Catalogue no. 12-001-X

Survey Methodology

June 2008

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

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

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-001-XIE, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select "Publications" > "Free Internet publications."

This product, Catalogue no. 12-001-XPE, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

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

Standards of service to the public

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

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2008

Published by authority of the Minister responsible for Statistics Canada

June 2008

Catalogue no. 12-001-XIE
ISSN 1492-0921

Catalogue no. 12-001-XPB
ISSN: 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-X au catalogue).

Note of appreciation

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 34, Number 1, June 2008

Contents

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 1-2
Statistics Canada, Catalogue No. 12-001-X

In This Issue

This issue of *Survey Methodology* includes papers on a variety of methodology topics.

In the first paper, Thompson and Wu consider the problem of obtaining inclusion probabilities, for the derivation of sampling weights, when modifications or compromises to the original sample design have been made due to practical constraints or limitations. The problem was motivated by the International Tobacco Control Policy Evaluation Survey of China which used a multi-stage unequal probability design for the selection of adult smokers and nonsmokers from seven cities. Due to refusal to participate by some districts, substitution units had to be selected after the original sample was selected. This substitution made it very difficult to calculate first order inclusion probabilities and practically impossible to calculate second order probabilities. In the paper the authors demonstrate, both theoretically and empirically, that the first and second order inclusions probabilities can be accurately estimated through Monte Carlo simulations.

Torabi and Rao derive the mean squared error (MSE) of a proposed new generalized regression estimator (GREG) of a small area mean under a two-level model and provide both theoretical and empirical comparisons between the new GREG and the best linear unbiased prediction estimator in terms of relative efficiency.

The paper by You discuses various cross-sectional and time series small area models for unemployment rate estimation for Canadian sub-provincial areas. In particular he considers an integrated non-linear mixed effects model under the hierarchical Bayes (HB) framework. An HB approach with the Gibbs sampling method is used to obtain estimates of posterior means and posterior variances of small area unemployment rates. The proposed HB model leads to reliable model-based estimates in terms of CV reduction. You also analyses the proposed model fitness and compares the model-based estimates to direct estimates.

The paper by Wang, Fuller and Qu studies small area estimation under a restriction. The authors study the impact of different augmented models in terms of MSE of the EBLUP. They consider small area models augmented with one additional explanatory variable for which the usual small area predictors achieve a self-calibrated property. The then consider small area models augmented with an added auxiliary variable that is a function of area size to reduce the bias when an incorrect model is used for prediction.

Nandram and Choi present an interesting approach to allocating undecided voters in surveys conducted prior to an election. Data from election polls are typically presented in two-way categorical tables with many polls taken before the actual election. They present the construction and analysis of a time-dependent nonignorable nonresponse model using Bayesian methods. They compare their model to extended versions (to include time) of ignorable and nonignorable nonresponse models introduced by Nandram, Cox and Choi (Survey Methodology, 2005). They also construct a new parameter to help predict the winner. The approach is illustrated using polling data from the 1998 race for governor of Ohio.

In their paper, Lazar, Meeden and Nelson develop a Bayesian approach to finite population sampling, through the use of a Polya posterior, when prior information is available in the form of partial knowledge about an auxiliary variable. The authors introduce the constrained Polya estimator and show that it has similarities with the generalized regression estimator under simple random sampling. However, their estimator does not require specification of a linear model. It is also related to empirical likelihood methods. Examples are used to illustrate the theory.

Zaslavsky, Zheng and Adams consider optimal sampling rates in element-sampling designs when the anticipated analysis is a survey-weighted linear regression and the estimands of interest are linear combinations of regression coefficients from one or more models. Methods are first developed assuming that exact design information is available in the sampling frame and then generalized to situations in which some design variables are available only as aggregates for groups of potential subjects, or from inaccurate or old data. Potential applications include estimation of means for several sets of overlapping domains, estimation for subpopulations such as minority races by disproportionate sampling of geographic areas, and studies in which characteristics available in sampling frames are measured with error.

The paper by Li explores the problem of estimating a finite population total using a nonlinear generalized regression estimator. The Box-Cox technique along with pseudo maximum likelihood estimation is used to obtain data-driven predictions. The author shows that the resulting regression estimator is design-consistent. Its performance is also evaluated through a simulation study.

Béguin and Hulliger extend the BACON algorithm to handle incomplete survey data. The BACON algorithm was developed to identify multivariate outliers using Mahalanobis distance. In the presence of missing values, the EM algorithm can be considered to estimate the covariance matrix at each iteration step of the BACON algorithm. The authors modify the EM algorithm to handle finite population sampling, which they call the EEM (Estimated Expectation Maximization) algorithm, and combine this algorithm with the BACON algorithm. This leads to the proposed BACON-EEM algorithm. It is then applied to two datasets and compared with alternative methods.

The paper by Jäckle and Lynn provides an empirical assessment of the effects of continued incentive payments on attrition, nonresponse bias and item nonresponse, and whether these effects change across waves of a multi-mode panel survey of young people in the UK. They test several hypotheses about the effects of incentives. They conclude that respondent incentives are an effective means of maintaining sample sizes of a panel, thus ensuring its value in terms of efficiency, especially for subgroup analyses. However, they also found that incentives had no effect on attrition bias.

Finally, Mohadjer and Curtin discuss challenges in designing and implementing a sample selection process that satisfies the goals of the National Health and Nutrition Examination Survey (NHANES). They describe how the sample design for NHANES must balance the requirement for efficient subdomain samples with the need for an efficient workload for the interview and examination staff at the Mobile Examination Centres (MEC), while keeping response rates as high as possible and costs down. The article elaborates on a number of unique features of the NHANES design and concludes with a brief summary of what has been achieved and some of the challenges facing future NHANES designs.

Harold Mantel, Deputy Editor

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 3-10
Statistics Canada, Catalogue No. 12-001-X

Simulation-based randomized systematic PPS sampling under substitution of units

Mary E. Thompson and Changbao Wu [1]

Abstract

The International Tobacco Control (ITC) Policy Evaluation Survey of China uses a multi-stage unequal probability sampling design with upper level clusters selected by the randomized systematic PPS sampling method. A difficulty arises in the execution of the survey: several selected upper level clusters refuse to participate in the survey and have to be replaced by substitute units, selected from units not included in the initial sample and once again using the randomized systematic PPS sampling method. Under such a scenario the first order inclusion probabilities of the final selected units are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In this paper we develop a simulation-based approach for computing the first and the second order inclusion probabilities when direct calculation is prohibitive or impossible. The efficiency and feasibility of the proposed approach are demonstrated through both theoretical considerations and numerical examples. Several R/S-PLUS functions and codes for the proposed procedure are included. The approach can be extended to handle more complex refusal/substitution scenarios one may encounter in practice.

Key Words: Inclusion probability; Horvitz-Thompson estimator; Rao-Sampford method; Relative bias; Unequal probability sampling without replacement.

1. Introduction

Construction of survey weights is the first critical step in analyzing complex survey data. It starts with the calculation of the first order inclusion probabilities, which is often straightforward if the original sampling design is well executed without any alterations and/or modifications. For instance, if the sample units are selected with inclusion probability (π) proportional to size (PPS or πps), then the inclusion probabilities are readily available from a simple re-scaling of the size variable. Among existing unequal probability without replacement PPS sampling procedures which are applicable for arbitrary fixed sample sizes, the randomized systematic PPS sampling method is the simplest one to implement. The procedure was first described in Goodman and Kish (1950) as a controlled selection method, and was refined by Hartley and Rao (1962) who studied the important and yet difficult problem of how to compute the second order inclusion probabilities. Let $x_i, i = 1, 2, \ldots, N$ be the values of the known size variable, where N is the total number of units in the population. Let $z_i = x_i / X$ where $X = \sum_{i=1}^{N} x_i$ and assume $nz_i < 1$ for all i. The randomized systematic PPS sampling procedure is as follows: Arrange the N population units in a random order and let $A_0 = 0$ and $A_j = \sum_{i=1}^{j} (nz_i)$ be the cumulative totals of nz_i in that order so that $0 = A_0 < A_1 < \ldots < A_N = n$. Let u be a uniform random number over [0, 1]. The n units to be included in the sample are those with indices j satisfying $A_{j-1} \le u + k < A_j$ for $k = 0, 1, \ldots, n-1$. Let s be the set of n sampled units and $\pi_i = P(i \in s)$ be the first order inclusion probabilities. The randomized systematic PPS sampling procedure satisfies the condition

$$\pi_i = nz_i, \ i = 1, 2, \ldots, N. \tag{1.1}$$

Several other without replacement sampling procedures which satisfy (1.1) for an arbitrary fixed sample size n were also proposed in the literature, including the well-known Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967) and those of Chao (1982), Chen, Dempster and Liu (1994), Tillé (1996) and Deville and Tillé (1998), among others.

The extensive research work on PPS sampling methods was largely stimulated by the Horvitz-Thompson (HT) estimator $\hat{T} = \sum_{i \in s} y_i / \pi_i$ for the population total $T = \sum_{i=1}^{N} y_i$ of a study variable y. The HT estimator is extremely efficient when y is highly correlated with the size variable x and the sampling procedure satisfies (1.1). It is the unique design unbiased estimator among the class of linear estimators $\sum_{i \in s} w_i y_i$ for T if the weights w_i depend only on i.

While a PPS sampling procedure can be desirable from a theoretical point of view, it is often difficult and/or sometimes impossible to execute due to practical constraints and limitations. Certain modifications and compromises will have to be made. The modified design, however, will no longer satisfy condition (1.1). Direct calculation of the final inclusion probabilities often becomes difficult or even impossible. Among common problems arising from survey practice which require alteration of the original sampling

1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: methomps@uwaterloo.ca; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: cbwu@uwaterloo.ca.

design, units refusal and substitution of units are the most frequently encountered ones. The scenario is well illustrated by the following example.

The International Tobacco Control (ITC) Policy Evaluation Survey of China (ITC China Survey) uses a multi-stage unequal probability sampling design for the selection of adult smokers and nonsmokers from seven cities. Each city has a natural hierarchical administrative structure

City → Street District→ Residential Block→ Household → Individual

which was conveniently integrated into the sampling design. At the upper levels, the randomized systematic PPS sampling method is used to select ten street districts from each city, with probability proportional to the population size of the district, and then two residential blocks are selected within each selected district, again using the randomized systematic PPS sampling method, with probability proportional to the population size of the block. Households and individuals within households are further selected, using a modified simple random sampling method. The original plan was to select 40 adult smokers and 10 adult nonsmokers from each of the 20 residential blocks, making the final sample with 800 smokers and 200 nonsmokers for each city.

A difficulty, however, arises in the execution of the survey: several selected upper level clusters (first Street Districts and then Residential Blocks) have refused to participate in the survey, due to time conflict with other activities or unavailability of human resources. These refusing clusters have to be replaced by substitute units, selected from units not included in the initial sample; one possibility is to use once again the randomized systematic PPS sampling method, to achieve the targeted overall sample size.

Under multi-stage sampling designs such as the one used for the ITC China survey, first order inclusion probabilities for individuals selected in the final sample can be calculated by multiplying the inclusion probabilities of units at different stages. When the randomized systematic PPS sampling method is modified due to substitution of units at a certain stage, the condition (1.1) no longer holds for the final sample at that stage. The first order inclusion probabilities under such a scenario are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In Appendix A, we provide a method of direct calculation (5.2) for the π_i when both the initial and the substitute samples are selected using the randomized systematic PPS sampling, assuming random refusal from the initial sample and no refusal from the substitute sample. The expression is valid conditional on the number of refusals and the population order used (after randomization) for the selection of the initial sample. It is apparent that even under such restrictive conditions and assumptions, the expression itself becomes computationally unfriendly with a not-so-large sample size.

In this paper we demonstrate, through both theory and numerical examples, that the first and the second order inclusion probabilities can be accurately estimated through Monte Carlo simulations when complete design information is available. Our numerical examples are motivated by the ITC China survey for which the randomized systematic PPS sampling serves as a baseline method but our theoretical results and the general methodology apply to other unequal probability without replacement sampling procedures as well. Section 2 presents results on the accuracy of simulation based methods. Numerical examples and comparisons are given in Section 3. Several R/S-PLUS functions and codes for the proposed procedure, originally developed for the ITC China survey, are included in Appendix C. Some additional remarks are given in Section 4.

2. Properties of simulation-based methods

When calculation of exact inclusion probabilities is impossible or prohibitive but complete design information is available, Monte Carlo simulation methods can easily be used to obtain estimates of the inclusion probabilities. Denote the completely specified probability sampling design by *p*. The simulation-based method is straightforward: select *K* independent samples, all following the same sampling design *p*; let M_i be the number of samples which include unit *i*. Then the first order inclusion probability $\pi_i = P(i \in s)$ can be estimated by $\pi_i^* = M_i / K$. For a particular *i*, the M_i follows a binomial distribution and the π_i^* satisfies $E(\pi_i^*) = \pi_i$ and $\mathrm{Var}(\pi_i^*) \le (4K)^{-1}$. Suppose for instance that we can afford to take *K* as big as 25×10^6, then $P(|\pi_i^* - \pi_i| < 0.001) \ge 0.99$ for any given π_i.

A more relevant measure of the accuracy of simulation-based methods is the performance of the Horvitz-Thompson estimator using the simulated inclusion probabilities. Let $\hat{T} = \sum_{i \in s} y_i / \pi_i$ and $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$. For a given sample, the relative bias of using $\tilde{T}$ in place of $\hat{T}$ is defined as $(\hat{T} - \tilde{T}) / \hat{T}$. Without loss of generality, we assume $y_i \ge 0$ for all *i*. It is shown in Appendix B that for any $\varepsilon > 0$ and the given sample *s*,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right). \qquad (2.1)$$

Note that $\sum_{i \in s} (1/\pi_i)$ is the Horvitz-Thompson estimator of the population size *N*, a practical lower bound for $P(|\hat{T} - \tilde{T}| / \hat{T} \le \varepsilon)$ with a small ε is given by

$$\Delta = 1 - \frac{2(N-n)}{K\varepsilon^2}. \qquad (2.2)$$

If one requires that $\varepsilon = 0.01$ and $\Delta = 0.98$, then for $N - n = 100$ the (theoretical) number of independent samples required for the simulation is $K = 10^8$. Since the lower bound given by (2.1) is conservative, and valid for any response variable, one would expect that a smaller K with values around 10^7 or even 10^6 should work well for most practical scenarios where $N - n \le 100$. This is supported by numerical examples presented in Section 3.

Estimation of the second order inclusion probabilities $\pi_{ij} = P(i, j \in s)$ imposes no additional difficulty except that the total number of simulated samples, K, required to achieve the same level of relative accuracy as for the first order case is bigger. Let M_{ij} be the number of simulated samples among the K independent samples which include both i and j. Let $\pi_{ij}^* = M_{ij}/K$ be the estimate for π_{ij}. Suppose the goal is to estimate a quadratic population quantity

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} q(y_i, y_j).$$

The Horvitz-Thompson type estimators of Q using π_{ij} or π_{ij}^* are respectively given by

$$\hat{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}} \text{ and } \tilde{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}^*}.$$

Following the same argument as that which leads to (2.1), we can show that

$$P\left(\frac{|\hat{Q} - \tilde{Q}|}{\hat{Q}} \le \varepsilon \right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \sum_{j \in s} \frac{1}{\pi_{ij}} - n^2 \right). \qquad (2.3)$$

Note that $\sum_{i \in s} \sum_{j \in s} (1/\pi_{ij})$ is a design-unbiased estimator of N^2, a practical lower bound for $P(|\hat{Q} - \tilde{Q}|/\hat{Q} \le \varepsilon)$ is given by $1 - 2(N+n)(N-n)/(K\varepsilon^2)$. Comparing this with Δ given by (2.2), it is apparent that we need a much bigger K to achieve the same lower bound, although in both cases the lower bounds are conservative, and the actual K required can be smaller. On the other hand, second order inclusion probabilities are used for the estimation of second order parameters such as the population variance or the variance of a linear estimator. The desired estimation accuracy is less critical than that for first order parameters such as the population total or mean, and therefore a number in between 10^6 and 10^7 for K should be acceptable for many practical situations.

The most critical issue for simulation-based methods is obviously the feasibility of computational implementation. Among other things, it depends largely on the chosen value of K, the complexity of the sampling design, and the computational power available. If $K = 10^6$ and one would like to have the simulation-based results within ten hours, then it is necessary to take 28 simulated samples for every single second. The randomized systematic PPS sampling is the most efficient unequal probability without replacement sampling procedure in terms of computational implementation. It only involves a simple random ordering and selecting a random starting point. Most other competing procedures involve either rejective methods or complicated sequential selections. It takes much longer to select simulated samples with these methods. A comparison of CPU times for computing the simulated π_i between the randomized systematic PPS sampling and the Rao-Sampford unequal probability sampling design is given in Section 4.

3. Numerical examples

The design information used in this section is adapted from the ITC China survey. The number of Street Districts (top level clusters) in each of the seven cities involved in the survey ranges from $N = 20$ to $N = 120$. Within each city $n = 10$ districts are selected using the randomized systematic PPS sampling method. In the case of refusals, substitute districts are selected from the ones not included in the initial sample, again using the randomized systematic PPS sampling method. For the purpose of illustration we use the design information from the smallest city (*i.e.*, $N = 20$). Additional comments on cases where N is large are given in Section 4.

3.1 First order inclusion probabilities

We first demonstrate the accuracy of the simulated π_i when the exact values of π_i are known. We then investigate the impact of substitution of units on the final π_i and the performance of the Horvitz-Thompson estimator for a population total using the simulated π_i. The simulated inclusion probabilities under substitution of units are compared to those assuming the modified design is still PPS sampling.

Example 1. Simulation-based π_i^* when there is no refusal. In this case the exact values of π_i are given by $\pi_i = nz_i$.

(i) Exact values of π_i:

```
0.5840 0.5547 0.6702 0.5331 0.3085 0.2652 0.3930 0.4180 0.6952 0.3471
0.5993 0.5393 0.8240 0.6868 0.4469 0.2191 0.4237 0.4180 0.7567 0.3163
```

(ii) Simulated π_i^*, $K = 10^5$:

```
0.5828 0.5545 0.6656 0.5339 0.3071 0.2656 0.3929 0.4205 0.6969 0.3474
0.6009 0.5429 0.8227 0.6865 0.4446 0.2186 0.4215 0.4179 0.7569 0.3194
```

(iii) Simulated π_i^*, $K = 10^6$:

```
0.5836 0.5558 0.6701 0.5336 0.3081 0.2654 0.3931 0.4180 0.6950 0.3469
0.5994 0.5394 0.8242 0.6864 0.4469 0.2186 0.4237 0.4172 0.7569 0.3166
```

The simulated π_i^* matches π_i to the second decimal point for $K = 10^5$ and to the third for $K = 10^6$ for most cases.

Example 2. To assess the performance of the Horvitz-Thompson (HT) estimator for a population total using the true π_i and the simulated π_i^* from Example 1, we generated the response variable from the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, N,$ where x_i is the size variable and ε_i are independent and identically normally distributed with mean 0 and variance σ^2. We considered three populations (three values of σ^2) where the population correlation coefficients between *x* and *y* are respectively 0.3, 0.5 and 0.8. For each of the three populations, $B = 2{,}000$ repeated samples of size $n = 10$ were selected using the randomized systematic PPS sampling, and for each sample three HT estimators were computed using the true π_i, the simulated π_i^* with $K = 10^5$ and the π_i^* with $K = 10^6$, respectively. The results, not reported here to save space, showed that all three HT estimators have relative bias less than 0.04% and almost identical mean squared errors.

Example 3. When there are refusals in the initial PPS sample and substitute units are selected from units not included in the initial sample using the same PPS sampling procedure, there are two questions of interest: (1) how to compute the inclusion probabilities π_i for the final sample; and (2) to what extent the substitution procedure has altered the original PPS sampling design. We can compute the simulated π_i^* and compare them with $\tilde{\pi}_i$ obtained by assuming a PPS sampling after the refusing units are removed from the sampling frame. In simulating the π_i^*, we assume for simplicity that there is no possible refusal from any unit outside the initial sample, and hence there is no refusal among the substitute units. The number of replications *K* is chosen as 10^6 for the simulation. We consider two scenarios where there are three refusing units in the population, and all are among the initial sample of size $n = 10$.

(i) Three large units refuse: Simulated π_i^* (first two rows) versus $\tilde{\pi}_i$ (last two rows) assuming PPS.

```
0.7231 0.6981 0.7947 0.6773 0.4354 0.3811 0.5339 0.5619 0.0000 0.4815
0.7363 0.6826 0.0000 0.8070 0.5919 0.3210 0.5678 0.5615 0.0000 0.4441

0.7560 0.7182 0.8677 0.6901 0.3994 0.3434 0.5088 0.5412 0.0000 0.4494
0.7759 0.6983 0.0000 0.8892 0.5786 0.2837 0.5486 0.5412 0.0000 0.4096
```

(ii) Three small units refuse: Simulated π_i^* (first two rows) versus $\tilde{\pi}_i$ (last two rows) assuming PPS.

```
0.6326 0.6049 0.7167 0.5829 0.0000 0.0000 0.4415 0.4668 0.7406 0.3937
0.6482 0.5901 0.8558 0.7330 0.4965 0.0000 0.4728 0.4664 0.7976 0.3590

0.6343 0.6025 0.7280 0.5790 0.0000 0.0000 0.4268 0.4540 0.7550 0.3770
0.6510 0.5858 0.8949 0.7459 0.4854 0.0000 0.4602 0.4540 0.8218 0.3436
```

It is apparent that the sizes of the refusing units have dramatic impact on the distribution of the final inclusion probabilities. If one ignores the alteration of the sampling design due to substitution of units and treats the design as if it is still a PPS sampling, then the inclusion probabilities for large units are inflated and the role of small units is down-played. This trend is more pronounced when there are large units among the refusals, *i.e.*, case (i) where $\pi_{14}^* = 0.8070$ compared to $\tilde{\pi}_{14} = 0.8897$ and $\pi_{16}^* = 0.3210$ to $\tilde{\pi}_{16} = 0.2837$.

3.2 Second order inclusion probabilities

There have been considerable research activities on the randomized systematic PPS sampling, mainly for obtaining second order inclusion probabilities π_{ij} and variance estimators. Hartley and Rao (1962) derived exact formulas for the π_{ij} when $n = 2$ and $N = 3$ or $N = 4$; Connor (1966) extended the results and derived the exact formula for general *n* and *N*, and the related computational procedure was later implemented in the Fortran language by Hidiroglou and Gray (1980). The procedure is quite heavy as evidenced by the 165 lines of Fortran code.

The most intriguing result is probably the asymptotic approximation to π_{ij} derived by Hartley and Rao (1962). In a recent paper Kott (2005) showed that the variance estimator of a Horvitz-Thompson estimator based on the Hartley-Rao approximation not only performs well under the design-based framework but also has good model-based properties. The Hartley-Rao approximation was initially derived based on the assumption that *n* is fixed and *N* is large and is correct to the order of $O(N^{-4})$ (Hartley and Rao 1962: Equation (5.15) on page 369). In a private conversation with J.N.K. Rao during the 23rd International Methodological Symposium of Statistics Canada, he pointed out that the approximation is still valid even if *n* is large, as long as n/N is small. For cases where *N* is not large and/or n/N is not small, such as the ITC China survey example considered here, the goodness of the Hartley-Rao approximation has not been documented.

When the randomized systematic PPS sampling procedure is altered due to substitution of units, it is virtually impossible to derive the second order inclusion probabilities or some sort of approximations. With the simulation-based approach, however, it remains straightforward to obtain very reliable estimates of the π_{ij} through a large number of simulated samples, given that the altered sampling procedure is completely specified. In what follows we examine the performance of variance estimators using the simulated π_{ij}^* when there is no alteration to the randomized systematic PPS sampling procedure. In this case $\pi_i = nz_i$ and the Hartley-Rao approximation $\tilde{\pi}_{ij}$ to π_{ij} can also take part in the comparison.

Example 4. We first compare π_{ij}^* to $\tilde{\pi}_{ij}$ for each of the individual entries. To save space, we only present the results for $i = 1, \ldots, 5$ and $j = 1, \ldots, 10,$ which are sufficient to

show the general picture. The Hartley-Rao approximation $\tilde{\pi}_{ij}$ is very close to the simulated π^*_{ij}, matching to the second decimal point for the majority of the entries. This is clearly an interesting observation given that $N = 20$ and $n = 10$.

