model seek to equalize the establishment survey costs based on each kind of sampling frames it ignores the differential costs of constructing and maintaining each kinds of frame.

Even in the absence of empirical data about the comparative costs of constructing and maintaining the frames, it is fair to say that the population survey-generated frame should be seriously considered as a potential design alternative whenever constructing and maintaining good quality stand-alone frames would be infeasible or exorbitantly expensive or time consuming, and/or when constructing and maintaining good quality population survey-generated establishment sampling frames would be relatively inexpensive. For example, the population survey-generated frame would be a particularly attractive as a potential design alternative to the stand-alone frame when the stand-alone frame would be difficult to construct and maintain because it was undergoing rapid changing due to births, deaths, and establishment mergers, and the population survey-generated frame costs would be relatively small either because it could be constructed and maintained as a by-product of an ongoing population sample survey (Wunderlich 1992) and/or as a by-product of an ongoing program of matching transactions of households enumerated in a population survey with their establishment records (Cohen 1998).

Another limitation of the model is the unrealistic assumption that the population survey that generates the establishment sampling frame is based on a single stage sample design in which households are selected with equal probabilities and with replacement. In fact, population surveys are virtually always based on multistage sample designs in which households are selected without replacement in the final sampling stage. Typically, the srs assumption tends to significantly understate the variance of the NS estimator, and therefore would have the effect of exaggerating the relative efficiency of the NS estimator compared to the HH estimator. On the other hand, the household sampling with replacement assumption would have the opposite effects, but would be modest (Sirken 2001) compared to the srs assumtion. The error model can be applied, however, to the other population survey sample designs that are not considered in this paper.

The error model presented in this paper identifies the critical parameters that determine the relative efficiency of establishment survey estimators depending on stand-alone and population survey-generated sampling frames. Values of these parameters will vary greatly between surveys and between variables and population domains in the same surveys. Unfortunately, empirical data are currently unavailable, and they are sorely needed to estimate the model's parameters under a broad range of survey conditions. Hopefully, this paper will stimulate interest in conducting establishment surveys that depend on population survey-generated establishment sampling frames, and will lead to improvements in designing establishment surveys that estimate the volume of transactions between establishments and populations.

ACKNOWLEDGMENTS

I thank the 2 referees and in particular an Assistant Editor for very helpful comments. The views expressed in this paper are solely those of the author and do not necessarily represent the official views or positions of the National Center for Health Statistics.

APPENDIX

When expressed as a function of P, the fraction of households with one or more transactions, the single stage population variance of the network sampling (NS) estimator of X

$$\sigma^2_{\text{NS1}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j \in A_i} M_{ij}\,\bar{X}_j - \frac{X}{N}\right)^2$$

decomposes into 2 parts

$$\sigma^2_{\text{NS1}^{\bullet}}(P) = P\sigma^2_{\text{NS1}^{\bullet}} + \sigma^2(P)\,E^2_{\text{NS1}^{\bullet}} \quad 0<P \leq 1$$

where

$$P = \frac{N^*}{N},$$

$$\sigma^2_{\text{NS1}^{\bullet}} = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j \in A_i} M_{ij}\,\bar{X}_j - \frac{X}{N^*}\right)^2$$

is the truncated single stage population variance of the NS estimator exclusive of the $N_0 = N - N^*$ households without transactions with establishments,

$$\sigma^2(P) = P(1-P)$$

is the variance of the binomial variable P, and

$$E^2_{\text{NS1}^{\bullet}} = (X/N^*)^2$$

is the expected value squared of the x-variate distributed over N^* households.

Proof

$$\sigma^2_{\text{NS1}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j \in A_i} M_{ij}\bar{X}_j - \frac{X}{n}\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N^*}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_j - \frac{X}{N^*}\right)^2 + \frac{N_0}{N}\left(\frac{X}{N}\right)^2. \qquad \text{(A.1)}$$

Add and subtract X/N^* to the first term on the right side of (A.1).

$$\frac{1}{N}\sum_{i=1}^{N^*}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_j - \frac{X}{n}\right)^2$$

$$= \frac{P}{N^*}\sum_{i=1}^{N^*}\sum_{j \in A_i}\left(M_{ij}\bar{X}_j - \frac{X}{N^*}\right)^2 + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2$$

$$= P\,\sigma^2_{\text{NS1}^{\bullet}}(P) + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2 \qquad \text{(A.2)}$$

Substitute (A.2) for the first term on the right side of (A.1).

$$\sigma^2_{\mathrm{NS1}}(P) = P\,\sigma^2_{\mathrm{NS1}^*} + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2 + (1-P)\left(\frac{X}{N}\right)^2$$

$$= P\,\sigma^2_{\mathrm{NS1}^*}(P) + \sigma^2(P)\,E^2_{\mathrm{NS1}^*} \qquad \text{(A.3)}$$

where

$$\sigma^2\,(P) = P\,(1-P), \text{ and } E^2_{\mathrm{NS1}^*} = \left(\frac{X}{N^*}\right)^2.$$

REFERENCES

COHEN, S.B. (1998). Sample design of the 1996 medical expenditure panel survey medical provider component. *Journal of Economic and Social Measurement*. 24, 25-53.

DICKER, M., and SUNSHINE, J.H. (1987). Family use of health care, United States, 1980. *National Health Care Utilization and Expenditure Survey*. Report No. 10. DHHS Pub. 87-20210.

JUDKINS, D., BERK, M., EDWARDS, S., MOHR, P., STEWART, K. and WAKSBERG, J. (1995). National Health Care Survey: List verses Network Sampling, Unpublished report. National Center for Health Statstics.

JUDKINS, D., MARKER, D., WAKSBERG, J., BOTMAN, S. and MASSEY, J. (1999). National Health Interview Survey: Research for the 1995-2004 redesign. National Center for Health Statitics. *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2. 126, 76-89.

KISH, L. (1982). Design effect. *Encyclopedia of the Statistical Sciences*. John Wiley & Sons, Inc. 2, 347-348.

LEAVER, S., and VALLIANT, R. (1995). Statistical problems in estimating the U.S. consumer price index. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A Christianson, M.J. Colledge, and P.S. Kott). New York: John Wiley & Sons, Inc.

MASSEY, L.T., MOORE, T.F., PARSONS, V. and TADRO, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2, 110.

SIRKEN, M., and SHIMIZU, I. (1999). Population based establishment surveys: The Horvitz-Thompson estimator. *Survey Methodology*. 25, 187-91.

SIRKEN, M., SHIMIZU, I. and JUDKINS, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 1, 470-473.

SIRKEN, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics*. John Wiley & Sons, Inc. 4, 2977-2986.

SIRKEN, M.G. (2001). The Hansen-Hurwitz estimator revisited: PPS sampling without replacement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. In print.

THOMPSON, S. (1992). *Sampling*. New York: John Wiley & Sons, Inc. 117-118.

WUNDERLICH, G.S. (Ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. Washington, DC: National Academy Press.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 191-198
Statistics Canada

A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys

LOUIS-PAUL RIVEST[1]

ABSTRACT

This paper suggests stratification algorithms that account for a discrepancy between the stratification variable and the study variable when planning a stratified survey design. Two models are proposed for the change between these two variables. One is a log-linear regression model; the other postulates that the study variable and the stratification variable coincide for most units, and that large discrepancies occur for some units. Then, the Lavallée and Hidiroglou (1988) stratification algorithm is modified to incorporate these models in the determination of the optimal sample sizes and of the optimal stratum boundaries for a stratified sampling design. An example illustrates the performance of the new stratification algorithm. A discussion of the numerical implementation of this algorithm is also presented.

KEY WORDS: Neyman allocation; Power allocation; Stratified random sampling.

1. INTRODUCTION

The construction of stratified sampling designs has a long history in the statistical sciences. Chapters 5 and 5A in Cochran (1977) review several techniques for splitting a population into strata. The construction of strata is a topic of current interest in the statistical literature. Recent contributions include Hedlin (2000) who revisits Ekman (1959) rule for stratification, and Dorfman and Valiant (2000) who compare model-based stratification with balanced sampling. Model based stratification, is discussed in Godfrey, Roshwalb, and Wright (1984) and in chapter 12 of Särndal, Swensson, and Wretman (1992).

In business surveys, populations have skewed distributions; a small number of units accounts for a large share of the total of the study variable. It is therefore appropriate to include all large units in the sample (Dalenius 1952; Glasser 1962). A good sampling design has one take-all stratum for big firms, where the units are all sampled, together with take-some strata for businesses of medium and small sizes. Typically the sampling fraction goes down with the size of the unit; small businesses get large sampling weights. The Lavallée and Hidiroglou (1988) stratification algorithm is often used to determine the stratum boundaries and the stratum sample sizes in this context (see for instance Slanta and Krenzke 1994, 1996). This algorithm uses a stratification variable, known for all the units of the population. It gives the stratum boundaries and the stratum sample sizes that minimize the total sample size required to achieve a target level of precision. It uses an iterative procedure, due to Sethi (1963), to determine the optimal stratum boundaries. The Lavallée and Hidiroglou algorithm does not account for a difference between the stratification and the survey variables. As time goes by, this difference increases and the sampling design provided by the Lavallée and Hidiroglou algorithm may fail to meet the precision criterion.

Stratification in situations where the survey variable and the stratification variable differ is considered in Dalenius and Gurney (1951), see also Cochran (1977, chapter 5A). Many authors have studied approximate formulae for determining stratum boundaries, and for evaluating the gain in precision resulting from stratification on an auxiliary variable. Some relevant contributions are Serfling (1968), Singh and Sukatme (1969), Singh (1971), Singh and Parkash (1975), Anderson, Kish and Cornell (1976), Oslo (1976), Wang and Aggarwal (1984) and Yavada and Singh (1984). Hidiroglou and Srinath (1993) and Hidiroglou (1994) suggest techniques to update stratum boundaries using a new stratification variable. However these papers do not explicitly provide stratification algorithms accounting for the discrepancy between the stratification variable and the survey variable. This paper fills this gap by constructing generalizations of the Lavallée and Hidiroglou (1988) algorithm that express the difference between these two variables in terms of a statistical model.

A brief review of stratified sampling and of sample allocation methods is first given. Models for the difference between stratification and survey variables are then proposed. The implementation of Sethi's algorithm, when the stratification and the survey variable differ, is then presented. Numerical illustrations are provided.

2. A REVIEW OF STRATIFIED RANDOM SAMPLING

Some of the standard notation of stratified random sampling that will be used in this paper is

L = the number of strata;

[1] Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4.

$W_h = N_h/N$ is for $h = 1, ..., L$ the relative weight of stratum h, N_h is the size of stratum h, and $N = \sum N_h$ is the total population size;

n_h is for $h = 1, ..., L$ the sample size in stratum h and $f_h = n_h/N_h$ is the sampling fraction;

$\bar{Y}_h$ and $\bar{y}_h$ are the population and sample means of Y within stratum h;

S_{yh} is the population standard deviation of Y within stratum h.

In this paper the strata are constructed using X, a stratification variable. Stratum h consists of all units with an X-value in the interval $(b_{h-1}, b_h]$, where $-\infty = b_0 < b_1 < ... < b_{L-1} < b_L = \infty$ are the stratum boundaries. The survey estimator for $\bar{Y}$ can be expressed as $\bar{y}_{st} = \sum W_h \bar{y}_h$; its variance is given by:

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2. \tag{2.1}$$

In business surveys, all the big firms are sampled; we choose stratum L as the take-all stratum so that $n_L = N_L$. For $h < L$, n_h, the sample size in take-some stratum h, can be written as $(n - N_L)a_h$ where n is the total sample size and a_h depends on the allocation rule. The two allocation rules that are considered in this paper are

- The power allocation rule

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^{L-1} (W_k \bar{Y}_k)^p} \tag{2.2}$$

where p is a positive number in $(0, 1]$;

- The Neyman allocation rule

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}}. \tag{2.3}$$

Solving (2.1) for n leads to

$$n = N W_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2 / a_h}{\mathrm{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2 / N}. \tag{2.4}$$

The optimal stratum boundaries are the values of $b_1, ..., b_{L-1}$ that minimize n subject to a requirement on the precision of $\bar{y}_{st}$ such as $\mathrm{Var}(\bar{y}_{st}) = \bar{Y}^2 c^2$ where c is the target coefficient of variation (CV). The range $c = 1\%$ to 10% is often used for business surveys.

3. SOME MODELS FOR THE DISCREPANCY BETWEEN THE STRATIFICATION AND THE SURVEY VARIABLE

In this section $\{x_i, i = 1, ..., N\}$ denotes the known stratification variable for the N units in the population. Many stratification algorithms, including Lavallée and Hidiroglou, suppose that $\{x_i, i = 1, ..., N\}$ also represents the values of the study variable. This section suggests statistical models to account for a difference between these two variables.

For the sequel, it is convenient to look at X and Y as continuous random variables and to let $f(x)$, $x \in R$ denote the density of X. The data $\{x_i, i = 1, ..., N\}$ can be viewed as N independent realizations of the random variable X. Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h, for $h = 1, ..., L-1$. Three models for the difference between X and Y are next given along with their conditional means and variances for Y.