(i) Simulated π^*_{ij}, $K = 10^6$:

```
0.0000 0.3121 0.3821 0.2975 0.1669 0.1442 0.2116 0.2249 0.3975 0.1873
0.3121 0.0000 0.3623 0.2816 0.1590 0.1372 0.2025 0.2141 0.3766 0.1784
0.3821 0.3623 0.0000 0.3469 0.1899 0.1640 0.2483 0.2659 0.4586 0.2153
0.2975 0.2816 0.3469 0.0000 0.1523 0.1312 0.1938 0.2061 0.3606 0.1717
0.1669 0.1590 0.1899 0.1523 0.0000 0.0742 0.1124 0.1197 0.1968 0.0988
```

(ii) Hartley-Rao approximation $\tilde{\pi}_{ij}$:

```
0.0000 0.3079 0.3769 0.2952 0.1668 0.1427 0.2143 0.2286 0.3921 0.1884
0.3079 0.0000 0.3569 0.2795 0.1579 0.1351 0.2029 0.2164 0.3712 0.1784
0.3769 0.3569 0.0000 0.3421 0.1932 0.1654 0.2484 0.2649 0.4544 0.2183
0.2952 0.2795 0.3421 0.0000 0.1514 0.1296 0.1946 0.2075 0.3559 0.1710
0.1668 0.1579 0.1932 0.1514 0.0000 0.0732 0.1099 0.1172 0.2010 0.0966
```

Example 5. For second order inclusion probabilities the main focus is on variance estimation. With fixed sample size, an unbiased variance estimator for the Horvitz-Thompson estimator $\hat{Y}_{\mathrm{HT}} = \sum_{i\in s} y_i/\pi_i$ is given by the well-known Yates-Grundy format,

$$v(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \qquad (3.1)$$

We consider the three synthetic populations described in *Example 2*. The true variance $V = \mathrm{Var}(\hat{Y}_{\mathrm{HT}})$ is obtained through simulation using $B = 10^5$ simulated samples and is computed as $B^{-1}\sum_{b=1}^{B}(\hat{Y}_b - Y)^2$, where Y is the true population total and $\hat{Y}_b$ is the Horvitz-Thompson estimator of Y computed from the b^{th} simulated sample. Three variance estimators in the form of (3.1), denoted respectively by v_1, v_2 and v_3, are examined, with the π_{ij} in (3.1) being respectively replaced by the Hartley-Rao approximation $\tilde{\pi}_{ij}$, the simulated π^*_{ij} for $K = 10^5$ and the π^*_{ij} for $K = 10^6$. The performance of these estimators is measured through the simulated relative bias $\mathrm{RB} = B^{-1}\sum_{b=1}^{B}(v^{(b)} - V)/V$ and the simulated instability $\mathrm{INST} = \{B^{-1}\sum_{b=1}^{B}(v^{(b)} - V)^2\}^{1/2}/V$, where $v^{(b)}$ is the variance estimate computed from the b^{th} sample, using another set of $B = 10^5$ independent samples. The results are summarized in Table 1 below. The three populations are indicated by the correlation coefficient ρ between y and x.

Table 1 Relative bias and instability of variance estimators

	RB(%)			INST		
Population	v_1	v_2	v_3	v_1	v_2	v_3
$\rho = 0.30$	6.1%	1.4%	-0.3%	0.66	0.65	0.65
$\rho = 0.50$	4.3%	2.5%	-1.1%	0.42	0.44	0.42
$\rho = 0.80$	2.6%	1.2%	-0.2%	0.61	0.60	0.60

In terms of relative bias, all three variance estimators are acceptable, with the one (v_1) based on the Hartley-Rao approximation $\tilde{\pi}_{ij}$ having the largest bias. For variance estimators using the simulated π^*_{ij}, increasing the value of K from 10^5 (*i.e.*, v_2) to 10^6 (*i.e.*, v_3) makes the bias to be negligible, although the one with $K = 10^5$ is clearly acceptable in practice. All three versions of the variance estimator have similar measures in terms of instability.

4. Some additional remarks

In theory, the simulation-based method for computing inclusion probabilities is applicable to any sampling design, as long as the complete design information is available. It is an effective approach to handling more complex substitution scenarios or other types of modifications to the original design. In the ITC China survey, one of the refusing units has to be substituted by a unit from a particular region of the city due to workload constraints and field work restrictions. In a Canadian national survey of youth, there were second and third round refusing units (schools) and hence substitute units before achieving the targeted sample size. As pointed out by an Associate Editor, a similar situation was also reported in the 57th Round of the National Sample Survey Organization, Government of India (www.mospi.gov.in) where a modification was made to the circular systematic sampling with probability proportional to size in order to select two distinct sub-samples. Gray (1973) described a method on increasing the sample size (number of psu's) when the initial sample was selected by the randomized systematic PPS method. Calculation of second order inclusion probabilities under the proposed procedure is difficult even for a very small sample size. In all these cases analytic solutions to the inclusion probabilities are either difficult to use or not available but the simulation-based approach can be applied without any extra difficulty.

The recent paper by Fattorini (2006) discussed the use of the simulation-based method for spatial sampling where the units are selected sequentially. When a PPS sampling design is altered due to one or more rounds of substitution of units, the modified design can also be viewed as sequential. Our theoretical results on the accuracy of simulation-based methods, however, are different from those of Fattorini. We have used a conditional argument and proposed to assess the performance of the estimator using the simulated inclusion probabilities for a given sample, which is of interest for practical applications.

The central issue related to simulation-based methods is the feasibility of computational implementation. The randomized systematic PPS sampling has a major advantage in computational efficiency. The Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967), for instance, is another popular PPS sampling procedure. It has several desirable features such as closed form expressions for the second order inclusion probabilities and

is more efficient than the randomized systematic PPS sampling (Asok and Sukhatme 1976). The following is a comparison of CPU times between the randomized systematic PPS sampling and the Rao-Sampford PPS sampling for simulating the first order inclusion probabilities. The sample size is fixed at $n = 10$ and the number of simulated samples is $K = 10^6$. The results are obtained using R on a dual-processor unix machine.

N	Systematic PPS	Rao-Sampford PPS
200	4.7 hours	7.5 hours
100	2.5 hours	5.0 hours
50	1.6 hours	4.4 hours
20	1.2 hours	8.9 hours

It is interesting to note that, although in general the Rao-Sampford procedure takes longer time to obtain the results, it takes much longer for the case of $N = 20$. This is because the Rao-Sampford method uses a rejective procedure and it usually takes many rejections to arrive at a final sample when the sampling fraction n/N is large. The randomized systematic PPS sampling, on the other hand, is not affected by this and the simulation-based method can provide results with desired accuracy in a timely fashion for $N = 400$ or even bigger. Several R/S-PLUS functions and major codes for the proposed approach are included in Appendix C and are applicable to other substitution scenarios after minor modifications.

One of the reasons for the use of the randomized systematic PPS sampling in selecting upper level clusters in the ITC China survey is that the final design is self-weighting. An interesting question arises when there are refusals: how to select the substitute units such that the final altered sampling design is still (approximately) self-weighting? In some other circumstances such as rotating samples, this is achievable; see, for instance, Fellegi (1963). How to accomplish this goal with the ITC China survey design is currently under investigation.

Acknowledgements

This research is partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors also thank the International Tobacco Control (ITC) Policy Evaluation Project and the ITC China Survey Project for assistance and support. The ITC project is supported in part by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897). Funding for the ITC China project is provided by the Ministry of Health and the Ministry of Finance of China.

Appendix A

A direct calculation under random refusal

Under the randomized systematic PPS sampling design and assuming random refusal, it is possible in principle to calculate the inclusion probabilities under a substitution rule directly. The starting point is to enumerate all possible initial samples and their probabilities based on the particular population order used to select the initial sample.

Recall that $A_0 = 0, A_j = \sum_{i=1}^{j} (nz_i)$ and $A_N = n$. For a chosen uniform starting value $u \in [0, 1]$, unit j is to be selected if

$$A_{j-1} \le u + k < A_j \tag{5.1}$$

for some $k = 0, 1, \ldots, n-1$. Let k_j be the largest integer less than A_j, and let the remainder e_j be given by $e_j = A_j - k_j$. Let $0 < e_{(1)} \le e_{(2)} \le \ldots \le e_{(N)}$ be the order statistics of the remainders, and let $k_{(1)}, \ldots, k_{(N)}$ be the corresponding k_j's. Note that $e_{(N)} = 1$. We could then generate N possible samples $s_1, \ldots, s_N$ with respective probabilities

$$e_{(1)}, e_{(2)} - e_{(1)}, \ldots, e_{(N)} - e_{(N-1)},$$

some of which may be 0. We begin by generating s_1. From each $j = 1, \ldots, N$, put j in s_1 if $A_{j-1} \le k < A_j$ for some $k = 0, 1, \ldots, n-1$, *i.e.*, s_1 is selected using $u = 0$ in (5.1). As we move u from 0 to 1, different possible samples can be identified sequentially. Now given $s_1, \ldots, s_m$, let s_{m+1} be the same as s_m except that the $(k_{(m)} + 1)^{\text{th}}$ element is advanced by 1. For example, suppose $n = 4$ and $s_m = \{1, 3, 6, 9\}$, and suppose $k_{(m)} = 0$, then $s_{m+1} = \{2, 3, 6, 9\}$. On the other hand, if $k_{(m)} = 2$, then $s_{m+1} = \{1, 3, 7, 9\}$. The sample s_{m+1} will have probability $e_{(m+1)} - e_{(m)}$.

By construction, $\pi_i = nz_i$ for $i = 1, \ldots, N$. If only first and second order inclusion probabilities are desired, a similar but simpler algorithm can be used to calculate the second order inclusion probabilities directly, conditional on the initial order. However, for applications where the probabilities of all samples are needed, the sample generation algorithm can be carried out. For example, for small populations, it is then also possible to calculate the first order inclusion probabilities when there is refusal and substitution. Suppose we first select a sample of size n with randomized systematic PPS sampling. Suppose n_1 of these agree to respond and an additional $n_2 = n - n_1$ are selected, again using randomized systematic PPS sampling, from those units not sampled the first time. Assume for simplicity that refusal in the first sample occurs at random, and that there is no refusal in the second substitute sample. Note that this is a different assumption from the one used in Example 3, where the set of refusals is considered to be non-random. The

inclusion probability for unit *i*, conditional on the assumed initial population order, is

$$nz_i \times \frac{n_1}{n} + \sum_{m:i\notin s_m} p_1(s_m)\frac{n_2 z_i}{\sum_{j:j\notin s_m} z_j}. \qquad (5.2)$$

The outer sum is taken over all samples s_m of size n, generated according to the procedure described above but without having unit i, with probabilities $p_1(s_m) = e_{(m)} - e_{(m-1)}$. The inner sum involved in the denominator is taken over all j not included in s_m from the outer sum. The unconditional inclusion probability can be obtained by appropriate averaging over all population orders which give distinct values. Clearly this is feasible only when the population is small, or when z takes a small number of values.

Appendix B

Derivation of (2.1)

In this appendix we show that for any $\varepsilon > 0$ and a given sample s,

$$P\left(\frac{|\hat{T}-\tilde{T}|}{\hat{T}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i\in s}\frac{1}{\pi_i} - n\right),$$

where $\hat{T} = \sum_{i\in s} y_i/\pi_i, \tilde{T} = \sum_{i\in s} y_i/\pi_i^*$, and π_i^* are the simulated first order inclusion probabilities based on K independent samples. Noting that $E(\pi_i^*) = \pi_i$ and $\mathrm{Var}(\pi_i^*) = \pi_i(1-\pi_i)/K$, by Chebyshev's inequality we have $P(|\pi_i^* - \pi_i| > c) \le \pi_i(1-\pi_i)/(Kc^2)$ for any $c > 0$. It follows that

$$\begin{aligned}
&P\left(\frac{|\pi_i^*-\pi_i|}{\pi_i^*} > \varepsilon\right) \\
&= P(\pi_i^* - \pi_i > \pi_i^*\varepsilon) + P(\pi_i^* - \pi_i < -\pi_i^*\varepsilon) \\
&= P(\pi_i^* - \pi_i > \varepsilon\pi_i/(1-\varepsilon)) + P(\pi_i^* - \pi_i < -\varepsilon\pi_i/(1+\varepsilon)) \\
&\le P(|\pi_i^* - \pi_i| > \varepsilon\pi_i/(1-\varepsilon)) + P(|\pi_i^* - \pi_i| > \varepsilon\pi_i/(1+\varepsilon)) \\
&\le \frac{(1-\varepsilon)^2\pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2} + \frac{(1+\varepsilon)^2\pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2} \\
&= \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\frac{1}{\pi_i} - 1\right).
\end{aligned}$$

If $y_i \ge 0$ for all i, then

$$|\hat{T}-\tilde{T}| \le \sum_{i\in s}\frac{y_i}{\pi_i}\frac{|\pi_i^*-\pi_i|}{\pi_i^*} \le \max_{i\in s}\left\{\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\right\}\hat{T}.$$

For any $\varepsilon > 0$ and the given sample s,

$$\begin{aligned}
P\left(\frac{|\hat{T}-\tilde{T}|}{\hat{T}} \le \varepsilon\right) &\ge P\left(\max_{i\in s}\left\{\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\right\} \le \varepsilon\right) \\
&\ge 1 - \sum_{i\in s} P\left(\frac{|\pi_i^*-\pi_i|}{\pi_i^*} > \varepsilon\right) \\
&\ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i\in s}\frac{1}{\pi_i} - n\right).
\end{aligned}$$

Appendix C

R/S-PLUS Implementation

C1. An R function for randomized systematic PPS sampling.

The input variables of the function are `x`: the population vector of size variable and `n`: the sample size. The function `syspps` returns the set of *n* selected units.

```
syspps<-function(x,n){
N<-length(x)
U<-sample(N,N)
xx<-x[U]
z<-rep(0,N)
for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
r<-runif(1)
s<-numeric()
for(i in 1:N){
if(z[i]>=r){
s<-c(s,U[i])
r<-r+1
 } }
return(s[order(s)])
}
```

C2. An R function for simulating the second order inclusion probabilities.

The input variables of the function are `x`: the population vector of size variable and `s`: the set of labels of units in the sample. The default sampling procedure is the randomized systematic PPS sampling method and the number of repeated samples is $K = 10^6$. The function `piij` returns an $n \times n$ matrix with the $(ij)^{\text{th}}$ entry being the simulated $\pi_{ij}^*, i, j \in s$.

```
piij<-function(x,s){
N<-length(x)
n<-length(s)
p<-matrix(0,n,n)
for(k in 1:1000000){
ss<-syspps(x,n)
for(i in 1:(n-1)){
for(j in (i+1):n){
if(min(abs(ss-s[i]))+min(abs(ss-s[j]))==0)
    p[i,j]<-p[i,j]+1
  } } }
p<-(p+t(p))/1000000
return(p)
   }
```

C3. An R function for PPS sampling under substitution of units.

```
syspppssub<-function(x,n,refus){
s<-syspps(x,n)
sub<-numeric()
for (i in 1:n){
if(min(abs(s[i]-refus))==0) sub<-c(sub,i)
}
m<-length(sub)
if(m>0){
s<-s[-sub]
U1<-(1:length(x))[-c(refus,s)]
x1<-x[-c(refus,s)]
s1<-syspps(x1,m)
s<-c(s,U1[s1])
}
return(s[order(s)])
}
```

The default procedure for the selection of the initial sample and the substitute sample is the randomized systematic PPS sampling. The following R function `syspppssub` is used for simulating the inclusion probabilities under substitution of units. The input variables are `x`: the population vector of size variable, `n`: the sample size, and `refus`: the set of refusing units from the initial sample. The function returns a set of units for the final sample.

C4. R codes for simulating the π_i under substitution of units.

```
pi<-rep(0,N)
for(i in 1:1000000){
s<-syspppssub(x,n,refus)
for(j in 1:N){
if(min(abs(s-j))==0) pi[j]<-pi[j]+1
 } }
pi<-pi/1000000
```

References

Asok, C., and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

Chen, X.H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Deville, J.C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

Goodman, R., and Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Gray, G.B. (1973). On increasing the sample size (number of psu's). Technical Memorandum, Statistics Canada.

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

Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statist. Association*, 3, 173-180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 11-17
Statistics Canada, Catalogue No. 12-001-X

Small area estimation under a two-level model

Mahmoud Torabi and J.N.K. Rao [1]

Abstract

Lehtonen and Veijanen (1999) proposed a new model-assisted generalized regression (GREG) estimator of a small area mean under a two-level model. They have shown that the proposed estimator performs better than the customary GREG estimator in terms of average absolute relative bias and average median absolute relative error. We derive the mean squared error (MSE) of the new GREG estimator under the two-level model and compare it to the MSE of the best linear unbiased prediction (BLUP) estimator. We also provide empirical results on the relative efficiency of the estimators. We show that the new GREG estimator exhibits better performance relative to the customary GREG estimator in terms of average MSE and average absolute relative error. We also show that, due to borrowing strength from related small areas, the EBLUP estimator exhibits significantly better performance relative to the customary GREG and the new GREG estimators. We provide simulation results under a model-based set-up as well as under a real finite population.

Key Words: BLUP estimator; GREG estimator; Mean squared error; Random effects; Small area means.

1. Introduction

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area statistics. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas, in particular, model-based indirect estimators. Unit level random effect models, including nested error linear regression models and two-level models, are often used in small area estimation to obtain efficient model-based estimators of small area means. Rao (2003) gives a comprehensive account of model-based small area estimation.

A two-level model is given by

$$y_{ij} = x'_{ij}\beta_i + e_{ij};$$

$$\beta_i = Z_i\beta + \nu_i,\ j = 1, ..., N_i;\ i = 1, ..., m \qquad (1)$$

where N_i is the number of units in the i^{th} area ($i = 1, ..., m$), y_{ij} is the response and x_{ij} is a $p \times 1$ vector of unit level covariates attached to the j^{th} unit in the i^{th} area. Further, Z_i is a $p \times q$ matrix of area level covariates, β is a $q \times 1$ vector of regression parameters, ν_i's are independent random vectors with mean zero and covariance Σ_ν, and e_{ij}'s are independent random variables with mean zero and variance σ_e^2 and independent of ν_i's. We can express the mean $\bar{Y}_i$ of i^{th} area as

$$\bar{Y}_i \approx \mu_i = \bar{X}'_i(Z_i\beta + \nu_i),$$

assuming N_i is large, where $\bar{X}_i$ is the known population mean of x_{ij} in the i^{th} area. The sample values $\{(y_{ij}, x_{ij}),\ j = 1, ..., n_i;\ i = 1, ..., m\}$ are assumed to obey the model (1), that is, there is no sample selection bias. The model for the sample is then given by

$$y_{ij} = x'_{ij}(Z_i\beta + \nu_i) + e_{ij},\ j = 1, ..., n_i;\ i = 1, ..., m. \qquad (2)$$

In matrix notation, (2) may be written as

$$y_i = X_i(Z_i\beta + \nu_i) + e_i,\ i = 1, ..., m$$

with $\text{Var}(y_i) = V_i = X_i\Sigma_\nu X'_i + \sigma_e^2 I_{n_i}$, where y_i is a $n_i \times 1$ vector and X_i is a $n_i \times p$ matrix. The two-level model (2) was first introduced by Moura and Holt (1999) in the context of small area estimation. This model effectively integrates the use of unit level and area level covariates into a single model, by modeling the random slopes β_i in (1) in terms of area level covariates Z_i.

Lehtonen and Veijanen (1999) proposed a model-assisted new generalized regression (GREG) estimator of a small area mean under the two-level model. Lehtonen and Veijanen (1999) showed that the new GREG estimator based on model (1) performs better than the customary GREG estimator based on a model with fixed $\beta_i = Z_i\beta$. Moura and Holt (1999) obtained the best linear unbiased prediction (BLUP) estimator of the small area mean μ_i and its MSE under the two-level model (2); see Section 2. Lehtonen, Särndal, and Veijanen (2003) studied the effect of model choice on different types of estimators (synthetic, GREG, and composite) of small area means.

In Section 3, we first derive the mean squared error (MSE) of the new GREG estimator and the customary GREG estimator (Section 2) under the two-level model (2), assuming known model parameters. We then compare the MSE of the GREG, new GREG and BLUP estimators, and obtain an explicit expression for the increase in MSE of the new GREG estimator relative to the MSE of the BLUP estimator. In Section 4, we provide empirical results on the relative efficiency of the estimators when the model parameters are estimated. We used a model-based set-up as

1. Mahmoud Torabi, Department of Pediatrics, University of Alberta, Edmonton, Alberta, T6G 2J3; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

well as a real finite population for the simulation study. Finally, some concluding remarks are given in Section 5.

2. BLUP and GREG estimators

The two-level model (2) is a special case of a general linear mixed model with block diagonal covariance structure. Therefore, assuming known model parameters, we may calculate the BLUP estimator of μ_i as

$$\tilde{\mu}_i^B = \bar{X}_i'(Z_i\beta + \tilde{\nu}_i), \tag{3}$$

where

$$\tilde{\nu}_i = \Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta),$$

and the superscript B on $\tilde{\mu}_i$ stands for BLUP estimator (Rao 2003, page 107). Similarly, the BLUP estimator of a non-sampled observation j in i^{th} area can be written as

$$\tilde{y}_{ij} = x_{ij}'(Z_i\beta + \tilde{\nu}_i). \tag{4}$$

On the other hand, a model-assisted GREG estimator of μ_i (or $\bar{Y}_i$) is given by

$$\tilde{\mu}_i^G = \frac{1}{N_i}\left[\sum_{j=1}^{N_i}\tilde{y}_{ij} + \sum_{j=1}^{n_i} w_{ij}(y_{ij} - \tilde{y}_{ij})\right],\ i = 1, \ldots, m \tag{5}$$

where $\tilde{y}_{ij}$ is the predictor of y_{ij} under the assumed model, and w_{ij} is the survey weight which equals N_i/n_i in the case of simple random sampling (SRS) within areas. We focus on SRS within areas in this paper.

Using (5) with $\tilde{y}_{ij} = x_{ij}'Z_i\beta$ as the predictor of y_{ij} under the model (1) with fixed $\beta_i = Z_i\beta$, we can write the customary GREG estimator as

$$\tilde{\mu}_i^G = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'Z_i\beta, \tag{6}$$

where the superscript G on $\tilde{\mu}_i$ denotes GREG (Särdnal, Swensson and Wretman 1992, page 225), $\bar{y}_i$ is sample mean of y_{ij} in the i^{th} area, and $\bar{x}_i$ is the sample mean of x_{ij} in the i^{th} area, respectively. Using the predictor (4) based on the two-level model (1) in (5), we get a new GREG estimator of μ_i (or $\bar{Y}_i$) as

$$\begin{aligned}\tilde{\mu}_i^{\text{LV}} &= [\bar{X}_i'(Z_i\beta + \tilde{\nu}_i) + (\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))] \\ &= \bar{y}_i + (\bar{X}_i - \bar{x}_i)'(Z_i\beta + \tilde{\nu}_i),\end{aligned} \tag{7}$$

where the superscript LV on $\tilde{\mu}_i$ denotes that it was first proposed by Lehtonen and Veijanen (1999). The estimators $\tilde{\mu}_i^B$, $\tilde{\mu}_i^G$ and $\tilde{\mu}_i^{\text{LV}}$ are linear in the y_{ij}'s and unbiased under the two-level model (1). In practice, we replace the parameters β, Σ_v and σ_e^2 in (3), (6) and (7) by suitable estimators. The resulting estimators are denoted by $\hat{\mu}_i^B$, $\hat{\mu}_i^G$ and $\hat{\mu}_i^{\text{LV}}$ respectively, where $\hat{\mu}_i^B$ is the empirical BLUP (EBLUP) estimator. Under normality assumption, $\hat{\mu}_i^B$ is the empirical best (EB) estimator. The EBLUP estimator of $\bar{Y}_i$ is given in Section 4.2.2. Note that $\hat{\mu}_i^G$ and $\hat{\mu}_i^{\text{LV}}$ are valid as estimators of $\bar{Y}_i$.

3. Mean squared error

The mean squared error (MSE) of the customary GREG estimator $\tilde{\mu}_i^G$ under the two-level model can be written as

$$\begin{aligned}\text{MSE}(\tilde{\mu}_i^G) &= E(\tilde{\mu}_i^G - \mu_i)^2 \\ &= E[\bar{y}_i + (\bar{X}_i - \bar{x}_i)'Z_i\beta - \bar{X}_i'(Z_i\beta + \nu_i)]^2 \\ &= E[(\bar{x}_i - \bar{X}_i)'\nu_i + \bar{e}_i]^2 \\ &= (\bar{x}_i - \bar{X}_i)'\Sigma_v(\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i},\end{aligned}$$

as stated in Theorem 1.

Theorem 1. *The MSE of the GREG estimator* (*6*) *is given by*

$$\text{MSE}(\tilde{\mu}_i^G) = (\bar{x}_i - \bar{X}_i)'\Sigma_v(\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i}. \tag{8}$$

Further, we may write the MSE of the BLUP estimator $\tilde{\mu}_i^B$ as follows:

$$\begin{aligned}\text{MSE}(\tilde{\mu}_i^B) &= E(\tilde{\mu}_i^B - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{\nu}_i - \nu_i)]^2 \\ &= \bar{X}_i'\left(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v\right)\bar{X}_i,\end{aligned}$$

as stated in Theorem 2.

Theorem 2. *The MSE of the BLUP estimator* (*3*) *is given by*

$$\text{MSE}(\tilde{\mu}_i^B) = \bar{X}_i'\left(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v\right)\bar{X}_i. \tag{9}$$

Theorem 3 gives the MSE of the new GREG estimator $\tilde{\mu}_i^{\text{LV}}$.

Theorem 3. *The MSE of the new GREG estimator* (*7*) *is given by*

$$\begin{aligned}\text{MSE}(\tilde{\mu}_i^{\text{LV}}) &= \text{MSE}(\tilde{\mu}_i^B) \\ &\quad + \left\{\bar{x}_i'\Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i'\Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i}\right\}.\end{aligned} \tag{10}$$

Proof of Theorem 3 is given in the Appendix.

By definition, we have $\text{MSE}(\tilde{\mu}_i^B) \le \text{MSE}(\tilde{\mu}_i^{\text{LV}})$ and (10) gives an explicit expression for the increase in MSE of $\tilde{\mu}_i^{\text{LV}}$ relative to the MSE of the BLUP estimator $\tilde{\mu}_i^B$.

4. Empirical results

4.1 Empirical comparison of MSE values

In order to study the efficiency of the new GREG estimator, we used data from Moura and Holt (1999) based on 38,740 households in the enumeration districts (small areas) in one county in Brazil. The income of household's head was treated as the response variable y. Two unit level independent variables were identified as the educational attainment of household's head (ordinal scale of 0-5) and the number of rooms in the household (1-11+). The following two-level model was assumed for this data:

$$y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + e_{ij},$$
$$j = 1, \ldots, N_i;\ i = 1, \ldots, m \qquad (11)$$

with

$$\beta_{i0} = \beta_0 + \nu_{i0};\ \beta_{i1} = \beta_1 + \nu_{i1};\ \beta_{i2} = \beta_2 + \nu_{i2}, \quad (12)$$

where

$$\nu = (\nu_{i0}, \nu_{i1}, \nu_{i2})' \overset{\text{iid}}{\sim} N_3(0, \Sigma_\nu),\ e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$$

and x_1 and x_2 respectively represent the number of rooms and the educational attainment of household's head (centered about their respective population means). Note that the model (12) for the random β_i-coefficients does not contain area level covariates Z.

Moura and Holt (1999) also studied another model with an area level covariate Z for modeling β_i's in (12). For this data, average number of cars per household in each area was used as a covariate z for modeling the random coefficients β_{i1} and β_{i2} corresponding to the variables x_1 and x_2, but not for the random intercept term, β_{i0}. The two-level model (11) with the area level covariate z is given by

$$y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij},$$
$$j = 1, \ldots, N_i;\ i = 1, \ldots, m \qquad (13)$$

with

$$\beta_{i0} = \beta_0 + \nu_{i0};$$
$$\beta_{i1} = \beta_1 + \alpha_1 z_i + \nu_{i1};\ \beta_{i2} = \beta_2 + \alpha_2 z_i + \nu_{i2}. \quad (14)$$

Moura and Holt (1999) fitted models (11)-(12) and (13)-(14) to the full data set mentioned above. We summarize their results in Table 1.

Table 1
Parameter estimates based on Moura and Holt's (1999) data set, where σ_0^2, σ_1^2 and σ_2^2 are the diagonal elements and σ_{01}, σ_{02} and σ_{12} are the off-diagonal elements of the covariance matrix Σ_ν

Parameter	Diagonal Covariance: Model (14) with z	Diagonal Covariance: Model (12) without z	General Covariance: Model (12) without z
β_0	8.442	8.688	8.456
β_1	0.451	1.321	1.223
β_2	0.744	2.636	2.596
α_1	3.779	-	-
α_2	1.659	-	-
σ_0^2	0.745	0.637	1.385
σ_1^2	0.237	0.471	0.234
σ_2^2	0.700	1.472	0.926
σ_{01}	-	-	0.354
σ_{02}	-	-	0.492
σ_{12}	-	-	0.333
σ_e^2	44.00	44.01	47.74

The area means of x_1 and x_2 were calculated from the whole data and treated as the population means $\bar{X}_{1i}$ and $\bar{X}_{2i}$. A random subsample of 10% of the records was selected from each small area. The overall sample size is $n = 3{,}876$ and the number of small areas is $m = 140$. Using the sample values of x_1, x_2 and z and the population means $\bar{X}_{1i}$ and $\bar{X}_{2i}$, we computed $\text{MSE}(\tilde{\mu}_i^G)$, $\text{MSE}(\tilde{\mu}_i^B)$ and $\text{MSE}(\tilde{\mu}_i^{\text{LV}})$ using (8), (9) and (10) respectively, treating the estimates of regression parameters, Σ_ν and σ_e^2 in Table 1 for the full data as true values. We then calculated the average MSE values over the areas:

$$\overline{\text{MSE}}_G = \frac{1}{m}\sum_{i=1}^m \text{MSE}(\tilde{\mu}_i^G),$$
$$\overline{\text{MSE}}_B = \frac{1}{m}\sum_{i=1}^m \text{MSE}(\tilde{\mu}_i^B)$$

and

$$\overline{\text{MSE}}_{\text{LV}} = \frac{1}{m}\sum_{i=1}^m \text{MSE}(\tilde{\mu}_i^{\text{LV}}).$$

We define the relative efficiency of $\tilde{\mu}^B$ over $\tilde{\mu}^G$ as EFF_B and the relative efficiency of $\tilde{\mu}^{\text{LV}}$ over $\tilde{\mu}^G$ as EFF_{LV}, where

$$\text{EFF}_B = \frac{\overline{\text{MSE}}_G}{\overline{\text{MSE}}_B};\qquad \text{EFF}_{\text{LV}} = \frac{\overline{\text{MSE}}_G}{\overline{\text{MSE}}_{\text{LV}}}.$$

We summarize the results in Tables 2 and 3. Tables 2 and 3 reveal that the new GREG estimator is slightly more efficient than the usual GREG estimator in terms of average MSE: $\text{EFF}_{\text{LV}} \le 112\%$. However, the new GREG estimator is substantially less efficient than the BLUP estimator, under the assumed two-level model. For example, for the model with z and diagonal covariance matrix (Table 2), $\text{EFF}_B = 292\%$ compared to $\text{EFF}_{\text{LV}} = 106\%$, and $\overline{\text{MSE}}_B = 0.62$ compared to $\overline{\text{MSE}}_{\text{LV}} = 1.72$.

Table 2
Comparison of small area estimators: relative efficiency (EFF) and average MSE (MSE) for the case of diagonal covariance matrix based on Moura and Holt's (1999) data set

Quality Measure	Model without z			Model with z		
	GREG	New GREG	BLUP	GREG	New GREG	BLUP
EFF	100%	112%	306%	100%	106%	292%
MSE	1.92	1.71	0.62	1.83	1.72	0.62

Table 3
Comparison of small area estimators: relative efficiency (EFF) and average MSE ($\overline{\text{MSE}}$) for the case of a general covariance matrix based on Moura and Holt's (1999) data set

Quality Measure	Model without z		
	GREG	New GREG	BLUP
EFF	100%	108%	253%
MSE	2.02	1.87	0.80

4.2 Simulation study

4.2.1 Simulation study under a model-based framework

In order to investigate the efficiency of the new GREG estimator with estimated model parameters, a small

simulation study based on the two-level models (11)-(12) and (13)-(14) was undertaken. We only considered a diagonal covariance structure Σ_v with diagonal elements σ_0^2, σ_1^2 and σ_2^2. We again used the data from Moura and Holt (1999). The estimates of $\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2, \sigma_0^2, \sigma_1^2, \sigma_2^2$ and σ_e^2 reported in Table 1 are treated as true values.

In our simulation study, we took (x_{1ij}, x_{2ij}, z_i) from Moura and Holt (1999) and then generated y_{ij} based on the models (11)-(12) and (13)-(14). By using the generated samples $(y_{ij}^{(b)}, x_{1ij}, x_{2ij}, z_i)$, $b = 1, ..., B = 1{,}000$, we calculated $\hat{\beta}^{(b)}$ by generalized least squares for the new GREG method as well as for the BLUP method. For the GREG method we used ordinary least squares to estimate β as $\hat{\beta}_{\text{ols}}^{(b)}$. In addition, $\hat{\Sigma}_v^{(b)}$ and $\hat{\sigma}_e^{2(b)}$ were computed based on the restricted maximum likelihood (REML) method. For each generated sample, we calculated

$$\mu_i^{(b)} = \bar{X}'_i(Z_i\beta + \nu_i^{(b)}),\ i = 1, ..., m;\ b = 1, ..., B.$$

We computed the new GREG estimator of $\mu_i^{(b)}$ as $\hat{\mu}_i^{\text{LV}(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)'(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})$, the GREG estimator of $\mu_i^{(b)}$ as $\hat{\mu}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)' Z_i \hat{\beta}_{\text{ols}}^{(b)}$ and the empirical BLUP (EBLUP) estimator of $\mu_i^{(b)}$ as $\hat{\mu}_i^{B(b)} = \bar{X}_i'(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})$, where $\hat{\nu}_i^{(b)} = \hat{\Sigma}_v^{(b)} X_i' \hat{V}_i^{-1(b)}(y_i^{(b)} - X_i Z_i \hat{\beta}^{(b)})$.

We then computed the average mean squared error ($\overline{\text{MSE}}_1$) and average absolute relative error ($\overline{\text{ARE}}_1$)

$$\overline{\text{MSE}}_1 = \frac{1}{m}\sum_i^m \text{MSE}_{1i} \text{ where } \text{MSE}_{1i} = B^{-1}\sum_{b=1}^B (\hat{\mu}_i^{(b)} - \mu_i^{(b)})^2,$$

$$\overline{\text{ARE}}_1 = \frac{1}{m}\sum_i^m \text{ARE}_{1i} \text{ where } \text{ARE}_{1i} = B^{-1}\sum_{b=1}^B |\hat{\mu}_i^{(b)} - \mu_i^{(b)}| / \mu_i^{(b)},$$

where $\hat{\mu}_i^{(b)}$ denotes $\hat{\mu}_i^{\text{LV}(b)}$, $\hat{\mu}_i^{G(b)}$ or $\hat{\mu}_i^{B(b)}$. We report the results in Table 4. Both models with area level covariate z and without z have slightly smaller values of $\overline{\text{MSE}}_1$ and $\overline{\text{ARE}}_1$ for the new GREG estimator relative to the GREG estimator. However, $\overline{\text{MSE}}_1$ and $\overline{\text{ARE}}_1$ are significantly smaller for the EBLUP estimator due to borrowing strength from related areas. Moreover, comparing Tables 2 and 4, we can see that the values of $\overline{\text{MSE}}_1$ in Table 4 are slightly larger than the corresponding values in Table 2 due to estimating model parameters.

Table 4
Comparison of small area estimators: average MSE ($\overline{\text{MSE}}_1$) and average absolute relative error ($\overline{\text{ARE}}_1$) under a model-based framework

Quality Measure	Model without z			Model with z		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{\text{MSE}}_1$	1.93	1.73	0.67	1.84	1.73	0.73
$\overline{\text{ARE}}_1$	0.14	0.13	0.08	0.13	0.12	0.08

4.2.2 Simulation study under a finite population framework

To study the performance of the estimators under a finite population framework, we created a synthetic finite population from the Brazilian data consisting of $n = 3{,}876$ sample values $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$. By duplicating the sample values $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$ five times, we treated the new (y, x_1, x_2, z) -data of size 19,380 as our real population.

We generated 500 independent samples ($B = 500$), each of size $n = 700$ and $n = 1{,}400$, by taking simple random samples of size $n_i = 5$ and $n_i = 10$ in each area $i = 1, ..., 140$. As before, for each sample we calculated $\hat{\beta}^{(b)}$ for the new GREG and the BLUP methods and $\hat{\beta}_{\text{ols}}^{(b)}$ for the GREG method. In addition, $\hat{\Sigma}_v^{(b)}$ and $\hat{\sigma}_e^{2(b)}$ were calculated based on the REML method. We also computed the population mean of y_{ij} for each area i as

$$\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i,\ i = 1, ..., 140,$$

where N_i is the population size in i^{th} area. Further, for each sample $b = 1, ..., B$, we calculated the new GREG estimate of the i^{th} area mean as $\hat{\bar{Y}}_i^{\text{LV}(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i^{(b)})'(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})$, the GREG estimate as $\hat{\bar{Y}}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i^{(b)})' Z_i \hat{\beta}_{\text{ols}}^{(b)}$ and the EBLUP estimate as $\hat{\bar{Y}}_i^{B(b)} = f_i\bar{y}_i + (1 - f_i)[\bar{X}_i^{*'}(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})]$, where

$$f_i = n_i/N_i,\ \bar{X}_i^* = \frac{N_i\bar{X}_i - n_i\bar{x}_i}{N_i - n_i},$$

and $\hat{v}_i^{(b)} = \hat{\Sigma}_v^{(b)} X_i' \hat{V}_i^{-1(b)}(y_i^{(b)} - X_i Z_i \hat{\beta}^{(b)})$.

The EBLUP estimator accounts for the finite population corrections f_i.

We computed the average mean squared error ($\overline{\text{MSE}}_2$) and average absolute relative error ($\overline{\text{ARE}}_2$) as

$$\overline{\text{MSE}}_2 = \frac{1}{m}\sum_i^m \text{MSE}_{2i},\quad \overline{\text{ARE}}_2 = \frac{1}{m}\sum_i^m \text{ARE}_{2i},$$

where

$$\text{MSE}_{2i} = \frac{1}{B}\sum_{b=1}^B (\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i)^2,\quad \text{ARE}_{2i} = \frac{1}{B}\sum_{b=1}^B |\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i| / \bar{Y}_i,$$

and $\hat{\bar{Y}}_i^{(b)}$ denotes $\hat{\bar{Y}}_i^{\text{LV}(b)}$, $\hat{\bar{Y}}_i^{G(b)}$ or $\hat{\bar{Y}}_i^{B(b)}$. We report the results in Tables 5 and 6 for $n_i = 5$ and $n_i = 10$ respectively. Both models with area level covariate z and without z are considered.

Table 5
Comparison of small area estimators: average MSE ($\overline{\text{MSE}}_2$) and average absolute relative error ($\overline{\text{ARE}}_2$) under a finite population framework ($n_i = 5$)

Quality Measure	Model without z			Model with z		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{\text{MSE}}_2$	11.03	10.02	6.50	10.76	10.06	7.06
$\overline{\text{ARE}}_2$	0.27	0.24	0.18	0.25	0.23	0.22

Table 5 shows that for $n_i = 5$ the new GREG estimator exhibits slightly better performance relative to the GREG estimator in the sense of smaller $\overline{\text{MSE}}_2$ and $\overline{\text{ARE}}_2$. On the other hand, Table 6 reveals that with $n_i = 10$ the GREG estimator has slightly better performance than the new GREG estimator in terms of $\overline{\text{MSE}}_2$ but not $\overline{\text{ARE}}_2$. However, the EBLUP estimator gives substantially smaller $\overline{\text{MSE}}_2$ and $\overline{\text{ARE}}_2$ than the GREG and the new GREG in both cases due to borrowing strength from related small areas. For example, for the model without z and $n_i = 5$, $\overline{\text{MSE}}_2 = 10.02$, 11.03 and 6.50 for the new GREG, the GREG and the EBLUP, respectively.

Table 6
Comparison of small area estimators: average MSE ($\overline{\text{MSE}}_2$) and average absolute relative error ($\overline{\text{ARE}}_2$) under finite population framework (n_i = 10)

Quality Measure	Model without z			Model with z		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{\text{MSE}}_2$	6.53	6.77	4.73	6.75	6.96	5.24
$\overline{\text{ARE}}_2$	0.20	0.18	0.15	0.19	0.18	0.19

5. Summary

In this paper, we derived the model mean squared error (MSE) of a two-level model-assisted new GREG estimator of a small area mean, proposed by Lehtonen and Veijanen (1999). In addition, we used a data set of Moura and Holt (1999) to demonstrate empirically that the BLUP estimator is substantially more efficient than the new GREG estimator in terms of model MSE, while the new GREG is only slightly more efficient than the customary GREG based on the regression model $y_i = X_i Z_i \beta + e_i, i = 1, ..., m.$ Moreover, using a simulation study under a model-based framework, we have shown that the new GREG estimator has consistently better performance relative to the GREG estimator in terms of average MSE, $\overline{\text{MSE}}$, and average absolute relative error, $\overline{\text{ARE}}$. However, due to borrowing strength from related small areas, EBLUP estimator exhibits significantly better performance relative to the GREG and the new GREG estimators. In addition, we conducted a simulation study under a finite population framework and showed that the EBLUP estimator outperforms the new GREG and the GREG estimators in terms of $\overline{\text{MSE}}$ and $\overline{\text{ARE}}$.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. It is based on a chapter in M.Torabi's Ph.D. thesis written under the supervision of J.N.K. Rao. The authors thank the referees and an associate editor for their helpful comments on the original version of this paper.

Appendix

Derivation of MSE($\tilde{\mu}_i^{\text{LV}}$):

$$\begin{aligned}\text{MSE}(\tilde{\mu}_i^{\text{LV}}) &= E(\tilde{\mu}_i^{\text{LV}} - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{\nu}_i - \nu_i)]^2 + E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 \\ &\quad + 2E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\bar{X}_i'(\tilde{\nu}_i - \nu_i)], \qquad \text{(A.1)}\end{aligned}$$

where the first term on the right hand side of (A.1) is the MSE of the BLUP estimator under the two-level model, given by (9). Moreover, we may write

$$\begin{aligned}E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 &= E[\bar{y}_i - \bar{x}_i'Z_i\beta]^2 \\ &\quad + E(\bar{x}_i'\tilde{\nu}_i)^2 - 2E[(\bar{y}_i - \bar{x}_i'Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)], \qquad \text{(A.2)}\end{aligned}$$

where

$$E[\bar{y}_i - \bar{x}_i'Z_i\beta]^2 = \text{Var}(\bar{y}_i) = \bar{x}_i'\Sigma_v\bar{x}_i + \frac{\sigma_e^2}{n_i},$$

and

$$\begin{aligned}E(\bar{x}_i'\tilde{\nu}_i)^2 &= \text{Var}(\bar{x}_i'\tilde{\nu}_i) + [E(\bar{x}_i'\tilde{\nu}_i)]^2 \\ &= \text{Var}\left[\bar{x}_i'\Sigma_v X_i'V_i^{-1}(\text{y}_i - X_iZ_i\beta)\right] \\ &\quad + \left[E\left(\bar{x}_i'\Sigma_v X_i'V_i^{-1}(\text{y}_i - X_iZ_i\beta)\right)\right]^2 \\ &= \bar{x}_i'\Sigma_v X_i'V_i^{-1}X_i\Sigma_v\bar{x}_i.\end{aligned}$$

In addition,

$$\begin{aligned}E[(\bar{y}_i - \bar{x}_i'Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)] &= E\left[\bar{y}_i\bar{x}_i'\Sigma_v X_i'V_i^{-1}(\text{y}_i - X_iZ_i\beta)\right] \\ &\quad - E\left[\bar{x}_i'Z_i\beta\bar{x}_i'\Sigma_v X_i'V_i^{-1}(\text{y}_i - X_iZ_i\beta)\right],\end{aligned}$$

where the second term is zero. Therefore, we may write

$$\begin{aligned}E[(\bar{y}_i - \bar{x}_i'Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)] &= E\left[\bar{y}_i\bar{x}_i'\Sigma_v X_i'V_i^{-1}\text{y}_i\right] \\ &\quad - \bar{x}_i'Z_i\beta\bar{x}_i'\Sigma_v X_i'V_i^{-1}X_iZ_i\beta,\end{aligned}$$

where the first term can be written as

$$\begin{aligned}&E\left[\frac{1'}{n_i}\text{y}_i\bar{x}_i'\Sigma_v X_i'V_i^{-1}\text{y}_i\right] = \\ &\qquad \frac{1'}{n_i}\left(X_i\Sigma_v\bar{x}_i + X_iZ_i\beta\bar{x}_i'\Sigma_v X_i'V_i^{-1}X_iZ_i\beta\right),\end{aligned}$$

using the following Lemma:

LEMMA 1 (Searle 1971). *If y is a $n \times 1$ vector with mean μ and variance-covariance matrix Σ and b is a $n \times 1$ vector, then* $E(yb'y) = \Sigma b + \mu b'\mu$.

Hence,

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i' Z_i \beta)(\bar{x}_i' \tilde{\nu}_i)] &= \bar{x}_i' \Sigma_v \bar{x}_i + \bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ &\quad - \bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ &= \bar{x}_i' \Sigma_v \bar{x}_i. \end{aligned}$$

Then we may write (A.2) as follows:

$$\begin{aligned} E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 &= \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \\ &\quad + \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - 2\bar{x}_i' \Sigma_v \bar{x}_i \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i}. \end{aligned} \quad \text{(A.3)}$$

Finally, we need to find the cross-product term of (A.1). We have

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] \\ &\quad - E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)], \end{aligned} \quad \text{(A.4)}$$

where the first term on the right side of (A.4) may be written as

$$\begin{aligned} E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E\left[\bar{y}_i \bar{X}_i'\left(\Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta) - \nu_i\right)\right] \\ &= E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} y_i\right] \\ &\quad - E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right] \\ &\quad - E[\bar{y}_i \bar{X}_i' \nu_i]. \end{aligned} \quad \text{(A.5)}$$

The first term of (A.5) is

$$\begin{aligned} E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} y_i\right] &= E\left[\frac{1'}{n_i} y_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} y_i\right] \\ &= \frac{1'}{n_i} X_i\left(\Sigma_v \bar{X}_i + Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right), \end{aligned} \quad \text{(A.6)}$$

the second term of (A.5) can be written as

$$\begin{aligned} E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right] &= \bar{x}_i' Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta, \end{aligned} \quad \text{(A.7)}$$

and the third term can be expressed as

$$\begin{aligned} E[\bar{y}_i \bar{X}_i' \nu_i] &= E[(\bar{x}_i' Z_i \beta + \bar{x}_i' \nu_i + e_i)\bar{X}_i' \nu_i] \\ &= E[\bar{x}_i' \nu_i \bar{X}_i' \nu_i] = \bar{x}_i' \Sigma_v \bar{X}_i. \end{aligned} \quad \text{(A.8)}$$

Therefore, substituting (A.6), (A.7) and (A.8) in (A.5), we have

$$E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.9)}$$

We now turn to the second term of (A.4). We have

$$\begin{aligned} E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E[\bar{x}_i' Z_i \beta \bar{X}_i' \tilde{\nu}_i] + E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \tilde{\nu}_i] \\ &\quad - E[\bar{x}_i' Z_i \beta \bar{X}_i' \nu_i] - E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \nu_i]. \end{aligned} \quad \text{(A.10)}$$

Then we obtain the following expression for the four terms on the right side of (A.10):

$$\begin{aligned} E[\bar{x}_i' Z_i \beta \bar{X}_i' \tilde{\nu}_i] &= E\left[\bar{x}_i' Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta)\right] = 0, \end{aligned} \quad \text{(A.11)}$$

$$\begin{aligned} E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \tilde{\nu}_i] &= \bar{x}_i' \Sigma_{\tilde{\nu}} \bar{X}_i + \bar{x}_i' E(\tilde{\nu}_i)\bar{X}_i' E(\tilde{\nu}_i) \\ &= \bar{x}_i' \mathrm{Var}\left[\Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta)\right]\bar{X}_i \\ &\quad + \bar{x}_i' E\left(\Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta)\right) \\ &\quad \times \bar{X}_i' E\left(\Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta)\right) \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i, \end{aligned} \quad \text{(A.12)}$$

$$E[\bar{x}_i' Z_i \beta \bar{X}_i' \nu_i] = 0 \quad \text{(A.13)}$$

and

$$\begin{aligned} E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \nu_i] &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta)\bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} y_i \bar{X}_i' \nu_i\right] \\ &\quad - E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1}(X_i Z_i \beta + X_i \nu_i + e_i)\bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \nu_i \bar{X}_i' \nu_i\right] \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i. \end{aligned} \quad \text{(A.14)}$$

Therefore, substituting (A.11)-(A.14) in (A.10), we get

$$E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.15)}$$

Hence, it follows from (A.4), (A.9) and (A.15) that

$$E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.16)}$$

It now follows from (A.1), (A.3) and (A.16) that

$$\mathrm{MSE}(\tilde{\mu}_i^{\mathrm{LV}}) = \mathrm{MSE}(\tilde{\mu}_i^{B}) + \left\{ \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \right\},$$

as stated in Theorem 3.

References

Lehtonen, R., and Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *IASS Satellite Conference on Small Area Estimation*, Riga: Latvian Council of Science, 121- 128.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44.

Moura, F.A.S., and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, 25, 73-80.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken: New York: John Wiley & Sons, Inc.

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

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 19-27
Statistics Canada, Catalogue No. 12-001-X

An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada

Yong You [1]

Abstract

The Canadian Labour Force Survey (LFS) produces monthly estimates of the unemployment rate at national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Urban Centers (UCs). However, for some sub-provincial areas, the direct estimates are not reliable since the sample size in some areas is quite small. The small area estimation in LFS concerns estimation of unemployment rates for local sub-provincial areas such as CMA/UCs using small area models. In this paper, we will discuss various models including the Fay-Herriot model and cross-sectional and time series models. In particular, an integrated non-linear mixed effects model will be proposed under the hierarchical Bayes (HB) framework for the LFS unemployment rate estimation. Monthly Employment Insurance (EI) beneficiary data at the CMA/UC level are used as auxiliary covariates in the model. A HB approach with the Gibbs sampling method is used to obtain the estimates of posterior means and posterior variances of the CMA/UC level unemployment rates. The proposed HB model leads to reliable model-based estimates in terms of CV reduction. Model fit analysis and comparison of the model-based estimates with the direct estimates are presented in the paper.

Key Words: Design effect; Hierarchical Bayes; Log-linear mixed effects model; Model checking; Sampling variance; Small area.