3.1 A Log-linear Model

The first model considers that $\log(Y) = \alpha + \beta_{\log} \log(X) + \varepsilon$, where ε is a normal random variable with mean 0 and variance $\sigma_{\log}^2$, which is independent from X, and α and $\beta_{\log}$ are parameters to be determined. When $\alpha = 0$, $\beta_{\log} = 1$ and $\sigma_{\log}^2 = 0$, one has $X = Y$; the survey and the stratification variables are the same. In general, $Y = e^{\alpha} X^{\beta_{\log}} e^{\varepsilon}$. The conditional moments of Y can be evaluated using the basic properties of the lognormal distribution (see Johnson and Kotz 1970), that is

$$E(e^{\varepsilon}) = e^{\sigma_{\log}^2/2} \text{ and } \mathrm{Var}(e^{\varepsilon}) = e^{\sigma_{\log}^2}(e^{\sigma_{\log}^2} - 1).$$

One has

$$E(Y|b_h \geq X > b_{h-1}) = \exp(\alpha + \sigma_{\log}^2/2)\, E(X^{\beta_{\log}}|b_h \geq X > b_{h-1})$$

while $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ is equal to

$$\mathrm{Var}(E(Y|X)\,|b_h \geq X > b_{h-1}) + E(\mathrm{Var}(Y|X)|b_h \geq X > b_{h-1})$$

$$= \exp(2\alpha + \sigma_{\log}^2)\{\mathrm{Var}(X^{\beta_{\log}}|b_h \geq X > b_{h-1})$$

$$+ (e^{\sigma_{\log}^2} - 1) E(X^{2\beta_{\log}}|b_h \geq X > b_{h-1})\}$$

$$= \exp(2\alpha + \sigma_{\log}^2)\{e^{\sigma_{\log}^2} E(X^{2\beta_{\log}}|b_h \geq X > b_{h-1})$$

$$- E(X^{\beta_{\log}}|b_h \geq X > b_{h-1})^2\}.$$

The parameter values $\beta_{\log}$ and $\sigma_{\log}$ can sometimes be calculated from historical data. Simple ad hoc values are $\beta_{\log} = 1$ and $\sigma^2_{\log} = (1 - \rho^2)\,\mathrm{Var}(\log(X))$. Here ρ is the assumed correlation between $\log(X)$ and $\log(Y)$. It can be set equal to predetermined values such as 0.95 or 0.99.

3.2 A Linear Model

In the survey sampling literature, the discrepancy between Y and X is often modeled with a heteroscedastic linear model,

$$Y = \beta_{\mathrm{lin}} X + \varepsilon, \tag{3.5}$$

where the conditional distribution of ε, given X, has mean 0 and variance $\sigma^2_{\mathrm{lin}}\ X^{\gamma}$, for some non negative parameter γ. Straightforward calculations lead to $E(Y|b_h \geq X > b_{h-1}) = \beta_{\mathrm{lin}}$ $E(X|b_h \geq X > b_{h-1})$ while $\mathrm{Var}(Y|b_h \geq X > b_{h-1}) = \beta^2_{\mathrm{lin}}$ $\{\mathrm{Var}(X|b_h \geq X > b_{h-1}) + (\sigma_{\mathrm{lin}}/\beta_{\mathrm{lin}})^2 E(X^{\gamma}|b_h \geq X > b_{h-1})\}$.

For an arbitrary $\gamma \geq 0$, the conditional variance of Y depends on three conditional moments of X. The generalization of Sethi's algorithm presented in section 5 does not work in this situation. Note however that when $\gamma = 2$, the conditional mean and variance of Y are proportional to those for the log-linear model with

$$\beta_{\log} = 1 \text{ and } \sigma^2_{\log} = \log(1 + (\sigma_{\mathrm{lin}}/\beta_{\mathrm{lin}})^2); \tag{3.6}$$

the proportionality factors are $\exp(\alpha + \sigma^2_{\log}/2)/\beta_{\mathrm{lin}}$ and $\exp(2\alpha + \sigma^2_{\log})/\beta^2_{\mathrm{lin}}$ for the conditional expectations and the conditional variances respectively. Thus the two models for the discrepancy between the stratification and the survey variable, either the log-linear model of section 3.1 or the linear model (3.5) with parameter $\gamma = 2$, lead, in section 5, to the same stratified design provided that (3.6) holds. In the later sections, the log-linear model is used to represent the change between X and Y. It should give good results when the true relationship between Y and X is modeled by (3.5) with $\gamma \approx 2$. When model (3.5) is assumed to hold with a smaller value of γ, the algorithm of section 5 can still be implemented when γ is set to either 0 or 1. This is however not pursued in this paper.

3.3 A Random Replacement Model

This model assumes that the stratification variable is equal to the survey variable, *i.e.*, $X = Y$, for most units. There is however a small probability ε that a unit changed drastically; its Y value then has $f(x)$ as density and is distributed independently of its X value. This is the approach used in Rivest (1999) to model the occurrence of stratum jumpers for which X is not representative of Y. More formally, this can be written as,

$$Y = \begin{cases} X \text{ with probability } 1 - \varepsilon \\ X_{\mathrm{new}} \text{ with probability } \varepsilon \end{cases},$$

where X_{new} represents a random variable with density $f(x)$ distributed independently of X. The conditional mean for Y under this model is given by

$$E(Y|b_h \geq X > b_{h-1}) = (1-\varepsilon)\ E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X),$$

while its conditional variance is equal to

$$\begin{aligned} &\mathrm{Var}(Y|b_h \geq X > b_{h-1}) \\ &\quad = (1-\varepsilon)E(X^2|b_h \geq X > b_{h-1}) + \varepsilon E(X^2) \\ &\quad - \{(1-\varepsilon)E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X)\}^2. \end{aligned}$$

4. AN EXAMPLE

Before addressing the technical details underlying the construction of the algorithms, it is convenient to look at an example. Consider the MU284 population of Särndal, Swensson and Wretman (1992), presenting data on 284 Swedish municipalities.

To build a stratified design for estimating the average of RMT85, the revenues from the 1985 municipal taxation, REV84, the real estate value according to 1984 assessment, is used as a stratification variable. One takes $L = 5$ and set the target CV at 5%. Two stratified designs obtained with the Lavallée and Hidiroglou algorithm are given in Table 1, for the power allocation with $p = 0.7$ and the Neyman allocation. Both have $n = 19$. When applied on survey variable RMT85, these two designs give estimators of total revenue with coefficients of variation of 8.3% and 7.3% respectively. Failing to account for a change between the survey and the stratification variables yields estimators that are more variable than expected.

Table 1
Stratified designs obtained with the Lavallée and Hidiroglou algorithm for the MU284 population using REV84 as stratification variable and a target CV of 5%

Power allocation with p = 0.7							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,251	874	56,250	86	1	0.01	19
stratum 2	2,352	1,696	100,898	82	2	0.02	19
stratum 3	4,603	3,114	351,547	65	3	0.05	19
stratum 4	10,606	6,442	2,027,436	41	3	0.07	19
stratum 5	59,878	19,631	275,502,518	10	10	1	19
Neyman allocation							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,273	878	57,260	87	2	0.02	19
stratum 2	2,336	1,701	99,688	81	2	0.02	19
stratum 3	4,619	3,114	351,547	65	3	0.05	19
stratum 4	11,776	6,921	3,724,610	46	7	0.15	19
stratum 5	59,878	28,418	426,851,844	5	5	1	19

To model the discrepancy between REV84 and RMT85, we use the log-linear model of section 3.1. There are outliers in the linear regression of log(RMT85) on log(REV84); they make the least squares estimates of $\beta_{\log}$ and $\sigma_{\log}$ unrepresentative of the relationship between the two variables. Robust estimates obtained with the Splus function lmRobMM are used instead. They are given by $\hat{\beta}_{\log} = 1.1$ and $\hat{\sigma}_{\log} = 0.2116$. Table 2 gives the stratified designs obtained with the generalized Lavallée and Hidiroglou algorithm for two allocation rules. They both give estimators of the total of RMT85 having a CV of 5.7%. This CV is still larger that 5%. Since there are outliers in the log-linear regression, the assumption of normal errors made in section 3.1 is not met. This might explain the failure to reach the target CV exactly. The increase in sample size for $n = 19$ to $n = 28$ is noteworthy! For both allocation methods the design obtained using the log-linear model has smaller take-all strata than Lavallée and Hidiroglou.

Table 2

Stratified designs obtained with the generalized Lavallée and Hidiroglou algorithm for the MU284 population using REV84 as stratification variable, a log-linear with $\beta_{\log} = 1.1$ and $\sigma_{\log} = 0.2116$ for the discrepancy between REV84 and RMT85, and a target CV of 5%

Log-linear model stratification algorithm with power allocation with $p = 0.7$							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,558	1,023	97,245	121	4	0.03	28
stratum 2	3,031	2,219	168,204	81	5	0.06	28
stratum 3	5,706	4,022	464,471	44	6	0.14	28
stratum 4	11,107	7,602	2,659,061	32	7	0.22	28
stratum 5	59,878	25,536	39,131,413	6	6	1	28
Log-linear model stratification algorithm with Neyman allocation							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,582	1,023	97,245	121	4	0.03	28
stratum 2	3,040	2,219	168,204	81	5	0.06	28
stratum 3	5,608	4,022	464,471	44	5	0.11	28
stratum 4	11,476	7,709	2,952,313	33	9	0.27	28
stratum 5	59,878	28,418	426,851,844	5	5	1	28

An alternative to the generalized Lavallée and Hidiroglou algorithm for the construction of stratified designs is to us their original algorithm with a smaller target CV. This increases the sample size thereby reducing the variance of the estimator of the total of the survey variable. When constructing a design for RMT85 using REV84 as a stratification variable, the standard Lavallée and Hidiroglou algorithm with power allocation rule ($p = 0.7$) and a target CV of 3.6%, yields a stratified design with $n = 28$. This design has the same sample size as those presented in Table 2. The CV of the estimator of the total RMT85 is 5.7%, the same as the CVs obtained with the designs of Table 2. The main difference between these designs is the size of the take-all stratum. The design constructed with the Lavallée and Hidiroglou algorithm has a take-all stratum of size $N_5 = 13$ as compared to $N_5 = 5$ and $N_5 = 6$ for the designs of Table 2. Allowing the stratification and the survey variables to differ appears to reduce the relative importance of the take-all stratum in the sampling design. Further investigations are needed to ascertain this hypothesis.

The stratification algorithm for the random replacement model of section 3.3 (with Neyman allocation) was also applied to REV84. Assuming changes in 2% of the units ($\varepsilon = 0.02$), the generalized Lavallée and Hidiroglou algorithm yields a stratified design with $n = 37$ sample units; the resulting estimator of total RMT85 has a CV of 5.5%. An interesting property of this stratified design is that the smallest sampling fraction is $\min_h f_h = 9.3\%$; it is much larger than $\min_h f_h$ for the designs of Tables 1 and 2. Despite the presence of outliers, the random replacement model does not describe the changes between REV84 and RMT85 as well as the log-linear model. This explains why a larger sample size, 37 instead of 28, is needed to get an estimator with a variance comparable to that obtained with the stratification based on a log-linear model.

5. A METHOD FOR CONSTRUCTING STRATIFICATION ALGORITHMS

The aim of a stratification algorithm is to determine the optimal stratum boundaries and sample sizes for sampling Y using the known values $\{x_i; i = 1, ..., N\}$ of variable X for all the units in the population. A model, such as those given in section 3, characterizes the relationship between X and Y. This section extends the stratification algorithm of Lavallée and Hidiroglou (1988) to situations where X and Y differ. It uses the log-linear model of section 3.1 to account for the differences between Y and X. Modifications to handle the random replacement model are easily carried out (see Rivest 1999).