1. Introduction

The unemployment rate is generally viewed as a key indicator of economic performance. In Canada, the unemployment rate estimates are produced monthly by the Labour Force Survey (LFS) of Statistics Canada. The LFS is a monthly survey of 53,000 households selected using a stratified, multistage design. Each month, one-sixth of the sample is replaced. Thus five-sixths of the sample is common between two consecutive months. This sample overlap induces correlations which can be exploited to produce better estimates by alternative methods such as model-based methods to borrow strength over time; more details will be discussed in Section 2. For a detailed description of the LFS design, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998). The LFS releases monthly unemployment rate estimates for large areas such as the nation and provinces as well as local areas (small areas) such as Census Metropolitan Areas (CMAs, *i.e.*, cities with population more than 100,000) and other Urban Centres (UCs) across Canada. Although national and provincial estimates get the most media attention, sub-provincial estimates of the unemployment rate are also very important. They are used by the Employment Insurance (EI) program to determine the rules used to administer the program. In addition, the unemployment rates for CMAs and UCs receive close scrutiny at local levels. However, many local areas do not have large enough samples to produce adequate direct estimates, since the LFS is designed to produce adequate or reliable estimates at the national level and provincial level. The estimated coefficient of variation (CV) level for the nation is about 2% and 4% to 7% for provinces. However, the CVs for CMAs and UCs range from about 7% to 50%. Some UCs have CVs even larger than 50%. The direct LFS estimates for some local areas are not reliable with very large CVs due to the small sample sizes for those areas. Therefore, alternative estimators, in particular, model-based estimators, are considered to improve the direct LFS estimates for small areas. The objective in this paper is to obtain a reliable model-based estimator that is an improvement over the direct LFS estimator in terms of small and stable CVs.

In general, direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains such as the nation and provinces. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide reliable direct estimates for specific small domains. In making estimates for small areas, it is necessary to borrow strength from related areas to form indirect estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as census counts and administrative records. It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data; see Rao (2003). The model-based estimators are indirect estimators in the sense that these estimators are obtained by using small

1. Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: yong.you@statcan.ca.

area models, direct estimates and related auxiliary variables. The model-based estimators are obtained to improve the direct design-based estimators in terms of precision and reliability, that is, smaller CVs. Traditional small area estimators borrow strength either from similar small areas or from the same area across time, but not both. In recent years, several approaches to borrowing strength simultaneously across both space and time have been developed. Estimators based on the approach developed by Rao and Yu (1994), Ghosh, Nangia and Kim (1996), Datta, Lahiri, Maiti and Lu (1999), Datta, Lahiri and Maiti (2002) and You, Rao and Gambino (2000, 2003), successfully exploit the two dimensions simultaneously to produce improved estimates with desirable properties for small areas. In particular, You *et al.* (2000, 2003) studied the model-based estimation of unemployment rates for local sub-provincial areas such as CMAs and Census Agglomerations (CAs) across Canada. They obtained efficient model-based estimators and adequate model fit for the LFS unemployment rate estimation. However, the model proposed by You *et al.* (2000, 2003) has some limitations. In this paper, we discuss these limitations and propose a new integrated model for the LFS unemployment rate estimation under hierarchical Bayes (HB) framework. The idea is to model the parameters of interest and the sampling variances together as suggested in You *et al.* (2003) and You and Chapman (2006). We will apply the proposed model to the 2005 LFS data and obtain the model-based unemployment rate estimates. Comparison of the HB estimates with the direct LFS estimates and model fit analysis will also be provided.

This paper is organized as follows. In Section 2, we present and discuss various small area models proposed in the literature for the unemployment rate estimation. In Section 3, we discuss the problem of smoothing and modeling the sampling covariance matrix. In Section 4, an integrated non-linear mixed effects model is proposed in a hierarchical Bayes framework, and the use of Gibbs sampling to generate samples from the joint posterior distribution is described. In Sections 5, we apply the proposed model to LFS data and obtain the HB estimates for small area unemployment rates. Model analysis and evaluation are also provided. And finally in Section 6 we offer some concluding remarks and future work directions.

2. Small area models

2.1 Cross-sectional model

Cross-sectional or area level models are used to produce reliable model-based estimates by combining area level auxiliary information and direct area level estimates. A basic area level model is the well-known Fay-Herriot model (Fay and Herriot 1979). This model has two components: (1) a sampling model for the direct survey estimates, and (2) a linking model that relates the small area parameters to area level auxiliary variables through a linear regression model. For the LFS monthly unemployment rate estimation, let θ_{it} denote the true unemployment rate for the i^{th} CMA/UC at a particular time (month) t, where $i = 1, ..., m$, where m is the number of CMA/UCs, and let y_{it} denote the direct LFS estimate of θ_{it}. Then the sampling model for y_{it} can be expressed as

$$y_{it} = \theta_{it} + e_{it}, \quad i = 1, ..., m, \tag{1}$$

where e_{it} is the sampling error associated with the direct estimator y_{it}. The sampling error is assumed to be normally distributed as $e_{it} \sim N(0, \sigma_{it}^2)$ where σ_{it}^2 is the sampling variance. The linking model for the true unemployment rate θ_{it} may be written as

$$\theta_{it} = x_{it}'\beta + v_i, \quad i = 1, ..., m, \tag{2}$$

where x_{it} is the auxiliary variable and v_i is area-specific random effect. For each time point (each month), we can use the Fay-Herriot model for the monthly direct estimates. The Fay-Herriot model combines cross-sectional information but does not borrow strength over the past time periods.

2.2 Cross-sectional and time series model

Because of the LFS sample design and rotation pattern, there is substantial sample overlap over six month time periods within each area. As a result, for a particular area i, the correlation between the sampling errors e_{it} and e_{is} $(t \neq s)$ need to be taken into account. You *et al.* (2000, 2003) proposed a cross-sectional and time series model for the LFS unemployment rate estimates. You *et al.* (2000, 2003) only used previous six months of data to predict the current month rate since the LFS sample rotation is based on a six month cycle. Each month, one sixth of the LFS sample is replaced. Thus after six months, the correlation between estimates is weak (see Section 2.1 for the lag correlation coefficients). Let $y_i = (y_{i1}, ..., y_{iT})'$, $\theta_i = (\theta_{i1}, ..., \theta_{iT})'$, and $e_i = (e_{i1}, ..., e_{iT})'$, where $T = 6$ here. By assuming that e_i follows a multivariate normal distribution with mean vector 0 and sampling covariance matrix $\sum_i$, we have

$$y_i \sim N\left(\theta_i, \sum\nolimits_i\right), \quad i = 1, ..., m.$$

Thus y_i is assumed to be design-unbiased for θ_i. The sampling covariance matrix $\sum_i$ is unknown in the model. Direct estimates of the sampling covariance matrices are available. It is customary to assume a known sampling variance in area level model-based small area estimation (Rao 2003). For example, the traditional Fay-Herriot model assumes the sampling variance known in the model. Usually

a smoothed estimator of the sampling variance is used. However, recent development on modeling the sampling variance provides an alternative approach to handle the problem of sampling variance; for example, see Wang and Fuller (2003), You and Dick (2004) and You and Chapman (2006). For the unemployment rate estimation, details on smoothing and modeling the sampling variances are given in Section 3.

To borrow strength across regions and time periods, and following You *et al.* (2000, 2003) we can model the true unemployment rate θ_{it} by a linear regression model with random effects through auxiliary variables x_{it}, that is,

$$\theta_{it} = x_{it}'\beta + v_i + u_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T, \qquad (3)$$

where v_i is a area random effect assumed to be $N(0, \sigma_v^2)$ and u_{it} is a random time and area component. We can further assume that u_{it} follows a random walk process over time period $t = 1, \ldots, T$, that is,

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \qquad (4)$$

where $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Then $\text{cov}(u_{it}, u_{is}) = \min(t, s)\, \sigma_\varepsilon^2$. The regression vector β and the variance components σ_v^2 and σ_ε^2 are unknown in the model and need to be estimated. Combining the model (1), (3) and (4), we obtain a linear mixed model with time components as

$$y_{it} = x_{it}'\beta + v_i + u_{it} + e_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T. \qquad (5)$$

You *et al.* (2003) showed that the cross-sectional and time series model (5) is better than the Fay-Herriot model in terms of smoothing the direct estimates and CV reduction over the direct estimates for the LFS unemployment rate estimation.

We have used a random walk model for u_{it}. Rao and Yu (1994) used a stationary autoregressive model for u_{it}. You *et al.* (2003) showed that the random walk model on u_{it} had provided better model fit to the unemployment rate estimation than the autoregressive AR(1) model. Datta *et al.* (1999) also used a random walk model to estimate the US unemployment rates at the state level.

2.3 Log-linear linking model

However, a limitation of the model (3) is that the linking model for the parameter of interest, the true unemployment rate θ_{it}, is a linear model with normal random effects. Since θ_{it} is a positive number between 0 and 1, and it is close to 0, the linear linking model with normal random effects may lead to negative estimates for θ_{it} for some small areas. To avoid this problem, You, Chen and Gambino (2002) proposed a log-linear linking model for θ_{it} as follows:

$$\log(\theta_{it}) = x_{it}'\beta + v_i + u_{it}, \quad i = 1, \ldots, m, t = 1, \ldots, T. \qquad (6)$$

You and Rao (2002) also studied the log-linear linking model for the Fay-Herriot model as the unmatched sampling and linking models with application in the Canadian census undercoverage estimation. The results of You and Rao (2002) and You *et al.* (2002) have shown that the log-linear linking model performs very well in the small area estimation problems. In this paper, we therefore will use the log-linear linking model (6) for the true unemployment rate θ_{it}.

3. Sampling variance

In general, we can obtain direct sampling variance estimates from survey data. However, these direct estimates are unstable if sample sizes are small. In area level models of small area estimation, the sampling variances are usually assumed to be known (*e.g.*, Fay and Herriot 1979; Datta *et al.* 1999; You and Rao 2002). If the sampling variances are assumed to be known in the model, then reliable (smoothed) estimates of sampling variances are constructed using other auxiliary data and models usually through generalized variance functions (*e.g.*, Dick 1995; Datta *et al.* 1999). In this paper alternatively, we model sampling variance covariance matrix using the direct estimates in a specific way such that we do not need to assume the sampling variances and covariances are known in the model. Thus we simplify the problem of smoothing unknown sampling variance and integrate the sampling variance modeling part into the whole model.

3.1 Smoothing sampling covariance matrix

You *et al.* (2000, 2003) used two steps to smooth the sampling covariance matrix. The first step is to obtain a smoothed or common CV for each CMA/UC by computing the average CVs for each CMA/UC over a certain time period, denoted as $\overline{\text{CV}}_i$, where $i = 1, 2, \ldots, m$. The second step is to obtain the average lag correlation coefficients over time and all CMA/UCs, denoted as $\bar{\rho}_{|t-s|}$ for the time lag $|t - s|$. This step involves intensive computation. We have used three years (1999 to 2001) of LFS data to compute the smoothed correlation coefficients. We treat the smoothed values over both time and space as the true values in the model. The one-month lag (lag-1) correlation coefficient is obtained as $\bar{\rho}_1 = 0.48$, lag-2 correlation coefficient is $\bar{\rho}_2 = 0.31$, lag-3 is $\bar{\rho}_3 = 0.21$, lag-4 is $\bar{\rho}_4 = 0.16$, lag-5 is $\bar{\rho}_5 = 0.11$ and $\bar{\rho}_6 = 0.1$. After lag 6, the lag correlation coefficient is less than 0.1. The lag correlation coefficients decrease as the lag increases. This is consistent with the rotation pattern of the LFS design. Figure 1 shows the smoothed lag correlation coefficients for the LFS unemployment rate estimates.

By using these smoothed CVs and lag correlation coefficients, a smoothed covariance matrix $\hat{\Sigma}_i$ can be obtained with diagonal elements $\hat{\sigma}_{it}^2 = (\overline{\mathrm{CV}}_i)^2 y_{it}^{2\,i}$ and off-diagonal elements $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|}\hat{\sigma}_{it}\hat{\sigma}_{is}$. The smoothed $\hat{\Sigma}_i$ is then treated as known in the model. The study of You *et al.* (2000, 2003) suggests that using the smoothed $\hat{\Sigma}_i$ in the model can significantly improve the estimates in terms of CV reduction compared to the HB estimates obtained using the direct survey estimates of Σ_i in the model. For more details of the result, see You *et al.* (2003).

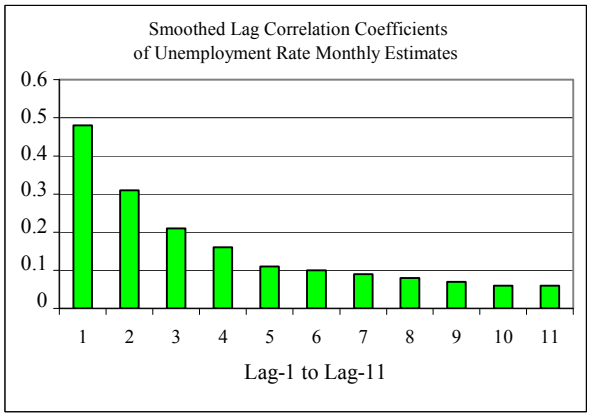

Figure 1 Smoothed unemployment rate lag correlation coefficients

3.2 Equal CV modeling approach

The main problem of the method of You *et al.* (2000, 2003) is that the smoothed sampling covariance matrices depend on the direct survey estimates y_{it}, whereas the y_{it}'s are not reliable for some small regions. Note that the true sampling variance can be written as $\sigma_{it}^2 = \theta_{it}^2(\mathrm{CV}_{it})^2$. Based on the assumption of common CV over time for a given area, You *et al.* (2003) suggested in their concluding remarks to use estimates of the form $\tilde{\sigma}_{it}^2 = \theta_{it}^2(\overline{\mathrm{CV}}_i)^2$ and $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$ for the smoothed variances and covariances respectively. Then the new smoothed sampling covariance matrix $\tilde{\Sigma}_i$ has diagonal elements $\tilde{\sigma}_{it}^2$ and off-diagonal elements $\tilde{\sigma}_{its}$. However, under this method, the sampling covariance matrix $\tilde{\Sigma}_i$ becomes unknown in the model, since $\tilde{\sigma}_{it}^2$ and $\tilde{\sigma}_{its}$ depend on the unknown parameters θ_{it}, whereas θ_{it} is related to a linking model. The advantage of this method is that the model structure of the sampling covariance matrix is clearly specified. This method is better than the smoothing method in the sense that the sampling covariance is clearly specified and not treated as known.

3.3 Equal design effects modeling approach

An alternative modeling approach is based on the assumption of common design effects as suggested in Singh, You and Mantel (2005) and Singh, Folsom and Vaish (2005) to smooth the sampling variance σ_{it}^2. The design effect (deff) for the ith area at time t may be approximately written as

$$\mathrm{deff}_{it} = \frac{\sigma_{it}^2}{\theta_{it}(1-\theta_{it})/n_{it}},$$

where n_{it} is the corresponding sample size. Then the sampling variance σ_{it}^2 can be written as $\sigma_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\mathrm{deff}_{it}/n_{it}$. Let $\tau_{it} = \mathrm{deff}_{it}/n_{it} = \sigma_{it}^2/(\theta_{it}(1-\theta_{it}))$. Then we can estimate τ_{it} using the direct estimates of θ_{it} and σ_{it}^2 as $\hat{\tau}_{it} = \hat{\sigma}_{it}^2/(y_{it}(1-y_{it}))$. For each area, based on the assumption of a common deff and a common sample size over time, we can obtain a smoothed average factor $\bar{\tau}_i$ as $\bar{\tau}_i = \sum_{t=1}^{T}\hat{\tau}_{it}/T$. Then a smoothed sampling variance can be obtained as $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\bar{\tau}_i$, which again depends on θ_{it} as well. The sampling covariance is still in the form of $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$, as in You *et al.* (2003). Note that $\bar{\tau}_i$ is a moving average of $\hat{\tau}_{it}$ over the time period T in the model. In practice, however, alternatively one may use a longer time series data to obtain more stable estimate of $\bar{\tau}_i$ for each area if necessary. In this paper, we will use the common design effects model for unemployment rate estimation based on the smoothed moving average factor $\bar{\tau}_i$ as we borrow information from the past time period T.

4. Hierarchical Bayes inference

In this section, we propose an integrated cross-sectional and time series log-linear model for the unemployment rate estimation. We apply the hierarchical Bayes approach to the model. Estimates of posterior means and posterior variances are obtained by using the Gibbs sampling method.

4.1 Integrated hierarchical Bayes model

We now propose the integrated cross-sectional and time series log-linear model in a hierarchical Bayes framework as follows:

- Conditional on $\theta_i = (\theta_{i1}, \ldots, \theta_{iT})'$, $[y_i \mid \theta_i] \sim \text{ind } N(\theta_i, \sum_i(\theta_i))$;
- Conditional on β, u_{it} and σ_v^2, $[\log(\theta_{it}) \mid \beta, u_{it}, \sigma_v^2] \sim \text{ind } N(x_{it}'\beta + u_{it}, \sigma_v^2)$;
- Conditional on $u_{i,t-1}$ and σ_ε^2, $[u_{it} \mid u_{i,t-1}, \sigma_\varepsilon^2] \sim \text{ind } N(u_{i,t-1}, \sigma_\varepsilon^2)$;
- $\sum_i(\theta_i)$ depends on θ_i with diagonal elements $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\bar{\tau}_i$ and off-diagonal elements $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$.
- Marginally β, σ_v^2 and σ_ε^2 are mutually independent with priors given as $\beta \propto 1$, $\sigma_v^2 \sim \mathrm{IG}(a_1, b_1)$, and $\sigma_\varepsilon^2 \sim \mathrm{IG}(a_2, b_2)$, where IG denotes an inverse gamma

distribution and a_1, b_1, a_2, b_2 are known positive constants and usually set to be very small to reflect our vague knowledge about σ_v^2 and σ_ε^2.

Remarks:

1. The proposed HB model has used a log-linear linking model for the small area parameter of interest θ_{it} as suggested in You *et al.* (2002) and You and Rao (2002).
2. The sampling covariance matrix $\sum_i$ is unknown in the model, and it is specified as a function of unknown small area parameter θ_i as suggested in You and Rao (2002) and You *et al.* (2003).
3. We have used the assumption of common design effects for small areas as suggested in Singh, You and Mantel (2005).
4. The proposed HB model overcomes the limitations of the model of You *et al.* (2000, 2003) in terms of log-linear modeling and specification of unknown sampling covariance matrix modeling. In particular, we model the unknown sampling covariance matrix through the small area parameters θ_i using smoothed estimates of design effects for each areas.

We are interested in estimating the true unemployment rate θ_{it}, and in particular, the current unemployment rate θ_{iT}. In the HB analysis, θ_{iT} is estimated by its posterior mean $E(\theta_{iT} \mid y)$ and the uncertainty associated with the estimator is measured by the posterior variance $V(\theta_{iT} \mid y)$. We use the Gibbs sampling method (Gelfand and Smith 1990; Gelman and Rubin 1992) to obtain the posterior mean and the posterior variance of θ_{iT}.

4.2 Gibbs sampling inference

The Gibbs sampling method is an iterative Markov chain Monte Carlo sampling method to simulate samples from a joint distribution of random variables by sampling from low dimensional densities to make inference about the joint and marginal distributions (Gelfand and Smith 1990). The most prominent application is for inference within a Bayesian framework. In Bayesian inference one is interested in the posterior distribution of the parameters. Assume that $y_i \mid \theta$ has conditional density $f(y_i \mid \theta)$ for $i = 1, ..., n$ and that the prior information about $\theta = (\theta_1, ..., \theta_k)'$ is summarized by a prior density $\pi(\theta)$. Let $\pi(\theta \mid y)$ denote the posterior density of θ given the data $y = (y_1, ..., y_n)'$. It may be difficult to sample from $\pi(\theta \mid y)$ directly in practice due to the high dimensional integration with respect to θ. However, one can use the Gibbs sampler to construct a Markov chain $\{\theta^{(g)} = (\theta_1^{(g)}, ..., \theta_k^{(g)})'\}$ with $\pi(\theta \mid y)$ as the limiting distribution. For illustration, let $\theta = (\theta_1, \theta_2)'$. Starting with an initial set of values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$, we generate $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ by sampling $\theta_1^{(g)}$ from $\pi(\theta_1 \mid \theta_2^{(g-1)}, y)$ and $\theta_2^{(g)}$ from $\pi(\theta_2 \mid \theta_1^{(g-1)}, y)$. Under certain regularity conditions, $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ converges in distribution to $\pi(\theta \mid y)$ as $g \to \infty$. Marginal inference about $\pi(\theta_i \mid y)$ can be based on the marginal samples $\{\theta_i^{(g+k)}; k = 1, 2, ...\}$ for large g.

For the proposed integrated HB model, to obtain the posterior estimation of unemployment rate, we implement the Gibbs sampling method by generating samples from the full conditional distributions of the parameters β, σ_v^2 and σ_ε^2, u_{it} and θ_i. These full conditional distributions are given in the Appendix. The distributions of β, σ_v^2 and σ_ε^2, u_{it} are standard normal or inverse gamma distributions that can be easily sampled. However, the conditional distribution of θ_i does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chib and Greenberg 1995) to update θ_i. Following You *et al.* (2002) and You and Rao (2002), the full conditional distribution of θ_i in the Gibbs sampler can be written as

$$\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto h(\theta_i) f(\theta_i),$$

where

$$h(\theta_i) = \left|\sum_i(\theta_i)\right|^{-1} \exp\{-\frac{1}{2}(y_i - \theta_i)'\sum_i^{-1}(y_i - \theta_i)\}$$

and

$$f(\theta_i) = \exp\left\{-\frac{1}{2\sigma_v^2}(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)'(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)\right\} \cdot \left(\prod_{t=1}^{T}\frac{1}{\theta_{it}}\right).$$

To update θ_i, we proceed as follows:

1. For $t = 1, ..., T$, draw $\theta_{it}^{(k+1)} \sim \log N(x_{it}'\beta^{(k+1)} + u_{it}^{(k+1)}, \sigma_v^{2(k+1)})$, then we have $\theta_i^{(k+1)} = (\theta_{i1}^{(k+1)}, ..., \theta_{iT}^{(k+1)})'$.
2. Compute the rejection probability

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\{\frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1\}.$$

3. Generate $\lambda \sim \text{Uniform}(0,1)$, if $\lambda < \alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$, then accept $\theta_i^{(k+1)}$; otherwise reject $\theta_i^{(k+1)}$ and set $\theta_i^{(k+1)} = \theta_i^{(k)}$.

To implement Gibbs sampling, we follow the recommendation of Gelman and Rubin (1992) and independently run $L (L > 2)$ parallel chains, each of length $2d$. The first d iterations of each chain are deleted. The convergence monitoring is based on the potential scale reduction factor as suggested in Gelman and Rubin (1992) and adopted by You *et al.* (2003) for estimating θ_{iT}. Details are given in You *et al.* (2003). Estimates of the posterior mean $E(\theta_{iT} \mid y)$ and the posterior variance $V(\theta_{iT} \mid y)$ are obtained based on the samples generated from the Gibbs sampler.

5. Application to LFS data

5.1 Estimation

We use the 2005 January to June LFS unemployment rate estimates, y_{it}, in our data analysis. In addition to the direct estimates y_{it} and the sampling covariance matrices used in the small area models, auxiliary administrative variables are needed in the models. For the unemployment rate estimation, local area employment insurance (EI) monthly beneficiary rate is used as auxiliary data x_{it} in the model. The beneficiary rate is calculated as the ratio of the number of persons applying EI benefit over the number of persons in the labour force. There are 72 CMA/UCs across Canada. One UC (Miramichi) does not have the EI data. So we consider $m = 71$ CMA/UCs in the model. Within each area, we consider six consecutive monthly estimates y_{it} from January 2005 to June 2005, so that $T = 6$. For the January to June 2005 data, the overall average (over 71 CMA/UCs and 6 months) unemployment rate is 0.076, and the overall average EI beneficiary rate is 0.059. For the proposed small area model, the parameter of interest θ_{iT} is the true unemployment rate for area i in June 2005, where $i = 1, \ldots, 71$. To implement the Gibbs sampler, we have used 10 parallel runs, each of length 2000. The first 1,000 iterations are deleted as "burn-in" periods. The hyper-parameters for variance components in the model are set to be 0.0001 to reflect the vague knowledge about σ_v^2 and σ_ε^2.

We now present the posterior estimates of the unemployment rates under the proposed integrated HB model given in section 4.1 using the Gibbs sampling method. Figure 2 displays the LFS direct estimates and the HB model-based estimates of the June 2005 unemployment rates for the 71 CMA/UCs across Canada. The 71 CMA/UCs appear in the order of population size with the smallest UC (Dawson Creek, BC) on the left and the largest CMA (Toronto, ON) on the right. For the point estimates, the HB estimates leads to moderate smoothing of the direct LFS estimates. For the CMAs with large population sizes and therefore large sample sizes, the direct estimates and the HB estimates are very close to each other as expected, particularly for Toronto, Montreal and Vancouver; for smaller UCs, the direct and HB estimates differ substantially for some regions.

Figure 3 displays the CVs of the estimates. The CV of the HB estimate is taken as the ratio of the square root of the posterior variance and the posterior mean. It is clear from Figure 3 that the direct estimates have very large CVs, particularly for the UCs, the CVs are very large and unstable. The HB estimates have very small and stable CVs compared to the direct estimates. The efficiency gain of the HB estimates is obvious, particularly for the UCs with smaller population sizes. More precisely, we computed the percent CV reduction for the HB estimators based on the data of June 2005. The percent CV reduction is computed as the difference of the direct CV and HB CV relative to the direct CV. The average CV reduction for UCs is 63% and the CV reduction for CMAs is 35%. As expected, the proposed model has achieved a large CV reduction over the direct estimates, particularly for smaller UCs with smaller sample sizes.