5.1 A Generalization of Sethi's (1963) Stratification Method

It is convenient to consider an infinite population analogue to equation (2.4) for n. Since the random variable X has a density $f(x)$, the first two conditional moments of Y given that $b_{h-1} < X \le b_h$ can be written in terms of

$$W_h = \int_{b_{h-1}}^{b_h} f(x)\,dx, \varphi_h = \int_{b_{h-1}}^{b_h} \alpha^{\beta} f(x)\,dx,$$

$$\text{and } \psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x)\,dx,$$

where β is the slope of the log-linear model given in section 3.1 (in this section β and σ represent parameters of the log-linear model of section 3.1, since there is no risk of confusion the subscript log is not used anymore). For stratification purposes, it is useful to rewrite (2.4) in terms of the conditional means and variances for Y,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \operatorname{Var}(Y|b_h \geq X > b_{h-1})/a_{h,X}}{\bar{Y}^2c^2 + \sum_{h=1}^{L-1} W_h \operatorname{Var}(Y|b_h \geq X > b_{h-1})/N}, \quad (5.7)$$

where $a_{h,X}$ denotes the allocation rule written in terms of the known X. For instance, under power allocation,

$$a_{h,X} = \frac{\{W_h E(Y|b_h \geq X > b_{h-1})\}^p}{\sum_{k=1}^{L-1} \{W_k E(Y|b_k \geq X > b_{k-1})\}^p},$$

for $h = 1, \ldots, L-1$. Given a model for the relationship between Y and X, $\operatorname{Var}(Y | b_h \geq X > b_{h-1})$ and $E\ (Y | b_h \geq X > b_{h-1})$ can be written in terms of W_h, φ_h, and ψ_h. Thus, the partial derivatives of n with respect to b_h can be evaluated, for $h < L-1$, using the chain rule,

$$\frac{\partial n}{\partial b_h} = \frac{\partial n}{\partial W_h}\frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \varphi_h}\frac{\partial \varphi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h}\frac{\partial \psi_h}{\partial b_h}$$

$$+ \frac{\partial n}{\partial W_{h+1}}\frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \varphi_{h+1}}\frac{\partial \varphi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}}\frac{\partial \psi_{h+1}}{\partial b_h}$$

Observe that

$$\frac{\partial W_h}{\partial b_h} = -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h)$$

$$\frac{\partial \varphi_h}{\partial b_h} = -\frac{\partial \varphi_{h+1}}{\partial b_h} = b_h^\beta f(b_h)$$

$$\frac{\partial \psi_h}{\partial b_h} = -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} f(b_h)$$

This leads to the following result, for $h < L-1$,

$$\frac{\partial n}{\partial b_h} = f(b_h)$$

$$\left\{\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right) + \left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}}\right) b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right) b_h^{2\beta}\right\}.$$

Similarly,

$$\frac{\partial n}{\partial b_{L-1}} = f(b_{L-1})\left\{-N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \varphi_{L-1}} b_{L-1}^\beta + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^{2\beta}\right\}.$$

The Sethi's (1963) algorithm is used to solve $\partial n / \partial b_h = 0$. It considers that the partial derivatives are proportional to quadratic functions in b_h^β. The updated value for b_h^β is given by the largest root of the corresponding quadratic function. When $h < L-1$, this gives

$$b_h^{\beta\,\text{new}} =$$

$$-\left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}}\right) \Big/ \left\{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\right\}$$

$$+ \frac{\left\{\left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}}\right)^2 - 4\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right)\right\}^{1/2}}{\left\{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\right\}},$$

while for $h = L-1$ we have

$$b_{L-1}^{\beta\,\text{new}} = \frac{-\frac{\partial n}{\partial \varphi_{L-1}} + \left\{\left(\frac{\partial n}{\partial \varphi_{L-1}}\right)^2 - 4\frac{\partial n}{\partial \psi_{L-1}}\left(\frac{\partial n}{\partial W_{L-1}} - N\right)\right\}^{1/2}}{\left(2\frac{\partial n}{\partial \psi_{L-1}}\right)}$$

The partial derivatives of n with respect to W_h, φ_h, and ψ_h depend on moments of order 0, 1, and 2 of x^β within stratum h. These moments are evaluated using the N x-values in the population. For instance,

$$\varphi_h = \frac{1}{N} \sum_{i: b_{h-1} < x_i \leq b_h} x_i^\beta.$$

Applications of this general method are provided next.

When using Sethi's algorithm, one typically has $L \geq 3$. Note however that it also works when $L = 2$. In this case, the algorithm is searching for the boundary between a take-all and a take-some stratum. Successive evaluations of $b_{L-1}^{\beta\,\text{new}}$ presented above yield an optimal boundary. When one assumes that the stratification and the study variable coincide, *i.e.*, $X = Y$, this boundary is nearly identical to that obtained with the algorithm presented in Hidiroglou (1986).

5.2 An Algorithm for Power Allocation

For the log-linear model of section 3.1, the conditional expectation is $E(Y|b_h \geq X > b_{h-1}) = C\varphi_h / W_h$ while the conditional variance is

$$\operatorname{Var}(Y|b_h \geq X > b_{h-1}) = C^2\{e^{\sigma^2}\psi_h / W_h - (\varphi_h / W_h)^2\},$$

where $C = \exp(\alpha + \sigma^2/2)$. Under the power allocation rule, $a_{h,X} = \varphi_h^p / \sum_{h=1}^{L-1} \varphi_k^p$, and formula (5.7) for n becomes

$$n = NW_L + \frac{\sum_{h=1}^{L-1} \varphi_h^p \sum_{h=1}^{L-1} (e^{\sigma^2} W_h \psi_h - \varphi_h^2) / \varphi_h^p}{\left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} \left(e^{\sigma^2} \psi_h - \varphi_h^2 / W_h\right) / N}.$$

The partial derivatives needed to implement the stratification algorithm are easily calculated; for $h \leq L-1$,

$$\frac{\partial n}{\partial W_h} = \frac{A e^{\sigma^2} \psi_h / \varphi_h^p}{F} - \frac{AB(\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{A\{-pe(\sigma^2 W_h \psi_h - \varphi_h^2)/\varphi_h^{p+1} - 2/\varphi_h^{p-1}\} + p\varphi_h^{p-1} B}{F}$$

$$+ 2\frac{AB\varphi_h/(n W_H)}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = e^{\sigma^2} \frac{A W_h / \varphi_h^p}{F} - e^{\sigma^2} \frac{AB/N}{F^2},$$

where

$$A = \sum_{h=1}^{L-1} \varphi_h^p, \ B = \sum_{h=1}^{L-1} \left(e^{\sigma^2} W_h \psi_h - \varphi_h^2\right) \Big/ \varphi_h^p,$$

and

$$F = \left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} \left(e^{\sigma^2} \psi_h - \varphi_h^2 / W_h\right) / N.$$

5.3 An algorithm for Neyman allocation

Under Neyman allocation, allocation rule (2.3) written in terms of W_h, φ_h, and ψ_h is

$$a_{h,X} = \frac{\left\{e^{\sigma^2} \psi_h W_h - \varphi_h^2\right\}^{1/2}}{\sum_{h=1}^{L-1} \left\{e^{\sigma^2} \psi_h W_h - \varphi_h^2\right\}^{1/2}}$$

and the formula for n is

$$n = NW_L + \frac{\left\{\sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h W_h - \varphi_h^2)^{1/2}\right\}^2}{\left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \varphi_h^2 / W_h)/N}.$$

The partial derivatives needed to implement Sethi's (1963) iterative algorithm are,

$$\frac{\partial n}{\partial W_h} = \frac{A e^{\sigma^2} \psi_h / (e^{\sigma^2} \psi_h W_h - \varphi_h^2)^{1/2}}{F} - \frac{A^2 (\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{-2A\varphi_h / \{e^{\sigma^2} W_h \psi_h - \varphi_h^2\}^{1/2}}{F} + \frac{2A^2 \varphi_h / (W_h N)}{F^2}$$

$$\frac{\partial n}{\partial \psi_h} = \frac{e^{\sigma^2} A W_h / \{e^{\sigma^2} W_h \psi_h - \varphi_h^2\}^{1/2}}{F} - e^{\sigma^2} \frac{A^2/N}{F^2},$$

where

$$A = \sum_{h=1}^{L-1} \left(e^{\sigma^2} \psi_h W_h - \varphi_h^2\right)^{1/2},$$

and

$$F = \left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} \left(e^{\sigma^2} \psi_h - \varphi_h^2 / W_h\right) \Big/ N.$$

6. NUMERICAL CONSIDERATIONS

Slanta and Krenzke (1994, 1996) encountered numerical difficulties when using the Lavallée and Hidiroglou algorithm with Neyman allocation: convergence was slow and sometimes the algorithm did not converge to the true minimum value for n. Indeed Schneeberger (1979) and Slanta and Krenzke (1994) showed that, for a particular bimodal population, the problem has a saddle; that is the partial derivatives are all null at boundaries b_h which do not give a true minimum for n.

When using the algorithms constructed in this paper, we also experienced the numerical difficulties alluded to in Slanta and Krenzke (1994). The algorithms constructed under power allocation were generally more stable than those using Neyman allocation; numerical difficulties were more frequent when the number L of strata was large. Furthermore, as the distribution for Y moved away from that of X, *i.e.*, as σ^2 increases, non convergence of the algorithm and failure to reach the global minimum for n were more frequent. In these situations, the stratification algorithm's starting values were of paramount importance. For instance, in Table 2, the design accounting for changes between Y and X obtained under Neyman allocation depends heavily on the starting values. The one presented in Table 2 uses the boundaries presented in Table 2 for the power allocation as starting values. Starting the algorithm with the boundaries obtained in Table 1 for the Lavallée Hidiroglou algorithm with Neyman allocation yields a different sampling design having $n = 29$.

A good numerical strategy is to run the stratification algorithm for several intermediate designs to get to a final sampling design, with the stratum boundaries obtained at one step used as starting values for the algorithm at the next step. The log-linear algorithm is always run in two steps; first run the Lavallée and Hidiroglou algorithm, setting $\sigma = 0$, and use these boundaries as starting value for the algorithm with a non null σ. Also use as starting value for Neyman allocation the corresponding boundaries found under power allocation with a p value around 0.7.

7. CONCLUSION

This paper has proposed generalizations of the Lavallée and Hidiroglou stratification algorithm that account for a difference between the stratification and the survey variables. Two statistical models have been introduced for this purpose. The new class of algorithms uses the Chain Rule to derive partial derivatives and Sethi's (1963) technique to find the optimal stratum boundaries.

The log-linear model stratification algorithm proposed in this paper was used successfully in several surveys designed at the Statistical Consulting Unit of Université Laval. For estimating total maple syrup production in a year, the number of sap producing maples for a producer was a convenient size variable. Historical data was used to estimate the parameters of the log-linear model linking sap producing maples and production volume. Another example is the estimation of the total maintenance deficit of hospital buildings in Quebec. The value of each building was the known stratification variable. The maintenance deficit was estimated to be in the range (20%, 40%) by experts. Solving $4\sigma_{\log} = \log(40\%) - \log(20\%)$ gives $\sigma_{\log} = \log(2)/4 = 0.17$ as a possible parameter value for the log-linear model of section 3.1. In these two examples accounting for changes between the stratification and the survey variables increased the sample size n by a fair percentage and yielded survey estimators whose estimated CVs were close to the target CVs.

Two SAS IML functions implementing the algorithm presented in this paper, for power and Neyman allocation, are available on the author's website at http: //www.mat. ulaval.ca/pages/lpr/. They allow user specified starting values for the stratum boundaries; they can be used to implement the numerical strategies presented in section 6.

ACKNOWLEDGMENTS

I am grateful to Nathalie Vandal and to Gaétan Daigle for constructing SAS IML programs for the stratification algorithms used in the paper. The constructive comments of the associate editor and of the referee are gratefully acknowledged.

REFERENCES

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

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

DALENIUS, T. (1952). The Problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*. 35, 61-70.

DALENIUS, T., and GURNEY, M. (1951). The Problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*. 34, 133-148.

DORFMAN, A.H., and VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*. 16, 139-154.

ECKMAN, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*. 30, 219-229.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statisticasl Institute*. 30, 28-32.

GODFREY, J., ROSHWALB, A. and WRIGHT, R.L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*. 2, 1-9.

HEDLIN, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*. 16, 15-29.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*. 40, 27-31.

HIDIROGLOU, M. (1994). Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 153-162.

HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*. 11, 397-405.

JOHNSON, N.L., and KOTZ, S. (1970). *Continuous Univariate Distribution-1*. New York: John Wiley & Sons, Inc.

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

OSLO, I.T. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*. 23, 15-25.

RIVEST, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 64-72.

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

SCHNEEBERGER, H. (1979). Saddle points of the variance of the sample mean in stratified sampling. *Sankhyā: The Indian Journal of Statistics, Series C*. 41, 92-96.

SERFLING, R.J. (1968). Approximate optimal stratification. *Journal of the American Statistical Association*. 63, 1298-1309.

SETHI, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*. 5, 20-33.

SINGH, R.J. (1971). Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*. 66, 829-834.

SINGH, R., and PARKASH, D. (1975). Optimal stratification for equal allocation. *Annals of the Institute of Statistical Mathematics*. 27, 273-280.

SINGH, R., and SUKATME, B.V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*. 21, 515-528.

SLANTA, J., and KRENZKE, T. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 693-698.

SLANTA, J., and KRENZKE, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Survey Methodology.* 22, 65-75.

WANG, M.C., and AGGARWAL, V. (1984). Stratification under a particular Pareto distribution. *Communications in Statistics, Part A – Theory and Methods.* 13, 711-735.

YAVADA, S., and SINGH, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics, Part A – Theory and Methods.* 13, 2793-2806.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 199-207
Statistics Canada

Multi-way Stratification by Linear Programming Made Practical

WILSON LU and RANDY R. SITTER[1]

ABSTRACT

Sitter and Skinner (1994) present a method which applies linear programming to designing surveys with multi-way stratification, primarily in situations where the desired sample size is less than or only slightly larger than the total number of stratification cells. The idea in their approach is simple, easily understood and easy to apply. However, the main practical constraint of their approach is that it rapidly becomes expensive in terms of magnitude of computation as the number of cells in the multi-way stratification increases, to the extent that it cannot be used in most realistic situations. In this article, we extend this linear programming approach and develop methods to reduce the amount of computation so that very large problems become feasible.

KEY WORDS: PPS sampling; Proportional allocation; Random grouping; Survey sampling.