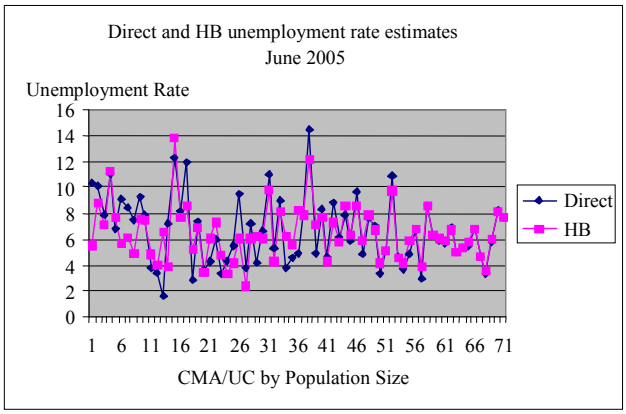

Figure 2 Comparison of direct and HB estimates

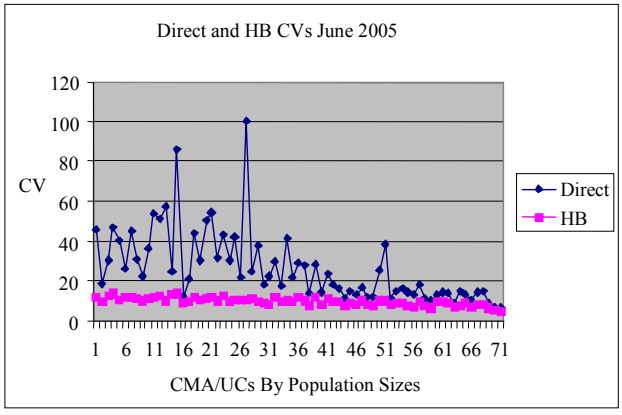

Figure 3 Comparison of direct and HB CVs

5.2 Model fit using posterior predictive distribution

To check the overall fit of the proposed model, we use the method of posterior predictive distribution. Let y_{rep} denote the replicated observation under the model. The posterior predictive distribution of y_{rep} given the observed data y_{obs} is defined as

$$f(y_{\text{rep}} \mid y_{\text{obs}}) = \int f(y_{\text{rep}} \mid \theta) f(\theta \mid y_{\text{obs}}) d\theta.$$

In this approach, a discrepancy measure $D(y, \theta)$ that depends on the data y and the parameter θ can be defined and the observed value $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ compared to the posterior predictive distribution of $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ with any significant difference indicates a model failure. Meng (1994) and Gelman, Carlin, Stern and Rubin (1995) proposed the posterior predictive p-value as

$$p = P(D(y_{\text{rep}}, \theta) \geq D(y_{\text{obs}}, \theta) \mid y_{\text{obs}}).$$

This is a natural extension of the usual p-value in a Bayesian context. If a model fits the observed data, then the two values of the discrepancy measure are similar. In other words, if the given model adequately fits the observed data, then $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ should be near the central part of the histogram of the $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ values if y_{rep} is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive p-value is expected to be near 0.5 if the model adequately fits the data. Extreme p-values (near 0 or 1) suggest poor fit. The posterior predictive p-value can be estimated as follows: Let θ^* represent a draw from the posterior distribution $f(\theta \mid y_{\text{obs}})$, and let y^*_{rep} represent a draw from $f(y_{\text{rep}} \mid \theta^*)$. Then marginally y^*_{rep} is a sample from the posterior predictive distribution $f(y_{\text{rep}} \mid y_{\text{obs}})$. Computing the p-value is relatively easy using the simulated values of θ^* from the Gibbs sampler. For each simulated value θ^*, we can simulate y^*_{rep} from the model and compute $D(y^*_{\text{rep}}, \theta^*)$ and $D(y_{\text{obs}}, \theta^*)$. Then the p-value is estimated by the proportion of times $D(y^*_{\text{rep}}, \theta^*)$ exceeds $D(y_{\text{obs}}, \theta^*)$.

For the proposed HB model, the discrepancy measure used for overall fit is given by $d(y, \theta) = \sum_{i=1}^{m} (y_i - \theta_i)' \sum_i^{-1} (y_i - \theta_i)$. This measure has been used by Datta *et al.* (1999) and You *et al.* (2003). We computed the p-value by combining the simulated θ^* and y^* from all 10 parallel runs. We obtained an estimated average p-value about 0.38. Thus we have no indication of lack of overall model fit.

The posterior predictive p-value model checking has been criticized for being conservative due to the double use of the observed data. The double use of the data can induce unnatural behaviour, as demonstrated by Bayarri and Berger (2000). They proposed alternative model checking p-value measures, named the partial posterior predictive p-value and the conditional predictive p-value. However, their methods are more difficult to implement and interpret (Rao 2002; Sinharay and Stern 2003). As noted in Sinharay and Stern (2003), the posterior predictive p-value is especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

To compare the proposed model with the model of You *et al.* (2003), we computed the divergence measure of Laud and Ibrahim (1995) based on the posterior predictive distribution. The expected divergence measure of Laud and Ibrahim (1995) is given by $d(y^*, y_{\text{obs}}) = E(k^{-1} \| y^* - y_{\text{obs}} \|^2 \mid y_{\text{obs}})$, where k is the dimension of y_{obs} and y^* is a sample from the posterior predictive distribution $f(y \mid y_{\text{obs}})$. Between two models, we prefer a model that yields a smaller value of this measure. As in Datta, Day and Maiti (1998) and You *et al.* (2003), we approximated the divergence measure $d(y^*, y_{\text{obs}})$ by using the simulated samples from the posterior predictive distribution. Using the Gibbs sampling multiple outputs, we obtained a divergence measure in the range of 8 to 9 for the proposed model, and about 12 to 14 for the model of You *et al.* (2003). Thus the divergence measure suggests a better fit of the proposed integrated HB model for the LFS unemployment rate estimation.

5.3 Bias diagnostic using regression analysis

To evaluate the possible bias introduced by the model, we use a simple method of ordinary least squares regression analysis for the direct LFS estimates and the HB model-based estimates. The regression method is suggested by Brown, Chamber, Heady and Heasman (2001). If the model-based estimates are close to the true unemployment rates, then the direct LFS estimators should behave like random variables whose expected values correspond to the values of the model-based estimates. We plot the model-based HB estimates as X and the direct LFS estimates as Y, and see how close the regression line is to $Y = X$. In terms of regression, basically we fit the regression model $Y = \alpha X$ to the data and estimate the coefficient α. Less biased model-based estimates should lead to the value of α close to 1. For the June 2005 data, let Y be the direct unemployment rate estimates, and X be the model-based HB estimates. We obtain the estimated α value as 1.0207 with standard error 0.0281. Figure 4 shows a scatter plot with the fitted regression line.

The regression result shows no significant difference from $Y = X$. Therefore, we conclude that the model-based estimates derived from the proposed model are consistent with the direct LFS estimates with no extra possible bias included. The result may also indicate no evidence of any bias due to possible model misspecification.

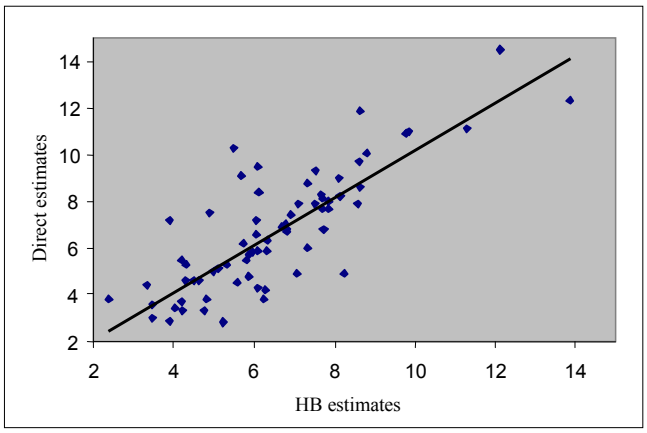

Figure 4 Scatter plot with regression line

6. Concluding remarks and future work

In this paper we have reviewed some small area models including the Fay-Herriot model and the cross-sectional and time series model of You *et al.* (2003). In view of the previous work, we have proposed an integrated non-linear cross-sectional and time series model to obtain model-based estimates of unemployment rates for CMA/UCs across Canada using the LFS data. The proposed model overcomes the limitations of the previous work. In particular, we can model the sampling variance as a function of the small area mean by assuming either a common CV for a given area or a common deff for a given area. Our data analysis has shown that the proposed model fits the data quite well. The hierarchical Bayes estimates, based on the model, improve the direct survey estimates significantly in terms of CV reduction, especially for UCs with small population sizes.

We plan to use alternative modeling approach for the sampling variance. Recently You and Dick (2004) and You and Chapman (2006) has used the HB approach to model the sampling variance directly without specifying the form of the sampling variance under the frame of the Fay-Herriot model. The model automatically takes into account the variability of estimating the sampling variances. In particular, You and Dick (2004) applied the model to the census undercoverage estimation problem and obtained efficient HB census undercoverage estimates for small domains across Canada. It will be interesting to adopt the same idea to the cross-sectional and time series model and compare the results with the current work. The purpose of comparison is to establish a reliable and easy-to-implement model for the LFS model-based unemployment rate estimation for small areas.

We plan to produce the model-based estimates for a relative long time period, for example, 24 months from 2004 to 2005. We will compare the 24 months model-based estimates with the 24 months direct estimates, particularly for the large CMAs to study the smoothing effects of the proposed model. The model-based estimates should follow the pattern of direct LFS estimates for large CMAs, which indicates that the smoothing effects on time series effects are reasonable. The purpose is to verify the robustness of the proposed model-based estimates over time.

Appendix

In the following, we present the full conditional distributions for the Gibbs sampler under the proposed HB model. Let $Y = (Y_1', ..., Y_m')'$, $X = (X_1', ..., X_m')'$, $\theta = (\theta_1', ..., \theta_m')'$, and $u = (u_1', ..., u_m')'$, with $Y_i' = (y_{i1}, ..., y_{iT})$, $X_i' = (x_{i1}, ..., x_{iT})$, $\theta_i' = (\theta_{i1}, ..., \theta_{iT})$, and $u_i' = (u_{i1}, ..., u_{iT})$, we obtain the full conditional distributions as follows:

- $\beta \mid Y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim$
  $N((X'X)^{-1}X'(\log(\theta) - u), \sigma_v^2 (X'X)^{-1})$;

- $\sigma_v^2 \mid Y, \beta, \sigma_\varepsilon^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_1 + mT/2, b_1 + \sum_{i=1}^{m}\sum_{t=1}^{T}(\log(\theta_{it}) - x_{it}'\beta - u_{it})^2\right)/2\right)$;

- $\sigma_\varepsilon^2 \mid Y, \beta, \sigma_v^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_2 + m(T-1)/2, b_2 + \sum_{i=1}^{m}\sum_{t=2}^{T}(u_{it} - u_{i,t-1})^2\right)/2\right)$;

- For $i = 1, ..., m$,
  $u_{i1} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{i1}) - x_{i1}'\beta}{\sigma_v^2} + \frac{u_{i2}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$, and $2 \le t \le T-1$,
  $u_{it} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1}, u_{i,t+1}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{it}) - x_{it}'\beta}{\sigma_v^2} + \frac{u_{i,t-1} + u_{i,t+1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$,
  $u_{iT} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,T-1}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{iT}) - x_{iT}'\beta}{\sigma_v^2} + \frac{u_{i,T-1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$,
  $\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto$
  $$\left|\sum\nolimits_i\right|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - \theta_i)'\sum\nolimits_i^{-1}(y_i - \theta_i)\right\}$$
  $$\times \exp\left\{-\frac{1}{2\sigma_v^2}\sum_{t=1}^{T}(\log(\theta_{it}) - x_{it}'\beta - u_{it})^2\right\}\left(\prod_{t=1}^{T}\frac{1}{\theta_{it}}\right).$$

Acknowledgements

The author would like to thank the Editor, the Associate Editor and one referee for their comments and suggestions. This work was partially supported by Statistics Canada Methodology Branch Research Block Fund.

References

Bayarri, M.J., and Berger, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency*: *A Methodological Perspective*, CD-ROM.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.

Datta, G.S., Day, B. and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā*, 60, 344-362.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.

Dick, P. (1995). Modeling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, 21, 45-54.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue No. 71-526.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Gelman, A., and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Laud, P., and Ibrahim, J. (1995). Predictive model selection. *Journal of Royal Statistical Society*, Series B, 57, 247-262.

Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical methods Conference proceedings*, *Washington*, *D.C.*, *www.fcsm.gov*.

Singh, A., You, Y. and Mantel, H. (2005). Use of generalized design effects for variance function modeling in small area estimation from survey data. Presentation at the 2005 Statistical Society of Canada Annual Meeting, Regina, SK.

Sinharay, S., and Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You. Y., Chen, E. and Gambino, G. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics*, Alexandria, VA: American Statistical Association. 3883-3888.

You, Y., and Dick, P. (2004). Hierarchical Bayes small area inference to the 2001 census undercoverage estimation. *2004 Proceedings of the American Statistical Association*, *Section on Government Statistics*, Alexandria, VA: American Statistical Association, 1836-1840.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. and Gambino, J. (2000). Hierarchical Bayes estimation of unemployment rates for sub-provincial regions using cross-sectional and time series data. *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*, 160-165.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 29-36
Statistics Canada, Catalogue No. 12-001-X

Small area estimation under a restriction

Junyuan Wang, Wayne A. Fuller and Yongming Qu [1]

Abstract

Small area prediction based on random effects, called EBLUP, is a procedure for constructing estimates for small geographical areas or small subpopulations using existing survey data. The total of the small area predictors is often forced to equal the direct survey estimate and such predictors are said to be calibrated. Several calibrated predictors are reviewed and a criterion that unifies the derivation of these calibrated predictors is presented. The predictor that is the unique best linear unbiased predictor under the criterion is derived and the mean square error of the calibrated predictors is discussed. Implicit in the imposition of the restriction is the possibility that the small area model is misspecified and the predictors are biased. Augmented models with one additional explanatory variable for which the usual small area predictors achieve the self-calibrated property are considered. Simulations demonstrate that calibrated predictors have slightly smaller bias compared to those of the usual EBLUP predictor. However, if the bias is a concern, a better approach is to use an augmented model with an added auxiliary variable that is a function of area size. In the simulation, the predictors based on the augmented model had smaller MSE than EBLUP when the incorrect model was used for prediction. Furthermore, there was a very small increase in MSE relative to EBLUP if the auxiliary variable was added to the correct model.

Key Words: Components-of-variance model; Best linear unbiased prediction; Calibration; Design consistent; Mixed linear models.

1. Introduction

There are situations in which it is desirable to derive reliable estimators for small geographical areas or small subpopulations from existing survey data. However, sample sizes for the areas may be such that the usual survey estimators yield unacceptably large standard errors. This makes it reasonable to use a model-based estimator. See Rao (2003) for a complete discussion of small area estimation.

A model for small area estimation is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i, \tag{1}$$

$$Y_i = y_i + e_i, \qquad i = 1, ..., n, \tag{2}$$

where y_i are unobservable small area means, Y_i are observable survey estimators, $\boldsymbol{x}_i'$ are known vectors, $\boldsymbol{\beta}$ is the vector of regression parameters, b_i are independent and identically distributed random variables with $E(b_i) = 0$ and $V(b_i) = \sigma_b^2$, and e_i are sampling errors with $E(e_i \mid y_i) = 0$ and $V(e_i \mid y_i) = \sigma_{ei}^2$. Combining (1) and (2), we obtain

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, ..., n, \tag{3}$$

which is a special case of the mixed linear model.

Assuming the variance components σ_b^2 and σ_{ei}^2 to be known, the best linear unbiased estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \left[\sum_{i=1}^{n}(\sigma_b^2 + \sigma_{ei}^2)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}(\sigma_b^2 + \sigma_{ei}^2)^{-1}\boldsymbol{x}_i Y_i\right], \tag{4}$$

where $\boldsymbol{X}' = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $\boldsymbol{Y}' = (Y_1, ..., Y_n)$, and $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y}) = \mathrm{diag}(\sigma_b^2 + \sigma_{e1}^2, ..., \sigma_b^2 + \sigma_{en}^2)$. Furthermore, the best linear unbiased predictor (BLUP) of y_i is

$$\tilde{y}_i^H = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \gamma_i(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}), \tag{5}$$

where

$$\gamma_i = (\sigma_b^2 + \sigma_{ei}^2)^{-1}\sigma_b^2. \tag{6}$$

See Henderson (1963) and Rao (2003). When the variance components are unknown, we replace the variance components in (4) and (6) with estimators to obtain $\hat{y}_i^H$, the empirical BLUP or EBLUP.

The survey estimator of the total of all survey areas is often judged to be of adequate precision. In such cases, the practitioner may prefer to use the design consistent estimator of the total and to require that the weighted sum of the small area predictors equal the design consistent estimator. Thus, it is desirable to have small area predictors $\hat{y}_i$ that satisfy

$$\sum_{i=1}^{n}\omega_i\hat{y}_i = \sum_{i=1}^{n}\omega_i Y_i, \tag{7}$$

where ω_i are sampling weights such that $\sum_{i=1}^{n}\omega_i Y_i$ is a design consistent estimator of the total (or mean). A number of procedures have been suggested for constructing predictors to satisfy (7). Such procedures are often called "benchmarking" or "calibration", *e.g.*, Mantel, Singh and Barcau (1993) and You and Rao (2003).

To review such procedures, let $\tilde{\boldsymbol{y}}^H = (\tilde{y}_1^H, ..., \tilde{y}_n^H)'$ denote the BLUP predictor of $\boldsymbol{y} = (y_1, ..., y_n)'$ defined in (5), where

1. Junyuan Wang, Director of Biostatistics, Global Biostatistics and Programming, Wyeth Research, 35 Cambridge Park Drive, Cambridge, MA 02140; Wayne A. Fuller, Professor, Department of Statistics, Iowa State University, Ames, IA 50011; Yongming Qu, Sr. Research Scientist, Eli Lilly and Company, Lilly Corporation Center, Indianapolis, IN 26285.

$$\tilde{\boldsymbol{y}}^H = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}, \tag{8}$$

$\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ are any solutions to

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1} \\ \boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{\Sigma}_e^{-1} + \boldsymbol{\Sigma}_b^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \\ \boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \end{bmatrix}, \tag{9}$$

$\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}_n$, and $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, ..., \sigma_{en}^2)$. Equation (9) is called the mixed model equation. Finding a solution to a mixed model equation (9) is equivalent to finding a solution to the minimization problem

$$\min_{\boldsymbol{\beta},\boldsymbol{b}} \{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b})'\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b}) + \boldsymbol{b}'\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}\}. \tag{10}$$

Pfeffermann and Barnard (1991) proposed the modified predictor

$$\hat{\boldsymbol{y}}^{\text{PB}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{PB}} + \hat{\boldsymbol{b}}^{\text{PB}}, \tag{11}$$

where $\hat{\boldsymbol{\beta}}^{\text{PB}}$ and $\hat{\boldsymbol{b}}^{\text{PB}}$ are any solutions to the minimization problem (10) with $\boldsymbol{\beta}$ and $\boldsymbol{b}$ subject to the constraint

$$\sum_{i=1}^{n} \omega_i(\boldsymbol{x}_i'\boldsymbol{\beta} + b_i) = \sum_{i=1}^{n} \omega_i Y_i. \tag{12}$$

This leads to the predictor

$$\hat{y}_i^{\text{PB}} = \tilde{y}_i^H + [\text{Var}(\tilde{y})]^{-1}\text{cov}(\tilde{y}_i^H, \tilde{y})\left[\sum_{j=1}^{n} \omega_j Y_j - \tilde{y}\right], \tag{13}$$

where $\tilde{y} = \sum_{i=1}^{n} \omega_i \tilde{y}_i^H$, $\text{cov}(\tilde{y}_i^H, \tilde{y}) = \omega_i \gamma_i \sigma_{ei}^2 + \sum_{j=1}^{n} \omega_j (1 - \gamma_i)(1 - \gamma_j) \boldsymbol{x}_i' V(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_j$, and $\text{Var}(\tilde{y}) = \sum_{i=1}^{n} \omega_i \text{cov}(\tilde{y}_i^H, \tilde{y})$.

Isaki, Tsay and Fuller (2000) imposed the restriction by a procedure that, approximately, constructs the best predictors of $n - 1$ quantities that are uncorrelated with $\sum_{i=1}^{n} \omega_i Y_i$. After some matrix operations, the Isaki-Tsay-Fuller (ITF) predictor can be rewritten as

$$\hat{y}_i^{\text{ITF}} = \hat{y}_i^H + \left[\sum_{j=1}^{n} \omega_j^2 \widehat{\text{Var}}(Y_j)\right]^{-1} \omega_i \widehat{\text{Var}}(Y_i)\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j^H\right), \tag{14}$$

where $\widehat{\text{Var}}(Y_i)$ is an estimator of $\sigma_b^2 + \sigma_{ei}^2$.

Note that the Pfeffermann-Barnard (PB) predictor (13), and the ITF predictor (14) have the form

$$\hat{y}_i^{\boldsymbol{a}} = \hat{y}_i + a_i\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j\right), \tag{15}$$

where $\sum_{i=1}^{n} \omega_i a_i = 1$. In other words, we may consider imposing restriction (7) to be an adjustment problem. To make an adjusted predictor $\hat{y}_i^{\boldsymbol{a}}$ satisfy (7), the difference $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ is allocated to small area predictor $\hat{y}_i$ using a_i.

Using the unit level model, You and Rao (2002) proposed an estimator of $\boldsymbol{\beta}$ such that the resulting predictors satisfy (7). They called such predictors self-calibrated. Applying their procedure to the area model (3), we have

$$\hat{y}_i^{\text{YR}} = \hat{\gamma}_i Y_i + (1 - \hat{\gamma}_i)\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\text{YR}}, \tag{16}$$

where

$$\hat{\boldsymbol{\beta}}_{\text{YR}} = \left[\sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1} \sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i Y_i. \tag{17}$$

Any predictor that has the self-calibrated property, such as the You and Rao (YR) predictor (16), is a predictor of the form (15) since the difference $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ is equal to zero.

We will derive the "best" predictor of the form (15) in Section 2. The results will lead to a unifying view of several BLUP based predictors. In Section 3, we propose an alternative approach that has the self-calibrated property. We will briefly discuss the mean square error (MSE) in Section 4 and use simulation studies to compare the predictors in Section 5. Conclusions and discussion will be given in Section 6.

2. "Best" linear unbiased predictor under a restriction

To find the "best" linear unbiased predictor for $\boldsymbol{y}$ that satisfies restriction (7), we first assume the parameters for the variance components are known. According to Lemma 1 of Pfeffermann and Barnard (1991), it is impossible to compare predictors that satisfy restriction (7) component-by-component to find the best one. Therefore, some kind of overall criterion is required. A natural criterion is

$$Q(\hat{\boldsymbol{y}}^{\boldsymbol{a}}) = \sum_{i=1}^{n} \varphi_i E(\hat{y}_i^{\boldsymbol{a}} - y_i)^2, \tag{18}$$

where the φ_i, $i = 1, ..., n$ are a chosen set of positive weights.

Theorem 1. *Assume the random effects model*

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, ..., n,$$

where the b_i have independent identical distributions with mean zero and variance σ_b^2, the e_i have independent distributions with mean zero and variance σ_{ei}^2, and $\boldsymbol{b} = (b_1, ..., b_n)'$ is independent of $\boldsymbol{e} = (e_1, ..., e_n)'$. Assume σ_b^2 and σ_{ei}^2 are known, and $\boldsymbol{\beta}$ is unknown. Let $\tilde{y}_i^H$ be the BLUP of y_i defined in (5)*. Let*

$$\hat{y}_i^{\breve{a}} = \tilde{y}_i^H + \breve{a}_i \left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H \right), \qquad (19)$$

where $\breve{a}_i = (\sum_{i=1}^n \varphi_i^{-1} \omega_i^2)^{-1} \varphi_i^{-1} \omega_i$ *and* ω_i *are the fixed weights of (7). Then* $\hat{\boldsymbol{y}}^{\breve{a}} = (\hat{y}_1^{\breve{a}}, ..., \hat{y}_n^{\breve{a}})'$ *is the unique predictor among all linear unbiased predictors that satisfy* (7) *and minimize criterion* (18).

Proof: See Appendix A.

Remark 1. When the variance components are unknown, we replace the variance components in (6) with suitable estimators to obtain the empirical BLUP or EBLUP, denoted by $\hat{y}_i^H$. Thus, we have the modified predictor

$$\hat{y}_i^{\breve{a}} = \hat{y}_i^H + \breve{a}_i \left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H \right). \qquad (20)$$

Remark 2. The criterion (18) defines a "loss function", where the choice of the weights φ_i depends on the problem under consideration. For example, a statistician can decide to assign higher weights to "more important" areas and lower weights to "less important" areas. Often, φ_i is function of the variance components. In some cases, one can choose φ_i so that the derived predictors have certain desirable properties. For example, $\varphi_i = [\widehat{\text{Var}}(Y_i)]^{-1}$ gives the ITF predictors, which are BLUP (in the traditional sense) in the subspace that is orthogonal to $(\omega_1, ..., \omega_n)'$ in the space spanned by $\boldsymbol{Y}$.