1. INTRODUCTION

In many practical survey situations, there are multiple stratifying variables available and thus the designer has the option of defining strata as cells formed as cross-classified categories of these variables. For examples, see Engle, Marsden and Pollock (1971), Hess, Riedel and Fitzpatrick (1976), Vihma (1981) and Skinner, Holmes and Holt (1994). This multi-way stratification often leads to situations where the desired sample size is less than or only slightly larger than the total number of stratification cells (particularly common when choosing primary sampling units (psu's) in stratified multi-stage designs) and hence conventional methods of sample allocation to strata may not be applicable.

An illustration, based on a hypothetical example of Bryant, Hartley and Jessen (1960), is given in Table 1. Communities (psu's) are classified by two stratifying factors, type and region, with three and five categories respectively. The desired sample size of $n = 10$ is less than the total number of cells, 15. This example also illustrates a related problem. The entries in Table 1 are the expected counts under proportional stratification, *i.e.*, the strata sample sizes are proportional to the population strata sizes. Under the sample size restrictions, the expected cell sample counts will not generally be integers. In cases with very small expected counts, rounding to integers will not lead to good choices while causing a serious violation of the property of proportional allocation. Non-integer margin totals are also typical and can cause their own difficulties. Goodman and Kish (1950) was the first to address this problem under the name of controlled selection, where they propose a sampling selection procedure which can be classified as random sytematic sampling (see Hess, Riedel and Fitzpatrick 1976; Waterton 1983). Bryant *et al.* (1960) presented a very simple method to randomly assign sample sizes for each cell in two-way stratification and gave two estimators based on that sampling scheme. However, since the expected cell sample sizes didn't include information of proportion of each cell (*i.e.*, the method is not a proper controlled selection technique, as only the probabilities of the marginal distributions are respected), these estimators may not have satisfactory MSE properties (see Sitter and Skinner 1994). Jessen (1970) points out that a further limitation of the method of Bryant *et al.* (1960) is that it is not possible to constrain specified cell sizes to be zero, which may be desired in some situations (see related methods under the label "lattice sampling", *e.g.* Jessen 1973, 1975). He proposes two methods for both two-way and three-way stratification but both methods are fairly complicated to implement and, as noted by Causey, Cox and Ernst (1985), may not lead to a solution. Inspired by the idea of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples (see also Lahiri and Mukerjee 2000), Sitter and Skinner (1994) proposed a linear programming approach which attempts to take advantage of the power of modern computing. This linear programming technique is simple in conception, is flexible to different situations, always has a solution and has better properties of the MSE. Its main practical constraint is that it becomes computationally intensive as the number of cells in the multi-way stratification increases, quickly to the point of infeasibility. In this paper we will present a simple method which will allow the linear programming technique to handle much larger problems. In section 2 we describe the linear programming method of Sitter and Skinner (1994) to introduce notation and briefly discuss its numerical limitations. In section 3.1, we first discuss some simple strategies to reduce the computational intensity of the method as motivation for the eventual proposal. In sections 3.2 and 3.3 we discuss the proposed method assuming integer margins

[1] Wilson Lu, Doctoral Student, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

and give some examples with from 80 to 300 stratification cells to illustrate the ability of the new methodology to handle large problems. In section 3.4, we describe the simple extention of the method to non-integer margins and illustrate by applying the method to a real example from the occupational health literature (Vihma 1981).

Table 1
Example from Bryant *et al.* (1960). Expected Sample Cell Counts Under Proportional Stratification (n = 10)

Region	Type of Community			
	Urban	Rural	Metropolitan	Total
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

2. THE LINEAR PROGRAMMING TECHNIQUE

2.1 The Basic Ideas

We introduce the linear programming method of Sitter and Skinner (1994) by considering the simplest kind of two-way stratification. Suppose that N units of a finite population are arranged in a two-way classification in R rows formed by categories of one variable and C columns by categories of another. Let N_{ij} denote the number of population units in the i-th row and the j-th column (*i.e.*, in the ij-th cell) of the two-way table and $P_{ij} = N_{ij}/N$ denote the proportion of the total population in the ij-th cell. Let $\bar{Y}$ denote the mean value of a survey characteristic y for the population and $\bar{Y}_{ij}$ denote the mean value of y for the ij-th cell.

The sample is selected as follows:

i) Sample sizes n_{ij} are randomly determined for each cell according to a pre-specified procedure. Letting s denote the $R \times C$ array $(n_{ij}, i = 1, \ldots, R,\ j = 1, \ldots, C)$, this procedure assigns a probability $p(s)$ to each s in the set S of possible such arrays and selects a single array, s, from S. We denote the dependence of n_{ij} on s by writing $n_{ij}(s)$.

ii) A simple random sample of $n_{ij}(s)$ units is then selected from the ij-th cell and the values of y obtained.

Restrict attention to designs of fixed sample size $n > 0$, that is, restrict to arrays $s \in S_n$ such that $\sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}(s) = n$. We would also like to restrict attention to proportionate stratification so that

$$\sum_{s \in S_n} n_{ij}(s)\, p(s) = nP_{ij} \quad \text{for} \quad i = 1, \ldots, R,\ j = 1, \ldots, C, \quad (1)$$

which implies that the simple unweighted sample mean $\bar{y}(s)$ is an unbiased estimator of $\bar{Y}$. We will refer to (1) as the expected proportional allocation (EPA) constraint.

The linear programming technique of Sitter and Skinner (1994) chooses a sampling design $p(s)$ which minimizes the expected lack of 'desirability' of the samples by solving the linear programming problem:

$$\min \sum_{s \in S_n} w(s) p(s) \quad (2)$$

subject to the constraint (1), where $w(s)$ is a loss function for the sample s, to be specified, and the $p(s)$ are the unknowns. Sitter and Skinner (1994) were exploiting the key observation of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples, that the objective function in (2) was linear in the $p(s)$'s (see also Lahiri and Mukerjee 2000).

In the objective function (2), the loss function $w(s)$ plays an important role. With a well defined $w(s)$, we have flexibility to explore the existence of an optimal solution to (2) within an economically sized S_n and, more importantly, to improve efficiency of estimation. Sitter and Skinner (1994) suggest choosing

$$w(s) = \sum_{i=1}^{R} \left(n_{i\cdot}(s) - nP_{i\cdot}\right)^2 + \sum_{j=1}^{C} \left(n_{\cdot j}(s) - nP_{\cdot j}\right)^2, \quad (3)$$

where $n_{i\cdot}(s) = \sum_j n_{ij}(s)$, $n_{\cdot j}(s) = \sum_i n_{ij}(s)$, $P_{i\cdot} = \sum_j P_{ij}$ and $P_{\cdot j} = \sum_i P_{ij}$. Obviously, the objective function (2) is actually $E(w(s))$ for any given design $p(s)$ and can be explained as the mean squared error of estimator $\bar{y}$ under an analysis of variance model (see Sitter and Skinner 1994). Then by solving the above linear programming problem, one can obtain minimized MSE in the sense of ANOVA while maintaining the EPA property of the $n_{ij}(s)$. One should note that if a design with objective function equal to zero is obtained, then all margin constraints are met. This would typically only be the case with integer margins.

Sitter and Skinner (1994) suggest that one simple way to reduce the size of S_n is to restrict the actual values that n_{ij} can take to be either $\lfloor nP_{ij} \rfloor$ or $\lfloor nP_{ij} \rfloor + 1$, where $\lfloor nP_{ij} \rfloor$ is the greatest integer less than or equal to nP_{ij}. By denoting $\tilde{n}_{ij} = n_{ij} - \lfloor nP_{ij} \rfloor$ and $r_{ij} = nP_{ij} - \lfloor nP_{ij} \rfloor$, one can then impose

$$E(\tilde{n}_{ij}) = r_{ij}, \quad (4)$$

where $\tilde{n}_{ij} = 0$ or 1 and $0 \le r_{ij} < 1$. Then the linear programming method can be applied to the $\tilde{n}_{ij}$ and finally $\lfloor nP_{ij} \rfloor + \tilde{n}_{ij}$ can be used as the actual cell sample sizes. Therefore, without loss of generality, we will assume that

$$n_{ij} = 0, 1 \quad \text{and} \quad 0 \le r_{ij} = nP_{ij} < 1. \quad (5)$$

2.2 Higher-way Stratification

The Sitter and Skinner (1994) approach extends straightforwardly to more stratifying factors by letting s denote the corresponding r-way array. The loss function would then include more terms, for example for three-way stratification equation (3) could be replaced by

$$w(s) = \gamma_1 \sum_{i=1}^{R_1} (n_{i..}(s) - nP_{i..})^2 + \gamma_2 \sum_{j=1}^{R_2} (n_{.j.}(s) - nP_{.j.})^2 + \gamma_3 \sum_{k=1}^{R_3} (n_{..k}(s) - nP_{..k})^2$$

in obvious notation, where γ_1, γ_2 and γ_3 might represent the relative importance of balancing on the three factors based on prior information (see Sitter and Skinner 1994).

2.3 Multi-stage Sampling

An important application of multi-way stratification is to the selection of primary sampling units (psu's) in multi-stage sampling, where it is more common to have several stratifying factors available.

In section 2.1, the inclusion probabilities of each unit are $E(n_{ij}(s)/N_{ij}) = n/N$. If psu's are selected with equal probability then the approach extends directly with the psu's the units and with the observed values of y replaced by unbiased estimators of the psu totals. However, if the psu's are to be selected with unequal probabilities, say nz_{ijk} for psu k in stratification cell $ij(z_{ijk}$ will typically equal $M_{ijk}/\sum_{ijk} M_{ijk}$, with M_{ijk} being some measure of size of psu k in cell ij), then the procedure can be easily modified by setting P_{ij} equal to $z_{ij.}/z_{...}$, where $z_{ij.} = \sum_k z_{ijk}$ and $z_{...} = \sum_{ijk} z_{ijk}$. Then, if $n_{ij}(s) > 0$, a sample of $n_{ij}(s)$ psu's in cell ij is selected by some probability proportional to z_{ijk} method.

2.4 An Example

The linear programming approach can be illustrated using the hypothetical example of Bryant *et al.* (1960) given in Table 1. First, this problem is simplified as shown in Table 2 to meet the assumption in (5). Then, a standard linear programming package is used to solve this reduced problem (2). Because integer margins of expected sample cell counts can be exactly matched by marginal totals of sample sizes $n_{i.}$ and $n_{.j}$, which means that the loss function $w(s)$ can acheive a minimum value of zero, the objective function in (2) for this example is also minimized at zero. The optimal solution of this problem is given in Table 3. It should be noted that this solution has been converted back to match the original example shown in Table 1.

Table 2
Modified Example from Bryant *et al.* (1960)

Region	Type of Community			
	Urban	Rural	Metropolitan	Total
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

Table 3
Linear Programming Solution to Example from Bryant *et al.* (1960)

s			$p(s)$	s			$p(s)$	s			$p(s)$
1	1	0		1	1	0		1	1	0	
1	0	0		0	0	1		0	0	1	
0	1	1	0.2	0	1	1	0.1	0	0	2	0.2
0	2	1		1	1	1		1	2	0	
1	0	1		1	1	0		1	1	0	
1	0	1		1	0	1		1	0	1	
0	1	0		0	1	0		0	0	1	
1	0	1	0.2	0	1	1	0.1	0	1	1	0.2
0	2	1		1	1	1		1	2	0	
1	1	0		1	1	0		1	1	0	

The linear programming method is simple and easy to use. Its main drawback is computational. The number of parameters in the resulting linear programming problem is the number of samples of size n from the $RC > n$ cells, $\binom{RC}{n}$, which becomes infeasibly large quite quickly. In the next section we will explore ways of improving the computational efficiency of the linear programming approach while maintaining all of its good properties.

3. THE LINEAR PROGRAMMING APPROACH MADE PRACTICAL

The basic idea of the linear programming approach is to obtain an optimal sampling design in terms of the (minimum) expected lack of "desirability" of the sample by directly solving a linear programming problem with $p(s)$, $s \in S_n$, as the unknowns while maintaining the EPA property. The only obstacle to this approach is that the number of elements in S_n is often very large and even with modern computing power it becomes difficult to carry out linear programming if the number of unknowns is large.

To reduce the magnitude of the computational task for this linear programming problem determined by the cardinality of S_n, we want to obtain a subset of S_n, say S_{n0}, which is nearly as representative as S_n but much smaller, and thus solve the following linear programming problem with a much smaller set of $p(s)$, $s \in S_{n0}$, as the unknowns:

$$\min \sum_{s \in S_{n0}} w(s) p(s). \tag{6}$$

Hopefully, in this way we can easily deal with larger practical problems without losing the good properties of the linear programming approach.

3.1 Some Motivating Strategies

The above strategy is easy to state, but it turns out not to be entirely obvious how to go about it. In fact, there are several different directions we can explore to determine such a subset $S_{n0} \subset S_n$. In this section, we will describe a basic method related to loss functions which was alluded to in Sitter and Skinner (1994) and describe how it modestly increases the size of problems that can be handled. We will then discuss some obvious directions to take which did not improve things much. By describing these misguided attempts, we motivate the eventual proposal.