Remark 3. If $\varphi_i = \omega_i [\text{cov}(\tilde{y}_i^H, \tilde{y}_\cdot)]^{-1}$, where $\tilde{y}_\cdot = \sum_{j=1}^n \omega_j \tilde{y}_j^H$, we have the predictor (13) derived by Pfeffermann and Barnard (1991). When $\varphi_i = [\widehat{\text{Var}}(\hat{y}_j^H)]^{-1}$, we have the predictor used by Battese, Harter, and Fuller (1988).

3. An alternative way to impose the restriction

We have discussed a family of predictors in which the total for the small area predictors is equal to the total of the direct survey estimates. Implicit in the imposition of restriction (7) is the possibility that the small area predictor of the total is biased due to a misspecified model (3). In practical applications, model misspecification is a valid concern since the true mechanism that generates $\boldsymbol{Y}$ is unknown.

A common misspecification occurs when the explanatory variables used in the model are not the same as the ones that generated $\boldsymbol{Y}$. Thus, the direction of the overall bias may not be the same as the direction of the bias for a particular small area. In this case, the predictors of form (15) may increase the bias for some small areas compared to the bias before adjustment. Mantel *et al.* (1993) concluded that "Generally the effect of benchmarking here is a slight improvement in the overall bias at the cost of some deterioration with respect to the other evaluation measures".

Since the bias is nonzero if there is nonzero correlation between ω_i and $(Y_i - \hat{y}_i)$, the bias can be reduced by including ω_i in the model. That is, for a given model, one approach is to use the augmented model

$$\boldsymbol{Y} = \boldsymbol{X}_1 \boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (21)$$

where $\boldsymbol{X}_1 = (\boldsymbol{X}, \boldsymbol{\omega})$ and $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$, to obtain the BLUP or EBLUP. With $\boldsymbol{\omega}$ in the model, the adjustment needed to meet restriction (7) will often be much smaller than the adjustment for the model without $\boldsymbol{\omega}$.

Using an augmented model approach, we can go one step further and construct predictors that satisfy restriction (7). First, assume the variances σ_{ei}^2 are known. Note that

$$\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H = \sum_{i=1}^n \omega_i (1 - \gamma_i)(Y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}) \qquad (22)$$

and $\omega_i (1 - \gamma_i) \text{Var}(Y_i) = \omega_i \sigma_{ei}^2$. Using the theory of the linear model, we can show that the predictor constructed with the augmented model

$$\boldsymbol{Y} = \boldsymbol{X}_2 \boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (23)$$

where $\boldsymbol{X}_2 = (\boldsymbol{X}, \boldsymbol{\omega}_e)$ and $\boldsymbol{\omega}_e = (\omega_1 \sigma_{e1}^2, ..., \omega_n \sigma_{en}^2)'$, has the self-calibrated property when the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ is used. Note that this approach gives a predictor that is different than the You-Rao predictor (16).

If the σ_{ei}^2 are unknown, we replace σ_{ei}^2 in $\boldsymbol{\omega}_e$ with its estimator $\hat{\sigma}_{ei}^2$. As long as the $\hat{\sigma}_{ei}^2$ in $\boldsymbol{\omega}_e$ is the same as the $\hat{\sigma}_{ei}^2$ used in constructing γ_i, the predictors have the self-calibrated property. If the σ_{ei}^2 has the form $\sigma_e^2 f(\boldsymbol{u}_i)$, where σ_e^2 is unknown, but $\boldsymbol{u}_i$ and $f(.)$ are known, one can construct the variables for model (23) using $\boldsymbol{\omega}_e = (\omega_1 f(\boldsymbol{u}_1), ..., \omega_n f(\boldsymbol{u}_n))'$ without estimating σ_e^2. For example, if $\sigma_{ei}^2 = m_i^{-1} \sigma_e^2$, the $\boldsymbol{\omega}_e$ for model (23) is $(m_i^{-1} \omega_1, ..., m_n^{-1} \omega_n)'$.

4. The MSE of the modified predictors

One can show that any predictor of the form (15) can be written as

$$\tilde{\boldsymbol{y}}^{\boldsymbol{a}} = \boldsymbol{Y} - \boldsymbol{C}_{\boldsymbol{a}}^{-1} \boldsymbol{B} \boldsymbol{C}_{\boldsymbol{a}} (\boldsymbol{I}_n - \boldsymbol{\Gamma})(\boldsymbol{Y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}) \qquad (24)$$

by letting $\boldsymbol{C}_{\boldsymbol{a}} = \boldsymbol{A}_{\boldsymbol{a}} \boldsymbol{T}$, where

$$\boldsymbol{A}_{\boldsymbol{a}} = \begin{pmatrix} 1 & \boldsymbol{0}_{n-1}' \\ -\boldsymbol{a}_{n-1} & \boldsymbol{I}_{n-1} \end{pmatrix},$$

$$\boldsymbol{a}_{n-1} = (a_2, ..., a_n)'.$$

Therefore, the estimator for the variance of $\tilde{\boldsymbol{y}}^{a}$ defined in (14) proposed in Isaki *et al.* (2000) can be used to estimate the MSE of any predictor of the form (15). Often, the MSE of an adjusted predictor is close to the MSE of the predictor before the adjustment.

The augmented model (23) has the self-calibrated property, thus the MSE can be estimated using the formula for usual EBLUP predictors.

5. Simulation study

5.1 Simulation setup

To study the empirical properties of small area predictors described in Section 2 and 3, we use data designed to simulate a large national survey in which state estimates are of interest. Table 1 contains the approximate populations of the 50 states of the United States in the year 2000. The sample sizes m_i given in the table are approximately proportional to the square roots of the state populations.

Table 1 Population and sample sizes for the simulation

State	Population (in 1,000)	Sample size (m_i)	State	Population (in 1,000)	Sample size (m_i)
1	33,640	58	26	4,000	20
2	21,160	46	27	3,610	19
3	19,360	44	28	3,240	18
4	16,000	40	29	3,240	18
5	12,250	35	30	2,890	17
6	12,250	35	31	2,890	17
7	11,560	34	32	2,560	16
8	10,240	32	33	2,560	16
9	8,410	29	34	2,250	15
10	8,410	29	35	1,960	14
11	7,840	28	36	1,690	13
12	7,290	27	37	1,690	13
13	6,250	25	38	1,690	13
14	6,250	25	39	1,210	11
15	5,760	24	40	1,210	11
16	5,760	24	41	1,210	11
17	5,760	24	42	1,210	11
18	5,290	23	43	1,000	10
19	5,290	23	44	810	9
20	5,290	23	45	810	9
21	4,840	22	46	810	9
22	4,410	21	47	640	8
23	4,410	21	48	640	8
24	4,410	21	49	640	8
25	4,000	20	50	490	7

A total of 10,000 samples were generated. Each sample in the simulation study was composed of observations generated from model

$$Y_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \tag{25}$$

where $\boldsymbol{x}_i' = (1,\ z_i)$, $\boldsymbol{\beta}' = (6.0,\ 3.0)$, $z_i = \mathrm{Pop}_i^{0.2} - \overline{\mathrm{Pop}^{0.2}}$, Pop_i is the population of state i in millions, $\overline{\mathrm{Pop}^{0.2}}$ is the mean of $\mathrm{Pop}_i^{0.2}$, $b_i \sim NI(0,\ 1)$, and $\varepsilon_{ij} \sim NI(0,\ 16)$. The b_i's and e_{ij}'s are independent. The model for the state observations becomes

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \tag{26}$$

where $Y_i = m_i^{-1}\sum_{j=1}^{m_i} Y_{ij}$ and $e_i = m_i^{-1}\sum_{j=1}^{m_i}\varepsilon_{ij}$. With the sample sizes given in Table 1, the γ_i defined in (6) is 0.784 for the largest state (California) and 0.304 for the smallest state (Wyoming).

To investigate the performance of five predictors, EBLUP, Pfeffermann-Bernard (PB), Isaki-Tsay-Fuller (ITF), You-Rao (YR), and augmented model (23) (AUG2), two estimation models were used. The first model is a misspecified model with $\boldsymbol{x}_i' = 1$ in the notation of (26). This model is called model (A). Correspondingly, the data generating model (26) with $\boldsymbol{x}_i' = (1,\ z_i)$ is called model (B).

Following the method outlined in Wang and Fuller (2003), the estimator of σ_b^2 is

$$\hat{\sigma}_b^2 = \max\{0.5[\hat{V}(\tilde{\sigma}_b^2)]^{0.5},\ \tilde{\sigma}_b^2\}, \tag{27}$$

where

$$\tilde{\sigma}_b^2 = \sum_{i=1}^{50} c_i\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2 - \hat{\sigma}_{ei}^2\right], \tag{28}$$

$$\hat{V}(\tilde{\sigma}_b^2) = \sum_{i=1}^{50} c_i^2\left\{\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2 - \hat{\sigma}_{ei}^2\right] - \tilde{\sigma}_b^2\right\}^2, \tag{29}$$

$c_i = (\sum_{i=1}^{50} m_i^{0.5})^{-1} m_i^{0.5}$, $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is the ordinary least squares estimator of the regression coefficient of Y_i on $\boldsymbol{x}_i$, k is the dimension of the vector $\boldsymbol{x}_i$, $\hat{\sigma}_{ei}^2 = m_i^{-1}s_i^2$, and $s_i^2 = (m_i - 1)^{-1}\sum_{j=1}^{m_i}(Y_{ij} - Y_i)^2$ is the sample variance of area i.

The EBLUP predictor is

$$\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} + \hat{\gamma}_i(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}), \tag{30}$$

where

$$\hat{\gamma}_i = (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1}\hat{\sigma}_b^2, \tag{31}$$

and

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left[\sum_{i=1}^{n}(\sigma_b^2 + \hat{\sigma}_{ei}^2)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}(\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1}\boldsymbol{x}_i Y_i\right] \tag{32}$$

is the generalized least squares estimator of $\boldsymbol{\beta}$. The restriction considered is $\sum_{i=1}^{50}\omega_i\hat{y}_i^{a} = \sum_{i=1}^{50}\omega_i Y_i$, where

$$\omega_i = \left(\sum_{i=1}^{50}\mathrm{Pop}_i\right)^{-1}\mathrm{Pop}_i. \tag{33}$$

With EBLUP predictor, PB and ITF are derived using (13) and (14). The YR predictor is derived use (16) with $\hat{\gamma}_i$ defined in (31) and $\hat{\boldsymbol{\beta}}_{\mathrm{YR}}$ defined in (17). The AUG2

predictor is of the form (30) with $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$ under the augmented model (A) and $\boldsymbol{x}_i' = (1, z_i, \omega_i \hat{\sigma}_{ei}^2)$ under the augmented model (B).

For each of the 10 predictors, the criterion

$$Q(\hat{\boldsymbol{y}}) = 0.02 \sum_{i=1}^{50} \varphi_i (\hat{y}_i - y_i)^2 \tag{34}$$

was calculated, where $\varphi_i = (\gamma_i m_i^{-1} \sigma_e^2)^{-1}$. Note that φ_i^{-1} is the variance of the predictor of y_i constructed with known parameters.

5.2 Simulation results

The estimator of σ_b^2 constructed under model (B) has a Monte Carlo mean of 1.001 with a standard deviation of 0.386. When model (A) is used for estimation, the Monte Carlo mean of the estimator of σ_b^2 is 1.720 with a standard deviation of 0.521. The mean square of $3z_i$ is 0.636. Thus, the estimation procedure using model (A) incorporates much of the fixed area effect of model (B) into the random effect. With a bigger $\hat{\sigma}_b^2$, the $\hat{\gamma}_i$ is bigger and a higher proportion of Y_i was used to construct the predictors, which partially offsets the impact of model misspecification. Under augmented model (A), the estimator of σ_b^2 is 1.197, *i.e.*, a much smaller portion of the fixed area effect of model (B) is incorporated into the random effect.

Table 2 contains some summary statistics for predictors based on 10,000 simulated samples. The empirical bias in $\sum \omega_i (Y_i - \hat{y}_i)$ is zero for all predictors when model (B) or its augmented model is used because the prediction model matches the data-generating model. The difference $\sum \omega_i (Y_i - \hat{y}_i)$ has a simulated standard deviation of 0.022 for the usual EBLUP predictor. The standard deviations of $\sum \omega_i (Y_i - \hat{y}_i)$ are 0 for the other four predictors because the predictors satisfy the restriction $\sum_{i=1}^{50} \omega_i \hat{y}_i = \sum_{i=1}^{50} \omega_i Y_i$.

Table 2 Monte Carlo properties of small area predictors (average of 10,000 samples generated by model B)

Quantity		EBLUP	PB	ITF	YR	AUG2
		Predictor constructed under model (A)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Mean	-0.100	0.000	0.000	0.000	0.000
	(SD)	(0.027)	(0.000)	(0.000)	(0.000)	(0.000)
$Q(\hat{\boldsymbol{y}})$	Mean	1.438	1.446	1.419	1.558	1.298
		Predictor constructed under model (B)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Mean	-0.000	0.000	0.000	0.000	0.000
	(SD)	(0.022)	(0.000)	(0.000)	(0.000)	(0.000)
$Q(\hat{\boldsymbol{y}})$	Mean	1.203	1.202	1.202	1.208	1.219

Prediction based on model (A), or its augmented model, is biased because the data generation model (B) contains a function of the population size. The simulated mean of the weighted difference $\sum \omega_i (Y_i - \hat{y}_{iA})$ is -0.100, where $\hat{y}_{iA}$ is the EBLUP predictor. The *t*-statistic for the weighted bias is -3.70. The simulated variance of the weighted mean of the predictions is

$$V\left\{\sum_{i=1}^{50} \omega_i \hat{y}_{iA}\right\} = 0.060.$$

The estimated mean square error of the model (A) prediction of $\sum \omega_i y_i$ for data generated by model (B) is

$$\begin{aligned} \text{MSE}\left\{\sum_{i=1}^{50} \omega_i (\hat{y}_{iA} - y_i)\right\} &= V\left\{\sum_{i=1}^{50} \omega_i \hat{y}_{iA}\right\} + \text{Bias}^2 \\ &= 0.060 + (-0.100)^2 = 0.070. \end{aligned} \tag{35}$$

The variance of $\sum \omega_i Y_i$ is

$$V\left\{\sum_{i=1}^{50} \omega_i Y_i\right\} = \sum_{i=1}^{50} \omega_i^2 (\sigma_b^2 + m_i^{-1} \sigma_e^2) = 0.0622. \tag{36}$$

Thus, the use of $\sum \omega_i Y_i$ as the estimator of $\sum \omega_i y_i$ is about 12.5% more efficient than the predictor $\sum \omega_i \hat{y}_{iA}$ based on the model (A). Due to calibration, the MSE of the four predictors (PB, ITF, YR, and AUG2) of $\sum \omega_i y_i$ have the same MSE as the MSE of the directly estimated mean $\sum \omega_i Y_i$. The squared bias would be a much larger proportion of the mean square error if there were more small areas.

The value of criterion Q for EBLUP is 1.20 under model (B). Thus, estimation of the parameters increased the average variance of the predictors about 20% relative to the use of known parameters. If the predictions are made using the known σ_{ei}^2, the value of criterion Q for EBLUP is 1.06. Therefore, estimation of $\hat{\sigma}_{ei}^2$ contributes the most to the increase in the variability. Because the bias is zero when model (B) is used for estimation, the adjustments that the restricted predictors make are small. Consequently, the adjustment predictors give criteria values very similar to those of the unadjusted predictors. The YR predictors have slightly larger criterion values than the corresponding PB and ITF predictors because the YR predictor uses an inefficient estimator of $\boldsymbol{\beta}$. Predictors based on the augmented model have slightly larger criterion value Q because the model has a redundant variable. The less than 2% increase in Q is on the order of n^{-1}, which is the expected loss from adding an unnecessary parameter in a least square prediction.

The value of criterion Q for the EBLUP under model (A) is 1.438 compared to 1.203 for the EBLUP under model (B). This is the penalty due to the model misspecification. Among adjustment procedures based on model (A), the ITF procedure has the smallest value for Q. There is little difference among the PB, ITF, and the EBLUP predictors. The Q for the You-Rao procedure is about 8% larger than that for the EBLUP. The predictor based on the augmented

model has a Q about 11% smaller than that of the EBLUP predictor. In a sense, the augmented model is less misspecified.

The augmentation approach not only calibrates the small area predictors, it also reduces the bias at the area level when the model is misspecified. Tables 3 and 4 contain Monte Carlo properties of predictors for some selected areas. In the tables, the estimated biases are normalized by $(\gamma_i m_i^{-1} \sigma_e^2)^{0.5}$, the square root of the MSE of the BLUP predictor with known parameters, and the estimated MSE's are normalized by $\gamma_i m_i^{-1} \sigma_e^2$. When the correct model (B) is used, the Monte Carlo bias at the individual area is close to zero. See Table 3. Also there is little difference in the MSE's of the different procedures, with the augmented model having slightly (less than 2%) larger MSE. Again, Table 3 shows that the calibration process has little effect on the MSE of the area predictors in compared to the EBLUP predictor.

Table 3
Monte Carlo Properties of individual area predictors using versions of model (B) (10,000 samples generated by model B)

State	Quantity	EBLUP	PB	ITF	YR	AUG2
1	Bias	0.011	0.012	0.012	0.013	0.011
	MSE	1.100	1.101	1.105	1.104	1.120
2	Bias	0.000	0.001	0.001	0.002	0.001
	MSE	1.072	1.072	1.072	1.076	1.092
14	Bias	-0.001	-0.001	-0.001	-0.001	-0.001
	MSE	1.058	1.058	1.058	1.058	1.074
26	Bias	0.015	0.015	0.015	0.015	0.018
	MSE	1.078	1.077	1.078	1.079	1.092
38	Bias	-0.005	-0.005	-0.005	-0.006	-0.003
	MSE	1.123	1.122	1.122	1.132	1.135
50	Bias	0.012	0.012	0.012	0.009	0.014
	MSE	1.222	1.222	1.222	1.247	1.246

If a misspecified model, such as model (A) is used, the bias in the sum of the EBLUP predictors as an estimator of the total is negative for the example because the states with a negative bias have large ω_i. See Table 4. The adjustment procedures such as PB or ITF allocate the bias to all the small areas. Thus, the adjustment reduces the negative bias of area predictors with large negative bias and increases the positive bias of predictors with large positive bias. This results in a smaller MSE for larger states and a slightly larger MSE for smaller states. The YR predictor has larger bias than the ITF predictor. On the other hand, predictors constructed with $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$, *i.e.*, the augmented model (A), are much superior to those constructed under model (A). The bias is reduced for areas, large or small.

Table 4
Monte Carlo Properties of individual area predictors using versions of model (A) (10,000 samples generated by model B)

State	Quantity	EBLUP	PB	ITF	YR	AUG2
1	Bias	-0.597	-0.225	-0.052	-0.473	-0.030
	MSE	1.471	1.165	1.130	1.331	1.139
2	Bias	-0.496	-0.227	-0.170	-0.358	-0.070
	MSE	1.330	1.134	1.115	1.207	1.124
14	Bias	-0.121	0.004	-0.031	0.055	-0.025
	MSE	1.100	1.085	1.086	1.089	1.105
26	Bias	0.057	0.157	0.115	0.249	0.053
	MSE	1.126	1.148	1.136	1.188	1.132
38	Bias	0.380	0.453	0.406	0.601	0.202
	MSE	1.340	1.405	1.361	1.571	1.233
50	Bias	0.922	0.980	0.931	1.178	0.537
	MSE	2.196	2.316	2.215	2.767	1.577

6. Conclusions

In this paper, several calibrated predictors are reviewed. We offer a fresh look at the benchmarking restriction (7). Imposing the restriction is viewed as an adjustment problem and a criterion that unifies the derivation of calibrated predictors is presented. The criterion approach to the problem opens the door for consideration of other predictors.

Implicit in the imposition of the restriction is the possibility that the small area model is misspecified and the predictors are biased. When the model is misspecified, the calibration adjustment only adjusts for the overall bias, not for the bias at the small area level. The augmented model approach leads to a self-calibrated predictor and reduces the bias at the small area level. Also, variance estimation for the externally calibrated predictors is relatively complex, while variance estimation for self-calibrated predictors is straightforward. In summary, if the bias is a concern, use of a self-calibrated augmented model is preferred to external calibration.

7. Appendix

Proof of Theorem 1: Let $\tilde{y}_i^H$ be the BLUP of y_i and let $\hat{y}_i$ be any linear unbiased predictor of y_i. By standard results for BLUP (see, for example, Robinson (1991) and Harville (1976)), we have

$$\text{cov}(\tilde{y}_i^H - y_i,\ \hat{y}_j - \tilde{y}_j^H) = 0, \quad \text{if} \quad i \neq j, \qquad \text{(A.1)}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. Let $R(\hat{\boldsymbol{y}}^a)$ denote the collection of all linear unbiased predictors that satisfy (7). For any $\hat{\boldsymbol{y}} \in R(\hat{\boldsymbol{y}}^a)$, by (A.1), we have

$$E\{(\hat{y}_i - y_i)^2\} = E\{(\tilde{y}_i^H - y_i)^2\} + E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \quad \text{(A.2)}$$

Therefore,

$$Q(\hat{\boldsymbol{y}}) = \sum_{i=1}^{n} \varphi_i E\{(\hat{y}_i - y_i)^2\}$$

$$= \sum_{i=1}^{n} \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + \sum_{i=1}^{n} \varphi_i E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \quad \text{(A.3)}$$

Since $\hat{\boldsymbol{y}}$ satisfies (7), we have $\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i$ and, for the $\hat{y}_i^{\breve{a}}$ defined in (19),

$$\hat{y}_i^{\breve{a}} = \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^{n} \omega_j (Y_j - \tilde{y}_j^H) \right]$$

$$= \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^{n} \omega_j (\hat{y}_j - \tilde{y}_j^H) \right].$$

By (A.1),

$$Q(\hat{y}_i^{\breve{a}}) = \sum_{i=1}^{n} \varphi_i E(\hat{y}_i^{\breve{a}} - y_i)^2$$

$$= \sum_{i=1}^{n} \varphi_i E\left\{ \left[(\tilde{y}_i^H - y_i) + \breve{a}_i \left(\sum_{j=1}^{n} \omega_j \hat{y}_j - \sum_{j=1}^{n} \omega_j \tilde{y}_j^H \right) \right]^2 \right\}$$

$$= \sum_{i=1}^{n} \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + E\left\{ \left[\sum_{j=1}^{n} \omega_j (Y_j - \tilde{y}_j^H) \right]^2 \right\} \sum_{i=1}^{n} \varphi_i \breve{a}_i^2. \quad \text{(A.4)}$$

Since $\breve{a}_i = (\sum_{i=1}^{n} \varphi_i^{-1} \omega_i^2)^{-1} \varphi_i^{-1} \omega_i$ in (A.4), we have

$$Q(\hat{y}_i^{\breve{a}}) = \sum_{i=1}^{n} \varphi_i E\{(\tilde{y}_i^H - y_i)^2\}$$

$$+ E\left\{ \left[\sum_{j=1}^{n} \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \left(\sum_{i=1}^{n} \varphi_i^{-1} \omega_i^2 \right)^{-1}. \quad \text{(A.5)}$$

Note that

$$E\left\{ \left[\sum_{j=1}^{n} \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \le \sum_{j=1}^{n} \sum_{k=1}^{n} \omega_j \omega_k g_j g_k$$

$$= \left(\sum_{i=1}^{n} \omega_i g_i \right)^2, \quad \text{(A.6)}$$

where $g_j = \{E[(\hat{y}_j - \tilde{y}_j^H)^2]\}^{0.5}$. By Cauchy's inequality,

$$\left(\sum_{j=1}^{n} \omega_j g_j \right)^2 \le \left(\sum_{i=1}^{n} \varphi_i^{-1} \omega_i^2 \right) \left(\sum_{i=1}^{n} \varphi_i g_i^2 \right)$$

$$= \left(\sum_{i=1}^{n} \varphi_i^{-1} \omega_i^2 \right) \left[\sum_{i=1}^{n} \varphi_i E\{(\hat{y}_j - \tilde{y}_j^H)^2\} \right]. \quad \text{(A.7)}$$

Combining (A.3), (A.5), (A.6), and (A.7), we have $Q(\hat{y}_i^{\breve{a}}) \le Q(\hat{\boldsymbol{y}})$.

To show the uniqueness of $\hat{y}_i^{\breve{a}}$, we need to check when the inequalities (A.6) and (A.7) become equalities. Inequality (A.6) becomes an equality if and only if

$$\hat{y}_j - \tilde{y}_j^H = c_j^0 + c_j^1 (\hat{y}_1 - \tilde{y}_1^H) \quad \text{(A.8)}$$

for some constants c_j^0 and c_j^1, $j = 2, ..., n$. Inequality (A.7) becomes an equality if and only if

$$\sqrt{\varphi_i^{-1} \omega_i^2} \sqrt{v_j g_j^2} - \sqrt{v_j^{-1} \omega_j^2} \sqrt{\varphi_i g_i^2} = 0, \quad \text{(A.9)}$$

or, equivalently,

$$\varphi_i^2 \omega_i^{-2} E\{(\hat{y}_i - \tilde{y}_i^H)^2\} = v_j^2 \omega_j^{-2} E\{(\hat{y}_j - \tilde{y}_j^H)^2\}. \quad \text{(A.10)}$$

Also,

$$\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i. \quad \text{(A.11)}$$

Combining (A.8), (A.10), and (A.11), we have that the equality holds if and only if $\hat{y}_j = \hat{y}_i^{\breve{a}}$. Thus, we have shown that $\hat{y}_i^{\breve{a}}$ is the unique linear unbiased predictor that satisfies (7) and minimizes criterion (18).