The major flexibility of the linear programming approach is derived from the choice of loss function $w(s)$. Thus, it is natural for us to consider the loss function first when we try to improve the computational efficiency of this approach. By observing the objective function of the linear programming problem (2), we suspect that the loss function $w(s)$ as coefficients of unknowns $p(s)$ will not be very large when the objective function has been minimized. In other words, all positive $p(s)$ in an optimal sampling design will only be assigned to samples having small lack of "desirability". Based on this observation, we hypothesize that the following subset might be a good replacement for S_n,

$$S_{n0} = \{ s \in S_n : w(s) = \sum_{i=1}^{R} (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^{C} (n_{\cdot j}(s) - nP_{\cdot j})^2 \leq w_0 \}, \quad (7)$$

where w_0 is a pre-determined positive constant. In the case of integer margins, one could even let $w_0 = 0$ and restrict to samples where the margins are matched. For example, the solution in Table 3 assigned positive probability to only 6 samples and for each of these the objective function was zero.

Lu (2000) develops nested linear programming strategies for solving this problem. For moderately sized problems such as 8 × 5 arrays (*i.e.*, 40 cells) this approach does well. However, for larger problems the size of resulting candidate sets becomes large very quickly, even in the integer margin case. Thus for large problems the technique faces the same problem as before-a huge candidate set that results in the difficulty of solving a linear programming problem with too many unknowns.

In reality, even a candidate sample set S_{n0} of the form in (7) is far larger than necessary for us to find an optimal solution. What we really need is a smaller but fairly representative subset, where by "small" we mean small enough to make it *possible* to solve the resulting linear programming problem and by "representative" we mean containing elements which promise that this linear programming problem is *feasible*.

Before going on to describe our eventual proposed solution to this problem, we would like to introduce some naive methods of obtaining such a "representative subset" that turned out not to work well. These are not that useful in practice, but they did inspire our thinking in proposing a more sophisticated approach.

1) Two Stage Optimization: First of all, we could try to break S_{n0} in (7) into many subsets which are small enough to be handled by linear programming respectively. Hopefully, optimal solutions from each of these smaller sets in the first stage optimization procedure can be combined to form the desired representative set of samples. Then we can just collect these optimal solutions together and apply linear programming once more. We applied this method to some simulated examples of size 6×6, 7×7, 8×8 and 9×9 as a method of preliminary investigation of its potential. Generally, in the first two cases the method worked very well and quickly, in the 8×8 case the method was time consuming and was not always able to obtain optimal solutions, and in the 9×9 case the method became infeasible.

2) Resampling from S_{n0}: We could also randomly select a proportion, say 10%, of the S_{n0} in (7) and hope this proportion is statistically representative of the complete set. Unfortunately, simulation results showed that the proportion obtained in this way is not "representative" enough, and the resulting linear programming problem often does not have any feasible solution. For example, the method of nested linear programming discussed previously was able to obtain matched integer margin solutions for simulated 8 × 5 arrays, however, these solutions were obtained much quicker by repeatedly sampling 10% of S_{n0} and applying the Sitter and Skinner (1994) method to this set until a feasible solution was obtained. However, when slightly larger cases were considered the method took an inordinate amount of time before finding a feasible solution, and quickly became impractical.

There are two problems with both these approaches. First, the size of S_{n0} becomes huge combinatorically and even complete enumeration becomes difficult. Having to first obtain S_{n0} and then cutting the problem into pieces will either quickly outstrip the practical limits on linear programming due to the size of the pieces or create a huge number of pieces. Second, both of these strategies are not in any way attempting to avoid samples which are particularly bad choices for meeting the EPA constraints. The question is, is there any way we can generate a fairly "representative" candidate sample subset without choosing such "useless" samples or, more generally, can we select candidate samples in which the frequency of an entry's appearance is more or less related to its desired expected sample counts?, and also can we do so without first having to enumerate a large S_{n0}? The general idea revolves around the fact that if we could randomly select a candidate subset directly from S_n without complete enumeration using an unequal probability selection procedure which simultaneously ensures that the objective function is minimized for every sample while ensuring that the EPA property is satisfied we will have solved the problem without resorting to linear programming at all. We have been working on finding such a selection procedure, but have yet to succeed. What we have been able to do is to develop such a procedure with approximate EPA (AEPA). We can then use it to randomly generate a candidate subset of samples, S_{n0}, and then apply a linear programming technique to this subset.

3.2 A Sampling Procedure with AEPA Property

In this section we first describe the approach as it applies to the case of integer margins. That is, the column totals, $n_{.j} = \sum_{i=1}^{R} r_{ij}$, and the row totals, $n_{i.} = \sum_{j=1}^{C} r_{ij}$, are integer valued. We go on to discuss how it can easily be adapted to the general case. In the linear programming approach, the goal is to minimize the expected lack of 'desirability' of the samples while maintaining the EPA property. We propose to accomplish this in two stages. First, we will develop an unequal probability selection procedure which selects samples which exactly match the integer margins and also have the AEPA property. We will then randomly generate a moderately sized set of such arrays and then apply a modified linear programming technique to this subset of all possible arrays. This will be repeated with larger and larger such sets. We will describe the sampling procedure and then we will discuss the modified linear programming technique.

Here is the basic idea for constructing such a sampling procedure: for a two-way table (assuming the expected cell sample sizes have been adjusted to lie between 0 and 1 as was done in going from Table 1 to 2), first we draw a sequence of population cells to produce $a_{11}, a_{12}, ..., a_{1C}$ in the first row using an unequal probability without replacement sampling procedure based on the expected counts of that row, where $a_{ij} = 1$ if the ij-th cell is selected and $= 0$ otherwise. Then we draw $a_{i1}, a_{i2}, ..., a_{iC}$ subsequently for $i > 1$ while keeping all $\sum_{k=1}^{i} a_{kj}$ less than or equal to the corresponding marginal column totals $n_{.j}$. The details of this sampling procedure are as follows:

Step 1: Randomly permute the rows and let $i = 1$. Given the first row of inclusion probabilities $r_{11}, r_{12}, ..., r_{1C}$, draw a sample of $n_{1.}$ cells out of C in the first row stratum using an unequal probability without replacement sampling procedure; record the first row of samples in terms of indicator variables $a_{11}, a_{12}, ..., a_{1C}$ as defined previously; let $A_j = a_{1j}$ for $j = 1, ..., C$.

Step 2: Let $i = i + 1$

Step 2.1: For $j = 1, ..., C$, do the following

a) Let $R_j = \sum_{k=1}^{i} r_{kj}$,

b) If $R_j - A_j \le 0$ let $a_{ij} = 0$,

c) If $R_j - A_j \ge 1$ let $a_{ij} = 1$,

Step 2.2: Let $J = \{ j : 0 < R_j - A_j < 1 \}$ and *rtot* $= \sum_{j=1}^{C} r_{ij} - \#\{ j : a_{ij} = 1 \}$. If *rtot* > 0 then $r'_{ij} = r_{ij} \times$ *rtot* $/ \sum_{j \in J} r_{ij}$, for $j \in J$. If there exists a $j_0 \in J$ such that $r'_{ij0} > 1$ then let $a_{ij0} = 1$ and go to Step 2.1. Otherwise go to Step 3.

Step 3: Draw a sample of *rtot* cells from J using an unequal probability without replacement sampling procedure and r'_{ij} to get a_{ij} for $j \in J$.

Let $A_j = \sum_{k=1}^{i} a_{kj}$ for $j = 1, ..., C$.

Step 4: If $i = R$, then stop; otherwise go to Step 2.

One aspect of this sampling procedure that should be noticed is that in Step 2, the way of re-calculating the i-th row of inclusion probabilities is not unique. However, the general rules that should be followed for this re-calculation are:

(a) $0 \le r'_{ij} \le 1$ and if $A_j = n_{.j}$, which means that there are enough units being selected from the j-th column, r'_{ij} should be set to 0; if $A_j = n_{.j} - (R - i + 1)$, which means that there will not be enough units to be selected for this column unless all of the remaining units are selected, r'_{ij} should be set to 1;

(b) keep $\sum_{j=1}^{C} r'_{ij} = \sum_{j=1}^{C} r_{ij} = n_{i.}$.

The method extends easily to non-integer margins. We delay detailed discussion, however, to the sequel.

We can now use the above method to generate a candidate set, S_{n0}, and apply the linear programming technique to this set. To see why we choose to modify the linear programming technique, realize that for the integer margin case every $s \in S_{n0}$ already attains the minimum in (2) so that a direct application of linear programming amounts to determining whether there is a feasible solution or not. Thus, if we generate say an S_{n0} of size 500 then 1,000 *etc*, and the linear programming package continues to find no feasible solution we really do not know if we are getting closer to a solution or not. Instead we choose to turn the optimization around and solve a dual problem

$$\min_{p(s)} \sum_{i,j} \Big| \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \Big| . \tag{8}$$

We know that $w(s) = 0$ for all $s \in S_{n0}$ and we are looking for a solution which yields a minimum of zero in (8). We have essentially switched the roles of the objective function and the EPA constraints in the original problem. The difficulty is that it is more difficult to use linear programming to handle (8). This can be done as follows. Set up constraints

$$\sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} + d_{ij} - e_{ij} = 0 \quad \text{for} \quad i = 1, ..., R \quad \text{and} \quad j = 1, ..., C, \tag{9}$$

where

$$d_{ij} \ge 0, e_{ij} \ge 0, d_{ij} e_{ij} = 0. \tag{10}$$

Then note that

$$\Big| \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \Big| = \begin{cases} d_{ij} \text{ if } \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} < 0 \\ e_{ij} \text{ if } \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \ge 0 \end{cases}$$

$$= d_{ij} + e_{ij}. \tag{11}$$

Thus, we can replace (8) by

$$\min_{p(s), d_{ij}, e_{ij}} \sum_{i,j} (d_{ij} + e_{ij}), \tag{12}$$

subject to

$$\sum_{s\in S_{n0}} n_{ij}(s)p(s) - r_{ij} + d_{ij} - e_{ij} = 0, d_{ij}, e_{ij}, p(s) \geq 0, d_{ij}\, e_{ij} = 0. \qquad (13)$$

3.3 Some Illustrating Examples with Integer Margins

In this section, two examples will be used to illustrate the sampling procedure. The first with a 10×8 array is described in detail to show the whole procedure. The second with a larger size (20×15) is given to demonstrate the size of problem that this method can handle (this is near the limit of the problem the proposed method can realistically handle). Any unequal probability without replacement sampling procedure can be used within the method. In Example 1 below, we chose to use the the random grouping method of Rao, Hartley and Cochran (1962), since it is simple and we really only need to approximately match the selection probabilities, which it does. However, the Rao-Hartley-Cochran method only works well up to problems of moderate size. In Examples 2 and 3 one should use a method which exactly matches the selection probabilities. There are many such available, but we chose to use one developed in Lu (2000).

Example 1. 10×8 **array with integer margins:** A two-way stratification problem with expected sample cell counts and sample size is given in Table 4.

Table 4
Expected Sample Cell Counts Under Proportionate Stratification ($n = 40$)

Row No.	Column No.								Marginal
	1	2	3	4	5	6	7	8	Row Total
1	0.41	0.55	0.58	0.80	0.23	0.61	0.70	0.12	4
2	0.52	0.15	0.07	0.90	0.28	0.10	0.37	0.61	3
3	0.72	0.15	0.65	0.73	0.39	0.34	0.85	0.17	4
4	0.70	0.55	0.46	0.10	0.41	0.05	0.24	0.49	3
5	0.07	0.63	0.45	0.81	0.52	0.02	0.70	0.80	4
6	0.61	0.33	0.79	0.21	0.02	0.61	0.67	0.76	4
7	0.88	0.48	0.73	0.69	0.44	0.64	0.86	0.28	5
8	0.22	0.14	0.85	0.37	0.69	0.45	0.49	0.79	4
9	0.85	0.44	0.80	0.76	0.31	0.71	0.60	0.53	5
10	0.02	0.58	0.62	0.63	0.71	0.47	0.52	0.45	4
Marginal Col Total	5	4	6	6	4	4	6	5	40

The basic steps of our sampling design are illustrated as follows:

Step 1. Obtain a representative candidate sample subset S_{n0} by using proposed sampling procedure with AEPA property to draw, say 500, samples (obtained within 3 minutes). The sample proportion of each cell is shown in Table 5, which can be compared to Table 4 to see how close these are to satisfying the EPA property.

Step 2. Solve the linear programming problem given by (12) and (13) to obtain

$$\min_{p(s),\, s\in S_{n0}} \sum_{i,j} \left|\sum_{s} n_{ij}(s)p(s) - nP_{ij}\right|. \qquad (14)$$

If the objective value of (14) is greater than zero, repeat Step 1 with a larger set S_{n0}. If the objective value of (14) is zero, stop, an optimal solution has been obtained.