Acknowledgements

This research was funded in part by cooperative agreement 68-3A75-14 between the USDA Natural Resources Conservation Service and Iowa State University.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of American Statistical Association*, 28-36.

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

Henderson, C.R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, (Eds., W.D. Hanson and H.F. Robinson), National Academy of Sciences–National Research Council, Washington, Publication 982, 141-163.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*, 31-42.

Mantel, H.J., Singh, A.C. and Barcau, M. (1993). Benchmarking of small area estimators. In *Proceedings of International Conference on Establishment Surveys*, American Statistical Association, Washington, DC, 920-925.

Pfeffermann, D., and Barnard, C.H. (1991). Some new estimators for small-area means with application to the assessment of Farmland values. *Journal of Business & Economic Statistics*, 73-84.

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley & Sons, Inc.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 15-51.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area estimators constructed with estimated area variances. *Journal of American Statistical Association*, 716-723.

You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 431-439.

You, Y., and Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 197-208.

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 37-49
Statistics Canada, Catalogue No. 12-001-X

A Bayesian allocation of undecided voters

Balgobin Nandram and Jai Won Choi [1]

Abstract

Data from election polls in the US are typically presented in two-way categorical tables, and there are many polls before the actual election in November. For example, in the Buckeye State Poll in 1998 for governor there are three polls, January, April and October; the first category represents the candidates (*e.g.*, Fisher, Taft and other) and the second category represents the current status of the voters (likely to vote and not likely to vote for governor of Ohio). There is a substantial number of undecided voters for one or both categories in all three polls, and we use a Bayesian method to allocate the undecided voters to the three candidates. This method permits modeling different patterns of missingness under ignorable and nonignorable assumptions, and a multinomial-Dirichlet model is used to estimate the cell probabilities which can help to predict the winner. We propose a time-dependent nonignorable nonresponse model for the three tables. Here, a nonignorable nonresponse model is centered on an ignorable nonresponse model to induce some flexibility and uncertainty about ignorabilty or nonignorability. As competitors we also consider two other models, an ignorable and a nonignorable nonresponse model. These latter two models assume a common stochastic process to borrow strength over time. Markov chain Monte Carlo methods are used to fit the models. We also construct a parameter that can potentially be used to predict the winner among the candidates in the November election.

Key Words: Markov chain Monte Carlo; Metropolis sampler; Multinomial-Dirichlet model; Time-dependent model; Two-way categorical table.

1. Introduction

It is a common practice to use two-way categorical tables to present survey data. Our application is to predict the winner in an election using tables constructed from a short series of polls taken before the actual election. For many surveys, there are missing data and this gives rise to partial classification of the sampled individuals. Little and Rubin (2002, section 1.3) give definitions of the three missing data mechanism (missing completely at random - MCAR, missing at random - MAR, missing not at random - MNAR); ignorable models are used to analyze data from MAR and MCAR mechanisms and nonignorable models for data from MNAR mechanisms. Thus, for the two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing). One may not know how the data are missing, and a model that includes some difference between the observed data and missing data (*i.e.*, nonignorable missing data) may be preferred. For a general $r \times c$ categorical table we address the issue of estimation of the cell probabilities of the two-way table. This problem is important because, with a substantial number of undecided voters, an election prediction based on only the partially observed data may be misleading.

As in Nandram, Cox and Choi (2005) essentially there are four two-way tables, one table with all complete cases and three supplemental tables. Of the three supplemental tables, the first has only row classification (item nonresponse), the second has only column classification (item nonresponse), and the third does not have any classification (unit nonresponse). We have extended the ignorable and nonignorable nonresponse models for two-way categorical tables of Nandram, *et al.* (2005) to accommodate a third category (*i.e.*, time in a short sequence of election polls). We have extended these models even further to include a time-dependent nonignorable non-response structure. The inclusion of the time-dependent structure can provide a more efficient prediction. A Bayesian method permits modeling different patterns of missingness under the ignorability and nonignorability assumptions, and a time-dependent nonignorable non-response model is obtained.

Our application is in Ohio governor's election, and there are several related problems. The sampled persons are categorized by two types of attributes and the cells of such categorical tables are analyzed. However, only partial classification of the individuals is available because some individuals are classified by at most one attribute, and others are left unclassified. Specifically, we use tabular data from the Ohio polls to study the relation between a measure of voters' status (likely to vote and unlikely to vote) and candidate preference (Fisher, Taft and other) to illustrate our methodology. It is interesting that voters' status is related to candidate preference. Also, it is desirable to make an adjustment for undecided voters because the proportion of undecided voters is usually high, and they often decide the final outcome of an election.

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester MA 01609. E-mail: balnan@wpi.edu.; Jai Won Choi, Department of Biostatistics, Medical College of Georgia, 1420 15th Street, Augusta, GA 30912. E-mail: jchoi@MCG.edu.

We do not know whether an ignorable nonresponse model or a nonignorable nonresponse model is appropriate, but one may have uncertainty about the ignorability of undecided voters in election polls. Referring to the Buckeye State Poll, Chen and Stasny (2003) stated that "The assumption of nonignorability of the nonresponse may be a reasonable assumption in this study because people might be reluctant to express their preference for an unpopular candidate, or if their current preferences are not firm or accurate enough for the standards of the interview." They also said that while Chang and Krosnick (2001) use ignorable models for their analyses, Chang and Krosnick (2001) suggested that nonresponse might be related to the unobserved data itself. Chen and Stasny (2003) fit three ignorable nonresponse models (A, B and C) and one nonignorable nonresponse model (D). We compare our results with theirs.

Nandram and Choi (2002 a, b) use an expansion model to study nonignorable nonresponse binary data. The expansion model, a nonignorable nonresponse model, degenerates into an ignorable nonresponse model (in the spirit of Draper 1995). This degeneracy occurs when a parameter in the nonignorable nonresponse model is set to a certain value; a good description of the centering idea is given in Nandram, *et al.* (2005, section 1.2). Because it is difficult to carry out this procedure as described, we use an alternative procedure as in Nandram, *et al.* (2005). This permits an expression of uncertainty about ignorability. This is the idea of centering a nonignorable nonresponse model on an ignorable nonresponse model, and we have used it in several of our papers to express uncertainty about ignorability or nonignorability. Here, for nonignorable nonresponse we attempt a related methodology, but the issues for a two-way categorical table are more complex, especially when a third category (*i.e.*, time) is included in these tables.

Using the approach of Chen and Fienberg (1974), Chen and Stasny (2003) describe the two issues we are discussing in this paper. For the two-way categorical tables they can handle item nonresponse only; unit nonresponse is excluded from their analysis. However, they assume that the data are missing at random and show how to obtain maximum likelihood estimators under their model. They also use a nonignorable nonresponse model (D), which they claim is their best model. It is noted in Little and Rubin (2002, chapter 15) that one issue of the nonignorable nonresponse model for this problem is that there are too many parameters, and many parameters are not identified, so they attempted a correction using hierarchical log-linear models. See Nandram, *et al.* (2005) for the case in which there are three supplemental tables.

Our methodology differs from those of Chen and Stasny (2003). The major difference is that we use a Bayesian approach. This permits us to use a method that does not rely on asymptotic theory, incorporate nonignorable missingness into the modeling and obtain time-dependent nonignorable model for estimating the proportion of voters for the three candidates. Looking to predict the winner more convincingly, we have also constructed a new parameter; it is relatively easy to analyze this parameter within the Bayesian paradigm. The Bayesian method permits modeling different patterns of missingness under two different assumptions (*i.e.*, ignorable and nonignorable missingness). Our idea is to start with an ignorable nonresponse model, which is then expanded into a nonignorable nonresponse model, and to the time-dependent nonignorable nonresponse model. It is worth noting that unit nonresponse is also included in our modeling which the other researchers consider as a separate problem using weighting adjustment (*e.g.*, see discussion in Kalton and Kasprzyk 1986). However, there can be nonignorability here as well, and one would need to include unit and item nonresponses simultaneously.

In this paper, our key contribution is to introduce a Bayesian method to analyze data from an $r \times c$ categorical table when there are both item and unit nonresponse, and the missing data mechanism can be nonignorable with a time-dependent structure. In Section 2, we describe the categorical data on voters' status and candidate preference with a time-dependent structure. In Section 3, we describe the methodology to obtain estimates of the cell probabilities incorporating the two types of missing data, and we show how to expand an ignorable nonresponse model into a nonignorable nonresponse model and time-dependent model. We also show how to use Markov chain Monte Carlo methods to fit the nonignorable nonresponse model. In Section 4, we analyze the Ohio election data to demonstrate the versatility of our methods. Finally, Section 5 has concluding remarks.

2. Data on 1998 Ohio Polls

The Center for Survey Research (CSR) at the Ohio State University conducted the Buckeye State Poll (BSP) during the 1998 election for Senator, Governor, Attorney General, State Secretary, Treasurer and Columbus Mayor. In certain months before the election, CSR conducted pre-election surveys as part of the BSP and included additional questions to collect information related to the respondent's likelihood of voting and candidate preference. In the BSP, households are sampled using the Random Digit Dialing (RDD) method, and one adult per household is selected to be interviewed using the last birthday method (Lavrakas 1993).

It is pertinent to briefly describe the RDD method. Polling firms make extensive use RDD, and the main goal of RDD is to develop a representative sample of the overall voter population. RDD sampling assumes that a representative sample cannot be obtained using listed telephone numbers in the directory. Each telephone number has 10 digits, the first three form the area code, the next three form the prefix (colloquially called the exchange), and the last four (suffix) identify a particular subscriber or a household (one household can have more than one phone number). The area codes are geographically based and typically identify localities in a state, and the exchanges can be geographically oriented. There are ten million numbers to dial but roughly less than 25% of these are real telephone numbers. Thus time and money are wasted in dialing unused numbers. We discuss this further in Section 3.

Chen and Stasny (2003) and Chang and Krosnick (2001) analyzed data from three BSP pre-election forecasting polls. Details of each of these three BSP pre-election surveys can be found in Table 1. These BSP pre-election surveys measured respondents' candidate preferences three times (January, April and October) for the November 1998 Ohio Governor race. In addition, respondents were asked for their self-reported likelihood of voting in the upcoming election using two questions. Chang and Krosnick (2001) also used filter variables (such as registered to vote, self-reported likelihood of voting, and voted in the last major election, *etc*.) to obtain those most likely to vote. Thus prediction is based only on the respondents likely to vote. Those registered to vote are classified into likely to vote, unlikely to vote and undecided. Chang and Krosnick (2001) showed that deterministic allocation of undecided respondents provide improvement in forecasting voters' candidate preferences, as compared to exclusion of all undecided respondents. Chen and Stasny (2003) used probability models to allocate the undecided voters and compared their forecasting with that of Chang and Krosnick (2001).

The data set in Chen and Stasny (2003) is slightly different because we use the undecided counts (unit nonresponse) on both variables. A voter can be undecided on at least one of the two categorical variables at each of the three polls. Chen and Stasny (2003) only study the data with undecided in exactly one variable, not both. In Table 1 for the undecided voters in both variables the counts for the January, April and October polls are respectively 5, 3 and 4; these numbers are bolded. In fact, the inclusion of these counts into our model, is an extension of the models in Chen and Stasny, and generalizes our methodology considerably.

We briefly describe the 2×3 categorical table of Ohio election data by voters' status (VS) and candidate preference (CAN). Here VS is a binary variable, and there are two levels: likely to vote and not likely to vote; CAN has three levels: Fisher, Taft, others. There are also undecided voters in VS and CAN. The bulk of the undecided voters come from voters who are "likely to vote" and "unlikely to vote" and the numbers are 173, 142 and 138 for January, April and October respectively; the undecided voters for Fisher, Taft and others are much smaller.

Table 1
Classification of October 1998 Buckeye State Poll by voting status and candidate

	Candidate				
Status	**Fisher**	**Taft**	**Other**	**Undecided**	**Total**
a. January, 1998					
Likely to vote	127	183	8	109	427
Not likely to vote	57	94	4	59	214
Undecided	0	2	0	**5**	7
Total	184	279	12	173	648
b. April, 1998					
Likely to vote	114	135	1	61	311
Not likely to vote	104	149	3	78	334
Undecided	2	6	0	**3**	11
Total	220	290	4	142	656
c. October, 1998					
Likely to vote	112	140	23	61	336
Not likely to vote	96	108	21	73	298
Undecided	7	11	1	**4**	23
Total	215	259	45	138	657

NOTE: These data are taken from Chang and Krosnick (2001); Chen and Stasny (2003) used a very similar data set; they did not use 5, 3, 4, the number of undecided voters in both variables.

In the January 1998 poll, about 73% of the voters are completely classified, 27% have no decision about candidate preference, only 1% did not know whether they would vote or not, and only five persons were completely unclassified among the 648 participants. The data set, used in our study, is presented in Table 1 as a 2×3 categorical table of voters' status and candidate preference. Our problem is to predict the winning candidate by estimating the proportion of final votes for each candidate.

The samples obtained in January, April and October are independent. There is no oversampling for a particular subpopulation or weighting of the original sample. Like many telephone surveys, RDD frame suffers from the common problem of undercoverage. As telephone coverage is not uniform over age, race, sex, income and geography, there is a need to poststratify the original sample to reduce the coverage bias by properly weighting the original data.

We perform a preliminary test of heterogeneity of the cell proportions across the three polls. Assuming a missing at random mechanism, we fill in the undecided votes. We assume that for each row (column) the undecided voters are filled in proportionally to the cell counts. Let n_{tjk} denote the adjusted cell counts with $n_t = \sum_{j=1}^{r}\sum_{k=1}^{c} n_{tjk}$, and let p_{tjk} denote the cell proportions. For a model of heterogeneous proportions, we assume that

$$\boldsymbol{n}_t \mid \boldsymbol{p}_t \overset{\text{ind}}{\sim} \text{Multinomial}(n_t, \boldsymbol{p}_t) \text{ and } \boldsymbol{p}_t \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{1}), t = 1, \ldots, T,$$

where $\mathbf{1}$ is a rc -vector of ones.

For a model of homogeneous proportions, we assume that

$$\boldsymbol{n}_t \mid \boldsymbol{p} \overset{\text{ind}}{\sim} \text{Multinomial}(n_t, \boldsymbol{p}), t = 1, ..., T, \text{ and } \boldsymbol{p} \sim \text{Dirichlet}(\mathbf{1}).$$

Then, the Bayes factor of heterogeneity versus homogeneity is

$$\text{BF} = \frac{1}{\{(rc-1)!\}^2}\left[\prod_{j=1}^{r}\prod_{k=1}^{c}\left\{\frac{\prod_{t=1}^{T} n_{tjk}!}{(\sum_{t=1}^{T} n_{tjk})!}\right\}\right]\frac{\{\sum_{t=1}^{T} n_t + rc - 1\}!}{\prod_{t=1}^{T}(n_t + rc - 1)!}.$$

Thus, using the adjusted cell counts, the logarithm of the Bayes factor (LBF) is approximately 12.4, showing very strong evidence for heterogeneity, and supporting our time-dependent model.

In a similar manner, we have computed the Bayes factors of $\boldsymbol{p}_1 = \boldsymbol{p}_2 \neq \boldsymbol{p}_3$ or $\boldsymbol{p}_1 \neq \boldsymbol{p}_2 = \boldsymbol{p}_3$ versus homogeneity; the LBFs are 7.6 and 4.4 respectively. Thus, the time-dependence occurs for both periods, January-April and April-October.

3. Methodology

We have constructed a time-dependent nonignorable nonresponse model for the 1998 Ohio Poll data. For comparison we have also considered two other models, an ignorable and a nonignorable nonresponse model. These latter two models are not time-dependent because we assume that the three time points come from the same stochastic process (*i.e.*, no correlation across time). Our main contribution is the time-dependent model. We have used the ignorable and nonignorable nonresponse models for a single time point in Nandram, *et al.* (2005). Although these two models are not appropriate in the present context, they are natural to motivate our time dependent nonignorable nonresponse model. Essentially we start with the ignorable nonresponse model which is expanded into a nonignorable nonresponse model, and we extend the nonignorable nonresponse model to a time-dependent model.

In RDD stratification and clustering are used to reduce the excess artificial numbers. Stratification by area code and some exchanges is used; geographic ordering (state or region) with systematic selection provides implicit stratification of exchanges. If an exchange is used to form a stratum, there are still ten thousand numbers to dial, still a large waste with numerous redundant numbers. The Mitofsky-Waksberg (see Waksberg 1978) procedure is a stratified two-stage cluster sampling design used to reduce the artificial numbers. Exchange areas are divided into equal size, and a random sample of exchanges is taken with replacement from those eligible (according to the measure of size of each exchange area). Within selected exchange area, a fixed number of telephone numbers is generated at random, without replacement and dialed. Thus, there is also differential probabilities of selection (*i.e.*, unequal cluster sizes) that must be considered in a comprehensive analysis. There are other variants of this procedure. RDD was adequate in 1998 Ohio election, but because of new technological innovations (*e.g.*, cellular phone, email, internet, *etc.*), the usefulness of RDD may be diminished. In this paper, our method and models do not include stratification, clustering or differential probabilities of selection.

Our models are used to estimate the proportions of voters voting for Fisher, Taft and other in the October poll. Then, assuming no catastrophic change in the November election, we predict the proportion of voters voting for Fisher, Taft and other. In this way we can predict the winner in the November election. We are excited by a referee's suggestion that one can use a mixture model to cover the possibility of a catastrophe.

In Sections 3.1 and 3.2 we describe the notations and the three models. In Section 3.3 we show how to fit the time-dependent nonignorable nonresponse model. The ignorable and nonignorable nonresponse models can be fit in a similar manner (see Nandram, *et al.* 2005 for details). In Section 3.4 we show how to specify the two parameters (μ_0 and c_0^2), and in Section 3.5 we show how to do estimation in the October poll and prediction in the November election.

3.1 Notation

Let $I_{tjk\ell} = 1$ if the ℓ^{th} voter belongs to the j^{th} row and k^{th} column of the two-way table at time t and $I_{tjk\ell} = 0$ otherwise, $t = 1, ..., T$, $j = 1, ..., r$, $k = 1, ..., c$, $\ell = 1, ..., L$. That is, $I_{tjk\ell} = 1$ denotes the cell of the $r \times c$ table that a voter belongs to. In our application $T = 3$, $r = 2$ and $c = 3$. Let $J_{ts\ell} = 1$ if the ℓ^{th} voter falls in table s ($s = 1, 2, 3, 4$) and $J_{ts\ell} = 0$ otherwise, $s = 1, ..., 4$, $\sum_{s=1}^{4} J_{ts\ell} = 1$; $J_{ts\ell}$ indicates which table an individual belongs to and $\boldsymbol{J}_{t\ell} = (J_{t1\ell}, J_{t2\ell}, J_{t3\ell}, J_{t4\ell})$.

Let the cell counts be $y_{tsjk} = \sum_{\ell=1}^{n} I_{tjk\ell} J_{ts\ell}$, $s = 1, 2, 3, 4$ for the four tables at each poll. Here y_{t1jk} are observed and y_{tsjk}, $s = 2, 3, 4$, $t = 1, ..., T$ are missing (*i.e.*, latent variables). For y_{t1jk} we know that $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t1jk} = n_{t0}$, the number of individuals with complete data. For y_{t2jk} we know that $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, where the row margins u_{tj}, $j = 1, ..., r$ are observed. For y_{t3jk} we know that $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, where the column margins v_{tk}, $k = 1, ..., c$ are observed. For y_{t4jk} we know that $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$ (unit nonresponse). In this analysis n_{t0}, $\boldsymbol{u}_t$, $\boldsymbol{v}_t$ and w_t are held fixed (*i.e.*, fixed margin analysis) and known.

Whenever it is convenient, we will use notations such as

$$\sum_{s,\,j,\,k} y_{tsjk} \equiv \sum_{s=1}^{4}\sum_{j=1}^{r}\sum_{k=1}^{c} y_{tsjk}, \quad \prod_{s,\,j,\,k} \pi_{tsjk} \equiv \prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \pi_{tsjk}$$

and $\boldsymbol{y}_{t(1)} = (\boldsymbol{y}_{t2}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$, $\boldsymbol{y}_{t(2)} = (\boldsymbol{y}_{t1}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$, *etc.*, where $\boldsymbol{y}_{ts} = (y_{tsjk},\ j = 1, ..., r, k = 1, ..., c, t = 1, ..., T, s = 1, 2, 3, 4)$. Also, we let $\boldsymbol{y}_1 = (\boldsymbol{y}_{11}, ..., \boldsymbol{y}_{T1})$ and $\boldsymbol{y}_{(1)} = (\boldsymbol{y}_{1(1)}, ..., \boldsymbol{y}_{T(1)})$ with $\boldsymbol{y}_{(1)} = (\boldsymbol{y}_{t(1)}, ..., \boldsymbol{y}_{t(4)})$. Also, $\sum_{s,\,j,\,k}^{4,\,r,\,c} y_{tsjk} = n_t$. We will also use $y_{ts\cdot\cdot} = \sum_{j,\,k} y_{tsjk}$, $y_{t\cdot jk} = \sum_s y_{tsjk}$, *etc.*, $\boldsymbol{y}_t = (\boldsymbol{y}_{t1}, \boldsymbol{y}_{t2}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$ and $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$.

3.2 Nonresponse models

Letting $\boldsymbol{I}_{t\ell} = (I_{tjk\ell}, t = 1, ..., T,\ j = 1, ..., r, k = 1, ..., c, \ell = 1, ..., L)$, for all models, we take

$$\boldsymbol{I}_{t\ell} \,|\, \boldsymbol{p}_t \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{p}_t\}, \tag{1}$$

where

$$\sum_{j=1}^{r}\sum_{j=1}^{c} p_{tjk} = 1,\ p_{tjk} \geq 0, t = 1, ..., T,\ j = 1, ..., r, k = 1, ..., c.$$

For the ignorable nonresponse model we take

$$\boldsymbol{J}_{t\ell} \mid \boldsymbol{\pi}_t \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}_t\}. \tag{2}$$

That is, there is no dependence on the cell status of an individual. For the nonignorable nonresponse models we take

$$\boldsymbol{J}_{t\ell} \,|\, \{I_{tjk\ell} = 1, I_{tj'k'\ell} = 0, j \neq j', k \neq k', \boldsymbol{\pi}_{tjk}\} \overset{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}_{tjk}\}. \tag{3}$$

Assumption (3) specifies that the probabilities an individual belongs to one of the four tables depend on the two characteristics (*i.e.*, row and column classifications) of the individual. In this manner we incorporate the assumption that the missing data is nonignorable. Note that conditional on the specified parameters in (1)-(3), one voter's behavior is correlated with another at the same time t, but there is independence over time. It is worth noting here that while the parameters in (2) are identifiable, those in (3) are not identifiable. This is where the difficulty in the nonignorable nonresponse model arises, and special attention is needed.

It follows from (1) and (2) that for the ignorable model

$$g(\boldsymbol{p}, \boldsymbol{\pi} | \boldsymbol{y}) \propto \prod_{t=1}^{T}\left[\left[\prod_{s=1}^{4} \pi_{ts}^{y_{ts\cdot\cdot}}\right]\left[\prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \frac{p_{tjk}^{y_{tsjk}}}{y_{tsjk}!}\right]\right] \tag{4}$$

subject to $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, $j = 1, ..., r$, $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, $k = 1, ..., c$, and $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$. Note that under ignorability the likelihood function in (4) separates into two pieces, one that contains the π_{ts} only and the other the p_{tjk}, and inference about these two parameters are unrelated; see Section 3.2 of Nandram, *et al.* (2005) for the original discussion of this model. Also, it follows from (1) and (3) that for the nonignorable nonresponse models the augmented likelihood function for $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \,|\, \boldsymbol{y}_1$ is

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \prod_{t=1}^{T}\left[\prod_{s,\,j,\,k}^{4,\,r,\,c} \frac{\pi_{sjk}^{y_{tsjk}}}{y_{tsjk}!}\right]\left[\prod_{j,\,k}^{r,\,c} p_{tjk}^{y_{t\cdot jk}}\right] \tag{5}$$

subject to $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, $j = 1, ..., r$, $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, $k = 1, ..., c$, and $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$; see Nandram, *et al.* (2005) for a description of identifiability in a similar situation.

For the ignorable and nonignorable nonresponse models, we take

$$\boldsymbol{p}_t \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{\mu}_2\,\tau_2),\ t = 1, ..., T+1, \tag{6}$$

where we consider prediction at $T+1$, one step ahead (November). The probabilistic structure in (6) permits a "borrowing of strength" across time. Note that the k-dimensional vector $\boldsymbol{x}$ has a Dirichlet distribution if $p(\boldsymbol{x}|\boldsymbol{\alpha}) = \prod_{j=1}^{k} x_j^{\alpha_j - 1}/D(\boldsymbol{\alpha})$, $x_j \geq 0$, $j = 1, ..., k$, $\sum_{j=1}^{k} x_j = 1$, where $D(\boldsymbol{\alpha})$ is the Dirichlet function and $\alpha_j > 0$, $j = 1, ..., k$. For a quick reference see Ghosh and Meeden (1997, pages 42, 50, 127) in connection with the Polya urn distribution, and more appropriately its use as a conjugate prior in multinomial sampling; starting with our first paper (*i.e.*, Nandram 1998) we have been using the Dirichlet-multinomial extensively in our research.