Table 5
Sample Cell Counts Under Prop. Stratification ($n = 40$)

Row No.	Column No.								Marginal
	1	2	3	4	5	6	7	8	Row Total
1	0.408	0.554	0.582	0.776	0.250	0.594	0.734	0.102	4
2	0.554	0.150	0.062	0.916	0.280	0.122	0.366	0.550	3
3	0.690	0.144	0.638	0.720	0.402	0.360	0.838	0.208	4
4	0.692	0.542	0.452	0.120	0.416	0.044	0.260	0.474	3
5	0.060	0.602	0.446	0.814	0.568	0.016	0.708	0.786	4
6	0.558	0.348	0.780	0.216	0.012	0.634	0.682	0.770	4
7	0.866	0.480	0.734	0.676	0.470	0.664	0.842	0.268	5
8	0.254	0.158	0.848	0.400	0.654	0.412	0.490	0.784	4
9	0.870	0.418	0.830	0.772	0.292	0.692	0.624	0502	5
10	0.026	0.564	0.636	0.658	0.714	0.416	0.500	0.486	4
Marginal Col Total	5	4	6	6	4	4	6	5	40

In this example, a candidate subset S_{n0} with 500 samples was sufficient to get objective value of 0.

Example 2. 20×15 **array with integer margins:** In this example, a 20×15 array with integer margins is given in Table 6.

The actual computation steps are given as follows:

First Iteration:

Step 1. Draw 500 samples to form S_{n0}.
Step 2. The objective value of (14) is 0.1659.

Second Iteration:

Step 1. Draw 500 samples to add to S_{n0}.
Step 2. The objective value of (14) is 0. The final sampling design is attained.

This procedure took approximately 30-60 seconds using a Fortran program on a Sun Ultra 10 workstation.

3.4 Extension to Non-Integer Margins

The method extends easily to non-integer margins. Merely replace $n_{i.}$ throughout the algorithm by $n_{i.}^*$ which takes value $\lfloor r_{i.} \rfloor + 1$ with probability $\alpha = r_{i.} - \lfloor r_{i.} \rfloor$ and takes value $\lfloor r_{i.} \rfloor$ with probability $1 - \alpha$. The only additional difficulty is that $E[w(s)]$ cannot attain zero. Thus, we do not have an obvious lower-bound reference point to ascertain whether we are close to the best solution or not. However, the above randomization strategy ensures that for every obtained AEPA sample we have

$$|n_{i\cdot}(s) - r_{i\cdot}| < 1 \quad \text{and} \quad |n_{\cdot j}(s) - r_{\cdot j}| < 1 \quad \text{for} \quad i = 1, \ldots, R,\ j = 1, \ldots, C. \tag{15}$$

This together with the EPA property, $E[n_{ij}(s)] = \sum_s n_{ij}(s)p(s) = r_{ij}$ implies that the lack of desirability function $w(s)$ defined in (3) has a constant expectation

$$E[w(s)] = \sum_i \left(r_{i\cdot} - \lfloor r_{i\cdot} \rfloor\right)\left(1 + \lfloor r_{i\cdot} \rfloor - r_{i\cdot}\right) + \sum_j \left(r_{\cdot j} - \lfloor r_{\cdot j} \rfloor\right)\left(1 + \lfloor r_{\cdot j} \rfloor - r_{\cdot j}\right). \tag{16}$$

The proof of this is given in Appendix 1. Thus, if (14) attains zero under the above strategy then the resulting solution will yield minimum $E[w(s)]$ as in (16).

Example 3. 27×3 real example with non-integer margins: We will illustrate the method using a real example from environmental health (Vihma 1981). This study was concerned with occupational health of workers in various industries in Finland. The population chosen for study consisted of 1,430 small industrial workplaces (5-49 employees) totalling 22,893 employees in Uusimaa, the southern most and most industrialized province of Finland. The primary sampling units were the workplaces and a sample of n=100 such were desired. This was all that could be afforded given the cost of the eventual survey. The workplaces were stratified by two stratification variables: type of industry (27 categories) and number of employees (3 categories). The expected sample cell counts under proportionate stratification are given in Table 7. The actual sampling scheme used in this study was based on the method of Bryant *et al.* (1960) after some grouping strata as it was the only method available at the time of this study.

We applied our method to this problem. The minimum achievable $E[w(s)]$ using our proposed strategy is 5.0418. The actual computation steps were as follows:

First Iteration:

Step 1. Draw 500 samples to form S_{n0}, randomly generating the $n^*_{i\cdot}$ independently for each sample.

Step 2. The objective value of (14) is 0.45088.

Second Iteration:

Step 1. Draw 500 samples to add to S_{n0}.

Step 2. The objective value of (14) is 0. The final sampling design is attained and achieved the minimum value $E[w(s)] = 5.0418$.

This procedure took approximately 30 seconds using a Fortran program on a Sun Ultra 10 workstation.

Table 6
Expected Sample Cell Counts Under Proportionate Stratification (n =151)

0.73	0.58	0.08	0.59	0.69	0.84	0.04	0.17	0.27	0.80	0.02	0.84	0.79	0.03	0.53	7
0.43	0.39	0.35	0.57	0.35	0.38	0.47	0.53	0.39	0.96	0.52	0.27	0.68	0.40	0.31	7
0.73	0.25	0.15	0.73	0.48	0.32	0.91	0.49	0.03	0.61	0.14	0.61	0.73	0.25	0.87	7
0.13	0.28	0.35	0.60	0.26	0.38	0.37	0.39	0.71	0.01	0.93	0.72	0.30	0.66	0.91	7
0.32	0.06	0.86	0.47	0.80	0.93	0.96	0.30	0.65	0.72	0.67	0.54	0.51	0.77	0.44	9
0.12	0.78	0.81	0.34	0.28	0.02	0.89	0.41	0.94	0.82	0.37	0.81	0.85	0.51	0.05	8
0.48	0.51	0.50	0.62	0.35	0.11	0.85	0.78	0.29	0.39	0.69	0.07	0.67	0.78	0.91	8
0.86	0.41	0.11	0.17	0.75	0.89	0.48	0.48	0.91	0.20	0.53	0.67	0.34	0.19	0.01	7
0.81	0.00	0.13	0.93	0.36	0.12	0.19	0.86	0.33	0.04	0.79	0.69	0.56	0.37	0.82	7
0.82	0.22	0.54	0.82	0.61	0.46	0.74	0.33	0.24	0.53	0.41	0.18	0.30	0.03	0.77	7
0.95	0.60	0.35	0.33	0.95	0.43	0.06	0.63	0.71	0.02	0.55	0.23	0.87	0.21	0.11	7
0.96	0.65	0.96	0.83	0.41	0.58	0.49	0.27	0.74	0.88	0.93	0.46	0.60	0.13	0.11	9
0.83	0.54	0.05	0.96	0.79	0.70	0.33	0.81	0.86	0.45	0.45	0.84	0.29	0.30	0.80	9
0.75	0.65	0.63	0.04	0.32	0.36	0.38	0.80	0.50	0.23	0.37	0.23	0.85	0.69	0.20	7
0.79	0.31	0.55	0.26	0.04	0.05	0.91	0.11	0.43	0.79	0.14	0.64	0.44	0.48	0.06	6
0.23	0.92	0.81	0.42	0.49	0.10	0.74	0.56	0.24	0.47	0.34	0.57	0.60	0.56	0.95	8
0.13	0.77	0.65	0.66	0.05	0.23	0.58	0.74	0.19	0.94	0.26	0.75	0.16	0.71	0.18	7
0.31	0.01	0.60	0.38	0.01	0.55	0.70	0.72	0.20	0.87	0.55	0.82	0.77	0.44	0.07	7
0.63	0.67	0.21	0.02	0.16	0.68	0.14	0.17	0.95	0.78	0.58	0.55	0.94	0.96	0.56	8
0.99	0.40	0.31	0.26	0.85	0.87	0.77	0.75	0.42	0.49	0.76	0.51	0.75	0.53	0.34	9
12	9	9	10	9	9	11	10	10	11	10	11	12	9	9	151

Table 7
Occupational Health Survey, Vihma (1981) Expected Sample Cell Counts Under Proportionate Stratification (n = 100)

Type of Industry	Number of Personnel			
	5-9	10-19	20-49	$r_{i\cdot}$
Food products	2.38	3.56	3.78	9.72
Food	0.35	0.14	0.56	1.05
Beverage	0.14	0.07	0.21	0.42
Textiles	1.33	1.26	1.46	4.05
Apparel	3.15	3.71	2.09	8.95
Leather	0.56	0.14	0.07	0.77
Footwear	0.07	0.07	0.21	0.35
Wood Products	2.37	1.89	0.91	5.17
Furniture	1.33	0.84	0.91	3.08
Paper Products	0.42	0.49	0.42	1.33
Printing	7.20	6.01	4.20	17.41
Industrial Chemicals	0.56	0.35	0.28	1.19
Chemical Products	1.82	1.54	1.53	4.89
Petrolium	0.14	0.07	0.00	0.21
Misc Coal and Petrol.	0.07	0.07	0.14	0.28
Rubber Products	0.14	0.21	0.07	0.42
Plastic Products	1.40	1.05	1.19	3.64
Glass Products	0.42	0.21	0.21	0.84
Non-Metal Minerals	1.12	0.98	0.84	2.94
Iron & Steel	0.14	0.07	0.35	0.56
Nonferrous Metal	0.35	0.14	0.28	0.77
Fabricated Metal	4.96	4.06	2.59	11.61
Machinery	2.80	1.96	3.21	7.97
Electrical	1.89	1.60	1.33	4.82
Transport Equipment	0.84	0.84	0.84	2.52
Scientific Equipment	0.56	0.42	0.49	1.47
Manufacturing Industries	1.68	0.91	0.98	3.57
$n_{\cdot j}$	38.19	32.66	29.15	100.00

5. CONCLUDING REMARKS

We propose a method for two-way stratification which extends the applicability of the linear programming approach of Sitter and Skinner (1994) to much larger problems. The method focuses on how to construct a small "representative" candidate sample set by using an unequal probability sampling procedure which generates candidate samples which nearly meet the AEPA constraints of the linear programming problem and then applying the linear programming method to this much smaller set.

It should be noted that the linear programming method extends easily to stratified multi-stage designs. Since there is no fundamental difference between the original linear programming approach and the extension proposed here, this is still true of the proposed method. In the same spirit, one can view discussion on issues around variance estimation of the resulting estimators in Sitter and Skinner (1994) as well.

One should also note that once one restricts to bracketing integers around the nP_{ij}'s, the problem is related to a controlled rounding problem (see Kelly, Golden and Assad 1993, and references therein), though we do not explore this aspect here.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

APPENDIX 1

Proof of (16): $n_{i\cdot}(s) - \lfloor r_{i\cdot} \rfloor \sim \text{Bernoulli}(r_{i\cdot} - \lfloor r_{i\cdot} \rfloor)$ and has variance $(r_{i\cdot} - \lfloor r_{i\cdot} \rfloor)(1 + \lfloor r_{i\cdot} \rfloor - r_{i\cdot})$. This implies

$$\begin{aligned}\sum_s \left(n_{i\cdot}(s) - r_{i\cdot}\right)^2 p(s) &= E(n_{i\cdot}(s) - r_{i\cdot})^2 \, V(n_{i\cdot}(s)) \\ &= V\left(n_{i\cdot}(s) - \lfloor r_{i\cdot} \rfloor\right) \\ &= \left(r_{i\cdot} - \lfloor r_{i\cdot} \rfloor\right)\left(1 + \lfloor r_{i\cdot} \rfloor - r_{i\cdot}\right),\end{aligned}$$

and by similar argument that $\sum_s (n_{\cdot j}(s) - r_{\cdot j})^2 \, p(s) = (r_{\cdot j} - \lfloor r_{\cdot j} \rfloor)(1 + \lfloor r_{\cdot j} \rfloor - r_{\cdot j})$.

Therefore, with $w(s)$ defined in (3),

$$\begin{aligned}E[w(s)] &= \sum_s w(s)p(s) = \sum_s \left\{ \sum_i \left(n_{i\cdot}(s) - r_{i\cdot}\right)^2 + \sum_j \left(n_{\cdot j}(s) - r_{\cdot j}\right)^2 \right\} p(s) \\ &= \sum_i \sum_s \left(n_{i\cdot}(s) - r_{i\cdot}\right)^2 p(s) + \sum_j \sum_s \left(n_{\cdot j}(s) - r_{\cdot j}\right)^2 p(s) \\ &= \sum_i \left(r_{i\cdot} - \lfloor r_{i\cdot} \rfloor\right)\left(1 + \lfloor r_{i\cdot} \rfloor - r_{i\cdot}\right) + \sum_j \left(r_{\cdot j} - \lfloor r_{\cdot j} \rfloor\right)\left(1 + \lfloor r_{\cdot j} \rfloor - r_{\cdot j}\right).\end{aligned}$$

REFERENCES

BRYANT, E.C., HARTLEY, H.O. and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*. 55, 105-124.

CAUSEY, B.D., COX, L.H. and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

ENGLE, M., MARSDEN, G. and POLLOCK, S.W. (1971). Child work and social class. *Psychiatry*. 34, 140-150.

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

HESS, I., RIEDEL, D.C. and FITZPATRICK, T.B. (1976). *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, second edition.

JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*. 65, 776-795.

JESSEN, R.J. (1973). Some properties of probability lattice sampling. *Journal of the American Statistical Association*. 68, 20-28.

JESSEN, R.J. (1975). Square and cubic lattice sampling. *Biometrics*. 31, 449-471.

KELLY, J.K., GOLDEN, B.L. and ASSAD, A.A. (1993). The controlled rounding problem: complexity and computational experience. *European Journal of Operational Research*. 65, 207-217.

LAHIRI, P., and MUKERJEE, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistical Sinica*. 10, 1171-1178.

LU, W. (2000). Multi-way stratification by linear programming made practical. M.Sc. Thesis, Simon Fraser University.

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*. Serie B, 24, 482-491.

RAO, J.N.K., and NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*. 77, 807-814.

RAO, J.N.K., and NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*. 60, 89-98.

SITTER, R.R., and SKINNER, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*. 20, 65-73.

SKINNER, C.J., HOLMES, D.J. and HOLT, D. (1994). Multiple frame sampling for multiple stratification. *International Statistical Review*. 62, 333-347.

VIHMA, T. (1981). Health hazards and stress factors in small industry-Prevalence study in the province of Uusimaa with special reference to the type of industry and the occupational title as classifications for the description of occupational health problems. *Scandinavian Journal of Work, Environment and Health*. 7, Suppl. 3, 1-149.

WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*. 32, 150-164.

Survey Methodology, December 2002
Vol. 28, No. 2, pp. 209-212
Statistics Canada

On the Use of Generalized Inverse Matrices in Sampling Theory

ROBBERT H. RENSSEN and GERARD H. MARTINUS[1]

ABSTRACT

In theory, it is customary to define general regression estimators in terms of full-rank weighting models, *i.e.*, the design matrix that corresponds to the weighting model is of full rank. For such weighting models, it is well known that the general regression weights reproduce the (known) population totals of the auxiliary variables involved. In practice, however, the weighting model often is not of full rank, especially when the weighting model is for incomplete post-stratification. By means of the theory of generalized inverse matrices, it is shown under which circumstances this consistency property remains valid. As a non-trivial example we discuss the consistent weighting between persons and households as proposed by Lemaître and Dufour (1987). We then show how the theory is implemented in Bascula.

KEY WORDS: Bascula; General regression estimator; Weighting.

1. INTRODUCTION

Weighting methods that are based on the general regression estimator are commonly used in sample surveys to adjust for both sampling error and non-sampling error, see *e.g.* Bethlehem and Keller (1987) and Särndal, Swensson, and Wretman (1992). One complication in the use of general regression estimators, however, is that many weighting models are based on incomplete post-stratification, resulting in design matrices that are not of full rank. Usually, this problem is solved by using a reduced design matrix. Such a reduced design matrix can be constructed by deleting redundant columns and properly adjusting the population totals. Often, the redundancy can be recognized rather easily beforehand by the specification of the weighting model. However, for some weighting models such a redundancy check may be impractical.

For example, suppose that we have a post-stratification based on the complete crossing between two categorical variables A and B, with known counts for the population of each cell. We may obtain small sample counts or no sample in some cells. Then we may derive new classifications, A' from A and B' from B, by merging categories, and define the following more parsimonious scheme: $A + B + A' \times B'$. According to this incomplete post-stratification we simultaneously calibrate on three sets of counts, namely the marginal counts of A, the marginal counts of B, and the cell counts of $A' \times B'$. Since A and A' (and also B and B') appear in different weighting terms, it is difficult to recognize redundancy by the specification of the weighting model. This paper gives the theoretical background, which is based on generalized inverse matrices, of reducing such a design matrix.

In section 2 we briefly describe some properties of generalized inverse matrices. In section 3 we define the general regression estimator for weighting models that need not be of full rank. Given a regularity condition that can be nicely interpreted in a calibration estimation context (see Deville and Särndal 1992) it is shown that this estimator is invariant with respect to the choice of the generalized inverse. At the end of section 3 the fulfillment of this regularity condition is discussed for some well-known weighting models, such as incomplete post-stratification and consistent weighting between persons and households. In section 4 we describe the algorithm, which is implemented in Bascula (see Nieuwenbroek 1997; Renssen, Nieuwenbroek and Slootbeek 1997) for calculating the regression weights. Finally, in section 5 we briefly discuss the weighting model of the Dutch Labour Force Survey.

2. GENERALIZED INVERSE MATRICES

We are mainly interested in the use of generalized inverses within the framework of the general regression estimator. Hence, we only give some properties of a generalized inverse of the form $\mathbf{X}' \Lambda \mathbf{X}$, where Λ is a diagonal matrix of order $n \times n$ with strictly positive diagonal entries and $\mathbf{X}$ a design matrix of order $n \times p$ that results from the weighting model. For a more extensive discussion on generalized inverse matrices we refer to Searle (1971) and Rao (1973).

Before giving these properties, we briefly review the definition of a generalized inverse. Consider a $p \times q$ matrix A of any rank and let $Ax = y$ be a system of consistent equations, *i.e.*, any linear relationship existing among the rows of A also exists among the corresponding elements of y. A generalized inverse of A is a $q \times p$ matrix A^- such that $x = A^- y$ is a solution of this system of equations. It is easy to verify that the existence of A^- implies $AA^-A = A$ (choose y as the i-*th* column of A). Conversely, if A^- satisfies $AA^-A = A$ and $Ax = y$ is consistent, then $A(A^-y) = A\,(A^-Ax) = Ax = y$ and hence A^-y is a solution. Thus, as an alternative definition, a generalized

[1] Robbert H. Renssen and Gerard H. Martinus, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands.

inverse matrix of A is any matrix A^- such that $AA^-A = A$.

Now, if $\mathbf{G}$ denotes a generalized inverse of $\mathbf{X}'\Lambda\,\mathbf{X}$, then the following properties of $\mathbf{G}$ are proven in Searle (1971) for $\Lambda = \mathbf{I}_n$:

(P1) $\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\Lambda\,\mathbf{X}$,

(P2) $\mathbf{XGX}'\Lambda\,\mathbf{X} = \mathbf{X}$ *i.e.*, $\mathbf{GX}'\Lambda$ is a generalized inverse of $\mathbf{X}$,

(P3) $\mathbf{XGX}'$ is invariant to the choice of $\mathbf{G}$,

(P4) $\mathbf{XGX}' = \mathbf{XG}'\mathbf{X}'$ whether $\mathbf{G}$ is symmetric or not.

The proofs of (P1) to (P4) for diagonal are almost identical to those of Searle (1971, chapter 1.5, theorem 7) and therefore not repeated here.

3. THE GENERAL REGRESSION ESTIMATOR

Consider a finite population U of N units from which a sample S of n units is drawn without replacement. Let π_k denote the first order inclusion probability of the k-*th* unit, $k = 1, ..., N$. We associate with each unit a vector of study variables $\mathbf{y}_k$. Then, the data matrix for the sampled units is given by $\mathbf{Y}_S = (\mathbf{y}_1, ..., \mathbf{y}_n)'$. We distinguish between study variables with known population totals (auxiliary variables) and study variables with unknown population totals. The start in the definition of a general regression estimator (Särndal *et al.* 1992) is the specification of the weighting model, *i.e.*, the choice of the set of auxiliary variables to be used in the estimation. Denoting this specific set of p variables by $\mathbf{x}$, we call the $n \times p$ matrix $\mathbf{X}_S = (\mathbf{x}_1, ..., \mathbf{x}_n)'$ the design matrix, which is, by definition, a column subset of $\mathbf{Y}_S$. The vector of known population totals of $\mathbf{x}$ is denoted by $\mathbf{t}_x$. Let $\mathbf{x}_{\mathrm{HT}} = \sum_{k \in S} \pi_k^{-1}\mathbf{x}_k$ denote the Horvitz-Thompson estimator for $\mathbf{t}_x$, then, given $\mathbf{x}$, the general regression estimator of the vector of population totals of the i-*th* study variable $\mathbf{y}_k^{(i)}$ is defined as

$$\hat{\mathbf{t}}_{\mathrm{greg}}^{(i)} = \mathbf{y}_{\mathrm{HT}}^{(i)} + \hat{\mathbf{B}}'(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}) \qquad (1)$$

with

$$\hat{\mathbf{B}} = \mathbf{G}_S\mathbf{X}_S'\Lambda_S\mathbf{Y}_S^{(i)}.$$

In terms of regression weights, this general regression estimator can also be written as

$$\hat{\mathbf{t}}_{\mathrm{greg}}^{(i)} = \sum_{k \in S} w_k\mathbf{y}_k^{(i)} \qquad (2)$$

with

$$w_k = \pi_k^{-1} + \lambda_k\mathbf{x}_k'\mathbf{G}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}).$$

Here, $\mathbf{G}_S$ denotes a generalized inverse of $\mathbf{X}_S^t\Lambda_S\mathbf{X}_S$ and $\Lambda_S = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ is some diagonal matrix with strictly positive entries.

Like the weighting model, the diagonal matrix Λ_S has to be specified by the user. Often, one takes $\Lambda_S = \Pi_S^{-1}\sum_S^{-1}$, where $\Pi_S = \mathrm{diag}(\pi_1, ..., \pi_n)$ and $\sum_S = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2)$ with σ_k^2 interpreted as the variance of independent random variables of which some of the study variables are supposed to be the outcome according to some super-population model, see Särndal *et al.* (1992). It is required that all σ_k^2 be known up to a common scale factor. An important special case is $\sigma_k^2 = \sigma^2$, *i.e.*, all the modeled variances are the same. This results in the regression estimator proposed by Bethlehem and Keller (1987). If the population units represent households (of size m_k) and if we take $\sigma_k^2 = m_k\sigma^2$ we arrive at the estimator proposed by Lemaître and Dufour (1987) to obtain consistent weights between person and households. From a different point of view, Alexander (1987) derived the GLS-P estimate, which results in essentially the same estimator.

Below we show that the regression weights are invariant to the choice of $\mathbf{G}_S$. To that purpose we make the following assumption:

(A1) there exists a n-vector $\mathbf{w}$ such that $\mathbf{X}_S'\mathbf{w} = \mathbf{t}_x$.

Clearly, this assumption states that $\mathbf{X}_S'\mathbf{w} = \mathbf{t}_x$ is a system of consistent equations. It is interesting to note that this system precisely corresponds to the set of calibrations equations when considering the general regression estimator as a special case of the calibration estimator (see *e.g.* Deville and Särndal 1992). If $\mathbf{X}_S'\mathbf{w} = \mathbf{t}_x$ is a system of consistent equations, then so is $\mathbf{X}_S'\mathbf{v} = (\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})$. This is easily seen by taking $\mathbf{v} = \mathbf{w} - \mathbf{d}_S$ with $\mathbf{d}_S = (\pi_1^{-1}, ..., \pi_n^{-1})'$. The invariance of the regression weights to the choice of $\mathbf{G}_S$, and hence the invariance of the general regression estimator can be shown as follows. Let $\mathbf{F}_S$ be some other generalized inverse of $\mathbf{X}_S'\Lambda_S\mathbf{X}_S$, different from $\mathbf{G}_S$. Then, we have

$$\begin{aligned}\mathbf{X}_S\mathbf{G}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}) &= \mathbf{X}_S\mathbf{G}_S\mathbf{X}_S'\mathbf{v} && \text{by (A1)}\\ &= \mathbf{X}_S\mathbf{F}_S\mathbf{X}_S'\mathbf{v} && \text{by (P3)}\\ &= \mathbf{X}_S\mathbf{F}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}). && \text{by (A1)}\end{aligned}$$

So, it holds that $\mathbf{x}_k'\mathbf{G}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})$ is invariant to $\mathbf{G}_S$ for all $k \in S$, implying that the regression weights are invariant to the choice $\mathbf{G}_S$.

The fact that these weights reproduce the population totals of the auxiliary variables follows from the following series of equations:

$$\begin{aligned}\sum_{k \in S} w_k\mathbf{x}_k &= \mathbf{x}_{\mathrm{HT}} + \sum_{k \in S}\mathbf{x}_k\lambda_k\mathbf{x}_k'\mathbf{G}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})\\ &= \mathbf{x}_{\mathrm{HT}} + (\mathbf{X}_S'\Lambda_S\mathbf{X}_S)\mathbf{G}_S(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})\\ &= \mathbf{x}_{\mathrm{HT}} + (\mathbf{X}_S'\Lambda_S\mathbf{X}_S)\mathbf{G}_S\mathbf{X}_S'\mathbf{v} && \text{by (A1)}\\ &= \mathbf{x}_{\mathrm{HT}} + \mathbf{X}_S'\mathbf{v} && \text{by (P2) and (P4)}\\ &= \mathbf{x}_{\mathrm{HT}} + (\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}) = \mathbf{t}_x. && \text{By (A1)}\end{aligned}$$

We close this section by having a closer look at the stated assumption for some well-known weighting models. In case of post-stratification in which the weighting model is described by a complete crossing of categorical variables, (A1) has a simple interpretation. Namely (A1) is satisfied if and only if empty post-strata in the sample correspond to empty post-strata in the population. Next, we consider incomplete post-stratification in which the weighting model consists of several terms, each term describing a complete crossing of categorical variables and so each term corresponding to a post-stratification. Then, a necessary condition for (A1) to be satisfied is that empty post-strata in the sample correspond to empty post-strata in the population for each of these terms. Unfortunately, this condition is not sufficient. For example, inconsistencies may still occur when we attempt to calibrate on a number of complete crossings larger than the sample size.

The assumption is less straightforward in case of consistent weighting between persons and households (see *e.g.* Lemaître and Dufour 1987). This is due to the redefinition of the auxiliary variable. For example, if $\mathbf{x}_k$ is a variable defined at the person level, and from this variable a new variable is defined on the household level, say $\mathbf{z}_k$, then (A1) should be defined in terms of $\mathbf{Z}_S = (\mathbf{z}_1, ..., \mathbf{z}_n)'$ instead of $\mathbf{X}_S$, *i.e.*, (A1) is satisfied if there exists an n-vector $\mathbf{w}$ such that $\mathbf{Z}_S^t \mathbf{w} = \mathbf{t}_x$. In many (regular) situations, the linear manifold spanned by $\mathbf{Z}_S$ will coincide with the linear manifold spanned by $\mathbf{X}_S$. In such situations the method of Lemaître and Dufour does not affect the validity of (A1). However, in specific cases this may not be true. The following simplified example illustrates this.

Let $\mathbf{x}_k$ denote sex of the k-*th* person, say $\mathbf{x}_k = (0, 1)'$ if the k-*th* person is a female and $\mathbf{x}_k = (1, 0)'$ if the k-*th* is a male. According to the method of Lemaître and Dufour (1987), let $\mathbf{z}_k$ denote the j-*th* household mean for $\mathbf{x}_k$ whenever k belongs to the j-*th* household. Furthermore, let the population consists of N_1 males and N_2 females, from which a sample of 10 households is drawn. Suppose that each sampled household consists of two persons, namely one male and one female. This gives $\mathbf{z}_k = (1/2, 1/2)'$ for all $k \in S$. For this example the linear manifold spanned by $\mathbf{Z}_S$ is a linear subspace of the linear manifold spanned by $\mathbf{X}_S$. If $N_1 = N_2$ then (A1) is satisfied. Otherwise, if $N_1 \neq N_2$ then (A1) is not satisfied. Especially, when the method of Lemaître and Dufour is applied on a relatively large weighting model, the linear manifold spanned by $\mathbf{Z}_S$ may be a proper subspace of the linear manifold spanned by $\mathbf{X}_S$. Then, (A1) only is satisfied if $\mathbf{t}_x$ accidentally belongs to this subspace.

4. CALCULATING THE REGRESSION WEIGHTS IN BASCULA

In the previous section we have shown that the general regression weights $w_k = \pi_k^{-1} + \lambda_k \mathbf{x}_k^t \mathbf{G}_S (\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})$ are invariant to the choice of $\mathbf{G}_S$. In this section we show how to compute these weights. To do so, we start with the Cholesky decomposition of the positive (semi) definite matrix $\mathbf{X}_S^t \Lambda_S \mathbf{X}_S$, see Seber (1977, page 322). If $\mathbf{X}_S$ is of full rank, then $\mathbf{X}_S^t \Lambda_S \mathbf{X}_S$ is positive definite and it can be expressed uniquely in the form $\mathbf{X}_S^t \Lambda_S \mathbf{X}_S = \mathbf{U}' \mathbf{U}$, where $\mathbf{U}$ is an upper triangular matrix with positive diagonal elements. Let a_{ij} denote the ij-*th* element of $\mathbf{X}_S^t \Lambda_S \mathbf{X}_S$, then $\mathbf{U}$ can be computed, row by row, according to

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2} \quad \text{for} \quad i = 1, ..., p$$

and

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj}}{u_{ii}} \quad \text{for} \quad j = i+1, ..., p. \tag{3}$$

If $\mathbf{X}_S$ has rank $r < p$, then an application of (3) will give r non-zero and $p - r$ zero diagonal elements of $\mathbf{U}$. If we find a zero diagonal element then we put its corresponding row and column elements at zero. Subsequently, by elementary row and column interchanges, we obtain the following upper triangular matrix:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Accordingly to the elementary row and column interchanges we also interchange the elements of $\mathbf{X}_S$ and $(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})$: $\mathbf{X}_S \mathbf{E}' = (\mathbf{X}_{1S} \mathbf{X}_{2S})$ and

$$\mathbf{E}(\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}}) = \begin{pmatrix} (\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}}) \\ (\mathbf{t}_{2x} - \mathbf{x}_{2\mathrm{HT}}) \end{pmatrix},$$

where, by construction, $\mathbf{X}_{1S}$ is of full rank and $\mathbf{E}$ is a non-singular matrix of order $p \times p$. But, since

$$\mathbf{G}_S' = \begin{pmatrix} (\mathbf{X}_{1S}^t \Lambda_S \mathbf{X}_{1S})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^{-1} (\mathbf{U}_1^t)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

is a generalized inverse of $(\mathbf{X}_{1S} \mathbf{X}_{2S})^t \Lambda_S (\mathbf{X}_{1S} \mathbf{X}_{2S})$, we have that $G_S = \mathbf{E}' \mathbf{G}_S' \mathbf{E}$ is a generalized inverse of $\mathbf{X}_S^t \Lambda_S \mathbf{X}_S$. Inserting this generalized inverse into $w_k = \pi_k^{-1} + \lambda_k \mathbf{x}_k^t \mathbf{G}_S (\mathbf{t}_x - \mathbf{x}_{\mathrm{HT}})$ gives

$$w_k = \pi_k^{-1} + \lambda_k (\mathbf{x}_{1k}^t \; \mathbf{x}_{2k}^t) \mathbf{G}_S' \begin{pmatrix} (\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}}) \\ (\mathbf{t}_{2x} - \mathbf{x}_{2\mathrm{HT}}) \end{pmatrix}$$

$$= \pi_k^{-1} + \lambda_k \mathbf{x}_{1k}^t \mathbf{U}_1^{-1} (\mathbf{U}_1^t)^{-1} (\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}}),$$

which is computed as follows. First $\mathbf{z} = (\mathbf{U}_1^t)^{-1} (\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}})$ is computed by solving the lower triangular system $\mathbf{U}_1^t \mathbf{z} = (\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}})$. Thereafter $\mathbf{u} = \mathbf{U}_1^{-1} \mathbf{z}$ is computed by solving the upper triangular system $\mathbf{U}_1 \mathbf{u} = \mathbf{z}$. Once

$\mathbf{u} = \mathbf{U}_1^{-1}(\mathbf{U}_1^{t})^{-1}(\mathbf{t}_{1x} - \mathbf{x}_{1\mathrm{HT}})$ is computed it is a simple matter to compute w_k.

5. THE DUTCH LABOUR FORCE SURVEY

To illustrate some of the issues stated in this paper, we briefly discuss the weighting model of the Dutch Labour Force Survey (LFS) of 1987 up to 2000. The target population of this survey consisted of the non-institutional population residing in the Netherlands and its sampling design was based on a stratified three-stage sampling with households as ultimate sampling units. For details we refer to Nieuwenbroek and Van der Valk (1996). Five categorical variables were involved into the weighting model, namely Sex (2 categories), Age (12 categories), Marital Status (2 categories), Region (15 categories), and Nationality (2 categories). Mainly based on consistency requirements, the desired weighting model was

$$\mathrm{Sex} \times \mathrm{Age} \times \mathrm{MaritalStatus} \times \mathrm{Region} \times \mathrm{Nationality}.$$

However, this weighting model resulted in too many small cell counts, which gave unstable estimators. Therefore, the reduced model

$$(\mathrm{Sex} \times \mathrm{Age} \times \mathrm{MaritalStatus} \times \mathrm{Region}) + (\mathrm{Sex} \times \mathrm{Age}^{+} \times \mathrm{Region} \times \mathrm{Nationality})$$

was used instead, where Age^{+} (2 categories) was obtained by grouping the categories in Age. This reduced weighting model resulted in a design matrix not of full rank for two reasons, namely 1) some columns of the design matrix completely consisted of zeros due to impossible combinations of the categorical variables and 2) there were linear combinations between the columns of the design matrix.

Now, the first kind of redundancy can be easily traced. If such columns are found, then their corresponding population totals should be zero. Bascula carries out a check on this condition. The second kind of redundancy is more difficult to trace. Linear combinations between columns may arise because one variable is incorporated into several weighting terms. For example, sex and region appear in both weighting terms of the LFS weighting model. The resulting linear combinations can be recognized beforehand by the name of the variable. For the age-variable, which also appears in both weighting terms, such a redundancy check beforehand is less obvious. These latter kinds of redundancy are traced by means of the Cholesky decomposition. Naturally, if any linear combinations are found, either by name beforehand or by the Cholesky decomposition, then the same linear combinations should also exists between the vector of population totals. Bascula also checks this condition.

ACKNOWLEDGEMENT

The authors wish to thank Jan van den Brakel, Nico Nieuwenbroek, and Jeroen Pannekoek for their careful reading and helpful comments. The authors also wish to thank the referee for his careful reading and valuable suggestions. Especially his remarks on assumption (A1) of a previous version and his suggestions to simplify several proofs have led to a considerable improvement of the paper.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*. 13, 183-198.

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

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

LEMÂITRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*. 13, 199-207.

NIEUWENBROEK, N.J. (1997). General regression estimator in Bascula: Theoretical background. Research paper no. 9737, Statistics Netherlands.

NIEUWENBROEK, N.J., and VAN DER VALK, J. (1996). Research paper no. 9629, Statistics Netherlands.

RAO, C.R. (1973). *Linear Statistical Inference And Its Applications* (2nd edition). New York: John Wiley & Sons, Inc.

RENSSEN, R.H., NIEUWENBROEK, N.J. and SLOOTBEEK, G. T. (1997). Variance module in Bascula: Theoretical background. Research paper no. 9712, Statistics Netherlands.

SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J.H. (1992). *Model-assisted Survey Sampling*. New York, Spinger-Verlag.

SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees during 2002. An asterisk indicates that the person served more than once.

A. Alavi, *Statistics Canada*
C. Alexander, *U.S. Bureau of the Census*
* J-F. Beaumont, *Statistics Canada*
* D.R. Bellhouse, *University of Western Ontario*
* D.A. Binder, *Statistique Canada*
* G.J. Brackstone, *Statistics Canada*
K. Brewer, *Australian National University*
J.M. Brick, *Westat, Inc.*
T. Buskirk, *University of Nebraska-Lincoln*
R. Chambers, *University of Southampton*
M. J. Cho, *United States Bureau of Labor Statistics*
C. Clark, *U.S. Bureau of the Census*
M. Cohen, *U.S. Bureau of Transportation Statistics*
M. Cruddas, *Office of National Statistics UK*
G. Datta, *University of Georgia*
* P. Dick, *Statistics Canada*
J. Dumais, *Statistics Canada*
J. Eltinge, *U.S. Bureau of Labor Statistics*
L. Ernest, *Bureau of the Labour Statistics*
S. Fienberg, *Carnegie - Mellon University*
J.-M. Fillion, *Statistics Canada*
W.A. Fuller, *Iowa State University*
* J. Gambino, *Statistics Canada*
* M. Ghosh, *University of Florida*
G. Glonek, *Flinders University*
J. Green, *Westat, Inc.*
S. Heeringa, *University of Michigan*
* M.A. Hidiroglou, *Statistics Canada*
H. Hogan, *U.S. Bureau of Labor Statistics*
* D. Judkins, *Westat, Inc.*
* G. Kalton, *Westat, Inc.*
D. Kostanich, *U.S. Bureau of the Census*
* P. Kott, *NASS*
* M. Kovačević, *Statistics Canada*
J. Kovar, *Statistics Canada*
P. Lahiri, *Joint Program in Survey Methodology*
* M.D. Larsen, *University of Chicago*
M. Latouche, *Statistics Canada*
* P. Lavallée, *Statistics Canada*
S. Linacre, *Official National Statistics*
* S. Lohr, *Arizona State University*
G. Maiti, *Iowa State University*
* H. Mantel, *Statistics Canada*
G. Meeden, *University of Minnesota*
J.-P. Morin, *Statistics Canada*
D. Norris, *Statistics Canada*
D. Paton, *Statistics Canada*
C.R. Perry, *NASS*
D. Pfeffermann, *Hebrew University*
T.E. Raghunathan, *University of Michigan*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
E. Rancourt, *Statistics Canada*
J. Reiter, *Duke University*
* L.-P. Rivest, *Université Laval*
S. Roehrig, *Carnegie - Mellon University*
P.-A. Salamin, *Swiss Federal Statistical Office*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
I. Şchiopu-Kratina, *Statistics Canada*
* R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
D.D. Smith, *U.S. Bureau of the Census*
* E. Stasny, *Ohio State University*
* A. Théberge, *Statistics Canada*
Y. Thibaudeau, *U.S. Bureau of the Census*
R.C. Tiwari, *National Cancer Institute*
J. Tourigny, *Statistics Canada*
* R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
W.E. Winkler, *U.S. Bureau of the Census*
K. Wolter, *National Opinion Research Center*
P. Wong, *Statistics Canada*
Y. You, *Statistics Canada*
* W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 2002 issues: H. Laplante (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Cousineau, C. Ethier, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 18, No. 2, 2002

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents
Volume 18, No. 3, 2002

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 30, No. 2, June/juin 2002, 177-346

CONTENTS TABLE DES MATIÈRES

Volume 30, No. 3, September/septembre 2002, 347-492

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.