We next describe the stochastic models for the $\boldsymbol{\pi}_{tjk}$. For the ignorable nonresponse model, we take

$$\boldsymbol{\pi}_t \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\mathbf{1}),\ t = 1, ..., T, \tag{7}$$

where $\mathbf{1}$ is a four-dimensional vector of ones. We need (7) because T is small (*i.e.*, $T = 3$ in our application). Thus, we use the uniform prior in R^4 (essentially noniformative); otherwise we will have to specify the unknown parameters of the Dirichlet distribution with virtually no data. For the nonignorable nonresponse models we take

$$\boldsymbol{\pi}_{tjk} \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{\mu}_1\tau_1),\ t = 1, ..., T,\ j = 1, ..., r, k = 1, ..., c. \tag{8}$$

First, we note that (8) provides a "borrowing of strength" across time. More importantly, because $\boldsymbol{\pi}_{tjk}$ are not identifiable so are $\boldsymbol{\mu}_1$ and τ_1. One possible way out of this dilemma is to "center" the nonignorable nonresponse model on the ignorable nonresponse model.

For the time-dependent model, we take

$$\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{p}_{t-1}\,\tau_2),\ t = 1, ..., T+1, \tag{9}$$

where $\boldsymbol{p}_0$ is also unknown. Note that

$$E\{\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2\} = \boldsymbol{p}_{t-1}, t = 1, ..., T+1;$$

so that $\{p_t\}$, *a priori*, is a martingale vector. Here T is small (*i.e.*, $T = 3$). Thus, this time-dependent structure seems more appropriate, and can potentially provide improved precision. Note also that we have taken $\boldsymbol{p}_0 \sim$ Dirichlet (**1**).

Finally, we specify prior densities for the hyper-parameters. First, we take

$$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \overset{\text{iid}}{\sim} \text{Dirichlet } (\mathbf{1}), \tag{10}$$

essentially noniformative prior densities.

Finally, τ_1 and τ_2 are independent and identically distributed random variables from

$$f(x) = 1/(1+x)^2, \; x \geq 0. \tag{11}$$

Again this is an essentially noniformative prior density. Note that $\boldsymbol{\mu}_1$ and τ_1 do not exist in the ignorable nonresponse model. Gelman (2006) recommended priors like (11) instead of the ill-behaved proper diffuse gamma priors.

For the nonignorable nonresponse models we need to be more careful to specify the prior density of τ_1 because $\boldsymbol{\pi}_{tjk}$ are not identifiable. Here we attempt to "center" the nonignorable nonresponse models on the ignorable nonresponse model. In (8) the parameter τ_1 tells us about the closeness of the nonignorable model to the ignorable model. For example, if τ_1 is small, the $\boldsymbol{\pi}_{tjk}$ will be very different, and if τ_1 is large, the $\boldsymbol{\pi}_{tjk}$ will be very similar. Thus, *a priori* inference will be sensitive to the choice of τ_1, and one has to be careful in choosing τ_1. We would like to choose a prior density for τ_1 so that the nonignorable nonresponse model is kept close to the ignorable nonresponse model. Thus, we take

$$\tau_1 \sim \text{Gamma } (1/c_0^2, \; 1/\mu_0 c_0^2), \tag{12}$$

where $E(\tau_1) = \mu_0$ and $\text{CV}(\tau_1) = c_0$, with CV the coefficient of variation; both μ_0 and c_0 are to be specified. We use the prior (12) because by an appropriate choice of μ_0 and c_0 it is possible to center the nonignorable nonresponse model on the ignorable nonresponse model. Of course, one can use other convenient proper priors with parameters like μ_0 and c_0 to facilitate the centering. In Section 3.4 we will use samples from the posterior density of τ_1 under the ignorable nonresponse model to specify μ_0 and c_0.

For each of the three models, it is easy to write down the joint prior density of the parameters. For example, for the time-dependent model the joint prior density is

$$p(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2) \propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2} \times \prod_{t=1}^{T} \left\{ \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2 - 1}}{D(\boldsymbol{p}_{t-1}\tau_2)} \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\prod_{s=1}^{4} \pi_{tsjk}^{\mu_{1s}\tau_1 - 1}}{D(\boldsymbol{\mu}_1\tau_1)} \right\}, \tag{13}$$

where $D(\cdot)$ is the Dirichlet function.

3.3 Fitting the time-dependent nonignorable nonresponse model

Combining the likelihood function in (5) with the joint prior density in (13) via Bayes' theorem, the joint posterior density of the parameters $\boldsymbol{\pi}$, $\boldsymbol{p}$ $\boldsymbol{\mu}_1$, τ_1 τ_2 and the latent variables $\boldsymbol{y}_{(1)}$ is

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2} \prod_{t=1}^{T} \left[\prod_{s,j,k}^{4,r,c} \frac{\pi_{tsjk}^{y_{tsjk}}}{y_{tsjk}!} \right] \left[\prod_{j,k}^{r,c} p_{tjk}^{y_{t\cdot jk}} \right] \times \prod_{t=1}^{T} \left\{ \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2 - 1}}{D(\boldsymbol{p}_{t-1}\tau_2)} \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\prod_{s=1}^{4} \pi_{tsjk}^{\mu_{1s}\tau_1 - 1}}{D(\boldsymbol{\mu}_1\tau_1)} \right\} \tag{14}$$

subject to $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, $j = 1, ..., r$, $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, $k = 1, ..., c$, and $\sum_{j=1}^{r} \sum_{k=1}^{c} y_{t4jk} = w_t$, $t = 1, ..., T$.

The posterior density in (14) is complex, so we will use Markov chain Monte Carlo methods to fit it. However, it is easy to fit the time-dependent model using the griddy Metropolis-Hastings sampler (our terminology) as we will describe. Also, in a similar manner using the griddy Gibbs sampler (Ritter and Tanner 1992), it is easy to fit the ignorable and the nonignorable nonresponse models. We obtain a sample from the joint posterior density in order to make inference about the parameters. Specifically, we need to make inference about $\boldsymbol{p}_t$. To run the Metropolis-Hastings sampler, we need the conditional posterior density of each of the parameters given the others.

First, we consider the conditional posterior probability mass functions of $\boldsymbol{y}_{ts}, s = 2,3,4, t = 1, ..., T$ given $\boldsymbol{y}_{t(s)}$, $\boldsymbol{p}_t$, $\boldsymbol{\pi}_{tjk}$, $j = 1, ..., r$, $k = 1, ..., c$. From (14) it is clear that under the conditional posterior density the $\boldsymbol{y}_{ts}, t = 1, ..., T, s = 2,3,4,$ are independent multinomial random vectors. Specifically, letting $\boldsymbol{p} = (p_{tjk}, t = 1, ..., T, j = 1, ..., r, k = 1, ..., c)$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_{tjk}, t = 1, ..., T, j = 1 ..., r, k = 1, ..., c)$,

$$\boldsymbol{y}_{t2j} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \overset{\text{ind}}{\sim} \text{Multinomial}(u_{tj}, \boldsymbol{q}_{tj}^{(2)}) \quad j = 1, ..., r),$$

$$\boldsymbol{y}_{t3k} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \overset{\text{ind}}{\sim} \text{Multinomial}(v_{tk}, \boldsymbol{q}_{tk}^{(3)}), \; k = 1, ..., c),$$

$$\boldsymbol{y}_{t4} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \sim \text{Multinomial}(w_t, \boldsymbol{q}_t^{(4)}), \tag{15}$$

where $q_{tjk}^{(2)} = \pi_{t2jk}\, p_{tjk} / \sum_{k'=1}^{c} \pi_{t2jk'}\, p_{tjk'}$, $k = 1, ..., c$, $q_{tjk}^{(3)} = \pi_{t3jk}\, p_{tjk} / \sum_{j'=1}^{r} \pi_{t3j'k}\, p_{tj'k}$, $j = 1, ..., r$ and $q_{tjk}^{(4)} = \pi_{t4jk} p_{tjk} / \sum_{j'=1}^{r} \sum_{k'=1}^{c} \pi_{t4j'k'} p_{tj'k'}$, $j = 1, ..., r, k = 1, ..., c, t = 1, ..., T$

The conditional posterior density of $\boldsymbol{\pi}_{tjk}$ is given by

$$\boldsymbol{\pi}_{tjk} \mid \{\boldsymbol{\mu}, \tau, \boldsymbol{y}_t\} \overset{\text{ind}}{\sim} \text{Dirichlet}(y_{t1jk} + \mu_{t1}\tau_1, y_{t2jk} + \mu_{t2}\tau_1, y_{t3jk} + \mu_{t3}\tau_1, y_{t4jk} + \mu_{t4}\tau_1) \quad (16)$$

with independence over $t = 1, ..., T$, $j = 1, ..., r$, $k = 1, ..., c$.

The conditional posterior density for $\boldsymbol{p}_t$, $t = 1, ..., T$ is more difficult. We note that

$$\pi(\boldsymbol{p}_0 \mid \text{else}, \boldsymbol{y}_1) \propto \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{1jk}^{p_{0jk}\tau_2 - 1}}{D(\boldsymbol{p}_0 \tau_2)} \quad (17)$$

and

$$\pi(\boldsymbol{p}_t \mid \text{else}, \boldsymbol{y}_1) \propto \left\{ \prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{y_{t\cdot jk} + p_{t-1jk}\tau_2 - 1} \right\} \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{t+1jk}^{p_{tjk}\tau_2 - 1}}{D(\boldsymbol{p}_t \tau_2)}, t = 1, ..., T, \quad (18)$$

where "else" refers to all of the parameters in $(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2, \boldsymbol{y}_{(1)})$ excluding $\boldsymbol{p}_0$ in (17) or $\boldsymbol{p}_t$ in (18). We show how to draw samples from (17) and (18) in Appendix A.

Next, we consider the hyper-parameters. Letting $\delta_s = \prod_{t=1}^{T} \prod_{j=1}^{r} \prod_{k=1}^{c} \pi_{sjk}$, and $\boldsymbol{\pi} = (\pi_{tjk}, t = 1, ..., T, j = 1, ..., r, k = 1, ..., c)$, the joint conditional posterior density of $\boldsymbol{\mu}_1, \tau_1$ is

$$p(\boldsymbol{\mu}_1, \tau_1 \mid \boldsymbol{\pi}) \propto \frac{\prod_{s=1}^{4} \delta_s^{\mu_{1s}\tau_1}}{\{D(\boldsymbol{\mu}_1 \tau_1)\}^{rcT}} \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2},$$

where

$$\sum_{s=1}^{4} \mu_{1s} = 1, \ \mu_{1s} \geq 0, s = 1, 2, 3, 4, \tau_1 > 0.$$

We do not need to get a sample directly from $p(\boldsymbol{\mu}_1 \mid \tau_1, \boldsymbol{\pi})$. But, letting $\boldsymbol{\mu}_{1(s)}$ denote the vector of all components of $\boldsymbol{\mu}_1$ except μ_{1s}, we have

$$p(\mu_{1s} \mid \boldsymbol{\mu}_{1(s)}, \tau_1, \boldsymbol{\pi}) \propto \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}} \frac{\delta_4^{(1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1}}{\{\Gamma((1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1)\}^{rcT}},$$

$$0 \leq \mu_{1s} \leq 1 - \sum_{s'=1, s' \neq s}^{3} \mu_{1s'}, s = 1, 2, 3. \quad (19)$$

We use a grid method to draw a sample from $p(\mu_{1s} \mid \boldsymbol{\mu}_{1(s)}, \tau_1, \boldsymbol{\pi})$. We started by using 50 grids (*i.e.*, we have divided the range of μ_{1s}, $(0, 1 - \sum_{s'=1, s' \neq s}^{3} \mu_{1s'})$, into 50 intervals of equal widths) to form an approximate probability mass function of μ_{1s}, $s = 1, 2, 3$. We first draw a random variable from this probability mass function to indicate which of the 50 intervals is selected. Then, for μ_{1s} we draw a uniform random variable in this interval. This procedure is efficient because μ_{1s} is bounded, the intervals are very narrow, and it is very "cheap" to construct the discrete probability mass function for each μ_{1s}, $s = 1, 2, 3$. Finally, μ_{14} is obtained from its conditional posterior density by taking $\mu_{14} = 1 - \sum_{s=1}^{3} \mu_{1s}$.

The conditional posterior density of τ_1 is

$$p(\tau_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \left[\prod_{s=1}^{4} \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}} \right] \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2}, \ \tau_1 > 0. \quad (20)$$

To draw a random deviate from (20), we proceed in the same manner as for (19), except that we transform τ_1 from the positive half of the real line to $(0, 1)$. (It is more convenient to perform a grid approximation to a density in a bounded interval.) Thus, letting $\tau_1 = \phi/(1 - \phi)$ in (20), we have

$$p(\phi \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \frac{1}{(1-\phi)^2} \left\{ \left[\prod_{s=1}^{4} \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}} \right] \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \right\}_{\tau_1 = \frac{\phi}{1-\phi}}, 0 < \phi < 1.$$

Again, we started by using 50 intervals of equal width to draw ϕ, and the random deviate for τ_1 is $\phi/(1 - \phi)$.

Letting $p = (p_{tjk}, t = 1, ..., T, j = 1, ..., r, k = 1, ..., c)$, the conditional posterior density of τ_2 is

$$\pi(\tau_2 \mid \boldsymbol{p}) \propto \frac{1}{(1+\tau_2)^2} \prod_{t=1}^{T} \left\{ \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2 - 1}}{D(\boldsymbol{p}_{t-1}\tau_2)} \right\}, \tau_2 > 0. \quad (21)$$

A sample is obtained in a manner similar to τ_1 in (20).

We have extensive experience in using the grid approximation. However, one has to be careful in using the grid approximation for parameters close to 0 or 1 in the in the interval $[0, 1]$. One would need to use a grid approximation in an interval near the boundary; this can be obtained by trial and error in looking at the output of the sampler as it progresses. If a parameter in $[0, 1]$ is likely to be away from 0 or 1, then the grid method works fine; this is the case for the μ_{1s}'s. However, for a parameter like τ_1 (can be very large), when transformed to ϕ in the interval $[0, 1]$, ϕ can be very large (near to 1). If the transformed value is like 0.999, one needs to adjust the grid search to be in an interval containing 0.999. This has to be done by trial and error; one needs to look at the output of ϕ as the sampler progresses, and adjust the interval accordingly. For example, if 100 grid points are equally spaced in $[0, 1]$ such as 0.01, 0.02, 0.03, …. 0.99, and the parameter is likely to be around 0.999, although we draw uniformly in the selected grid interval, these grid points are not going to be very efficient.

The Metropolis-Hastings sampler is executed by drawing a random deviate from each of (15), (16), (17), (18), (19), (20) and (21) iterating the entire procedure until convergence. This is an example of the griddy Metropolis-Hastings sampler (Ritter and Tanner 1992). We obtain a sample from the posterior densities corresponding to the ignorable and nonignorable nonresponse models in a similar manner. For all models, we use a sample of $M = 1{,}000$ from the posterior densities to do estimation and prediction. We monitored the algorithm for convergence by looking at the trace plots of each parameter versus iteration order and we studied the autocorrelation coefficient. We used a griddy Gibbs sampler to fit the ignorable and nonignorable nonresponse models. We used a "burn in" of 1,000 iterates and we took every tenth thereafter. This procedure works well.

However, for the time-dependent model, we used a griddy approximation to the conditional posterior of $\boldsymbol{p}_0$, but Metropolis steps for $\boldsymbol{p}_t, t = 1,...,T$. The Metropolis steps did not work well because the jumping probabilities are 0.67, 0.65 and 0.73 for the three conditional posterior densities of $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$, but they are recommended to be between 0.25 and 0.50 (Gelman, Roberts and Gilks 1996); tuning did not help. So we used grid approximations to these three conditional posterior densities as well. The grid approximations are very accurate. In all grid approximations, we started with 50 grids, and we increased the number of grids until our estimates of all $\boldsymbol{p}_T$, $\boldsymbol{p}_{T+1}$, $\boldsymbol{\mu}_2$, τ_2 do not change. We found that 200 grids were adequate in all cases (*i.e.*, for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$). Also, we found that although the Metropolis-Hastings sampler did not work as well as we wanted, the estimates of the cell proportions are virtually the same from both samplers. The Metropolis-Hastings sampler was run for 25,000 iterations with a "burn in" of 5,000 and thinning by choosing every twentieth.

Finally, we stored the sample from the joint posterior density for further analysis. Specifically, for the ignorable and the nonignorable nonresponse models, we need the sample of size M from $\{(\boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}, \boldsymbol{p}_{T-1}^{(h)}, \boldsymbol{p}^{(h)}), h = 1,...,M\}$, and for the time-dependent model we need the sample of size M from $(\tau_2^{(h)}, \boldsymbol{p}_{T-1}^{(h)}, \boldsymbol{p}^{(h)}), h = 1,...,M$.

3.4 Specification of μ_0 and c_0^2

Finally, we describe how to specify μ_0 and c_0 in (12). This is important because it permits us to "center" the nonignorable nonresponse model on the ignorable nonresponse model (*i.e.*, an expansion model). This procedure is in the spirit of Nandram, *et al.* (2005).

We have drawn a sample of $\boldsymbol{\pi}_t^{(h)}, t = 1,...,T, h = 1,...,M$, $M = 1{,}000$ iterates from the ignorable nonresponse model, and computed $\boldsymbol{\pi}^{(h)} = \sum_{t=1}^{T} \boldsymbol{\pi}_t^{(h)}/T, h = 1,...,M$. Then, using the griddy Gibbs sampler, we fit the model

$$\boldsymbol{\pi}^{(h)} \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1\tau_1),$$

$$\boldsymbol{\mu}_1 \sim \text{Dirichlet}(\mathbf{1}),\ p(\tau_1) = 1/(1+\tau_1)^2, \tau_1 > 0,$$

with *a priori* $\boldsymbol{\mu}_1$ and τ_1 independent, to obtain a sample $\tau_1^{(h)}, h = 1,...,M$. We have drawn 1,500 iterates with a "burn in" of 500 to get $M = 1{,}000$ iterates.

Finally, taking $a = M^{-1}\sum_{h=1}^{M} \tau_1^{(h)}$ and $b = (M-1)^{-1}\sum_{h=1}^{M} (\tau_1^{(h)} - a)^2$, we set

$$c_0 = \sqrt{b}/a \text{ and } \mu_0 = a.$$

For the election data, our procedure gives $c_0 = 0.031$ and $\mu_0 = 2.431$. This specification will hold the nonignorable nonresponse model close to the ignorable nonresponse model, thereby providing a possible centering mechanism.

To study sensitivity to the misspecification of the prior density of τ_1, we use two constants, κ_1 and κ_2, such that *a priori*

$$\tau_1 \sim \text{Gamma}(1/\kappa_1^2 c_0^2,\ 1/\kappa_1^2 \kappa_2 \mu_0 c_0^2)$$

with varying values of κ_1 and κ_2. It is worth noting that $E(\tau_1) = \kappa_2\mu_0$ and $\text{CV}(\tau_1) = \kappa_1 c_0$; thus increasing κ_2 means increasing τ_1 which, in turn, means increasing precision *a priori* but not necessarily *a posteriori*. We will study the sensitivity to the specification of κ_1 and κ_2 when we describe the data analysis.

3.5 Estimation and prediction

We show how to improve estimation (*i.e.*, Rao-Blackwellization) in the October poll, and how to do prediction in the November election.

For the ignorable and nonignorable nonresponse models,

$$g(\boldsymbol{p}_T \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_T \mid \boldsymbol{\mu}_2, \tau_2)\, \pi(\boldsymbol{\mu}_2, \tau_2 \mid \boldsymbol{y}_1) d\boldsymbol{\mu}_2\, d\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g(\boldsymbol{p}_T \mid \boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}), \tag{22}$$

where $\boldsymbol{p}_T \mid \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}(\boldsymbol{\mu}_2\tau_2)$, and for the time-dependent model,

$$g(\boldsymbol{p}_T \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}, \tau_2)\, \pi(\boldsymbol{p}_{T-1}, \tau_2 \mid \boldsymbol{y}_1) d\,\boldsymbol{p}_T\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g(\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}^{(h)}, \tau_2^{(h)}), \tag{23}$$

where $\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}, \tau_2 \sim \text{Dirichlet}(\boldsymbol{p}_{T-1}\tau_2)$.

We obtain (predict) the cell proportions for November as follows. The ignorable or nonignorable nonresponse model, posterior density of $\boldsymbol{p}_{T+1}$ is

$$g(\boldsymbol{p}_{T+1} \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2, \tau_2)\, \pi(\boldsymbol{\mu}_2, \tau_2 \mid \boldsymbol{y}_1) d\boldsymbol{\mu}_2\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g(\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}), \tag{24}$$

where $\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}(\boldsymbol{\mu}_2\, \tau_2)$. For the time-dependent

$$g(\boldsymbol{p}_{T+1} \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T, \tau_2)\, \pi(\boldsymbol{p}_T, \tau_2 \mid \boldsymbol{y}_1)\, d\,\boldsymbol{p}_T\, d\,\tau_2$$

$$\approx \frac{1}{M} \sum_{h=1}^{M} g(\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T^{(h)}, \tau_2^{(h)}), \qquad (25)$$

where $\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T, \tau_2 \sim \text{Dirichlet}(\boldsymbol{p}_T\, \tau_2)$.

Thus, by (22), (23), (24) and (25), estimation and predict-tion are straight forward. For example, consider the time-dependent model. For estimation, by (24) for each h, we draw a random deviate $\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}^{(h)}, \tau_2^{(h)} \sim \text{Dirichlet}(\boldsymbol{p}_{T-1}^{(h)} \tau_2^{(h)})$, denoted by $\boldsymbol{p}_T^{(h)}, h = 1, \ldots, M$. For prediction, by (25) for each h, we draw a random deviate $\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T^{(h)}$, $\tau_2^{(h)} \sim \text{Dirichlet}(\boldsymbol{p}_T^{(h)} \tau_2^{(h)})$, denoted by $\boldsymbol{p}_T^{(h)}, h = 1, \ldots, M$. Thus, inference about $\boldsymbol{p}_T$ and $\boldsymbol{p}_{T+1}$ is made in the usual manner. The procedure is similar for the ignorable and nonignorable nonresponse models.

4. Data analysis

In this section we compare our models with those of Chen and Stasny (2003) and the actual (November election) outcomes. We have introduced a new parameter to help predict the outcome of the election. We also study extensively sensitivity of inference to choices of κ_1 and κ_2. Based on our procedure, we have specified the coefficient of variation, $c_0 = 0.031$, and the mean, $\mu_0 = 2.431$, of the prior distribution of τ_1.

In Table 2 we compare inference about the proportions of October voters allocated to the three candidates by our models and those of Chen and Stasny (2003). In this table the results are based on the prior $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0\, c_0^2)$ (*i.e.*, $\kappa_1 = \kappa_2 = 1$). We also present the actual proportions taken from Chang and Krosnick (2001). The actual proportions are (0.45, 0.50, 0.05) for Fisher, Taft and other. Using our time-dependent nonresponse model these proportions are estimated to be (0.41, 0.50, 0.09). These compare favorably with the actual outcomes. The corresponding estimates are (0.41, 0.51, 0.08) for the ignorable nonresponse model and (0.40, 0.50, 0.09) for the nonignorable nonresponse model. The best result of Chen and Stasny (2003) is obtained from their Model D, and their estimates are (0.42, 0.51, 0.07). We have provided 95% credible intervals for our estimates, but within the approach of Chen and Stasny (2003) it is relatively more difficult to provide similar intervals. Also, in Table 2 we present estimates of the predicted proportions for the November elections. The point predictors are similar to the point estimates except for the predicted proportion going to Taft under the ignorable nonresponse model. However, as expected the 95% credible intervals for the predicted proportions are much wider. For example, under the time-dependent model 95% credible interval for the proportion voting for Taft in the October poll is (0.41, 0.60) and for prediction it is (0.21, 0.78). Thus, while the point estimates and predictions do indicate the winner, the variability indicates no difference between Taft and Fisher. We will look at this further.

Table 2
Comparison of the proportion of likely voters for the October 1998 poll and prediction for November 1998 election for different models with actual outcome

Status	Fisher	Taft	Other
Sample Estimate	0.41	0.51	0.08
Approximate 95% CI	(0.35, 0.47)	(0.45, 0.57)	(0.05, 0.11)
Actual Outcome	0.45	0.50	0.05
a. Estimation			
Chen/Stasny models A,B,C	0.41	0.51	0.08
Chen/Stasny model D	0.42	0.51	0.07
Chen/Stasny model E	0.41	0.51	0.08
Ignorable model	0.41	0.51	0.08
95% CI	(0.35, 0.46)	(0.46, 0.57)	(0.05, 0.12)
Nonignorable model	0.41	0.50	0.09
95% CI	(0.32, 0.51)	(0.40, 0.60)	(0.05, 0.17)
Time-dependent model	0.41	0.50	0.09
95% CI	(0.32, 0.52)	(0.41, 0.60)	(0.05, 0.16)
b. Prediction			
Ignorable model	0.41	0.54	0.05
95% CI	(0.15, 0.70)	(0.25, 0.81)	(0.00, 0.22)
Nonignorable model	0.42	0.52	0.06
95% CI	(0.15, 0.70)	(0.22, 0.79)	(0.00, 0.28)
Time-dependent model	0.41	0.50	0.09
95% CI	(0.15, 0.71)	(0.21, 0.78)	(0.00, 0.31)

NOTE: $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0\, c_0^2)$, where $c_0 = 0.031$ and $\mu_0 = 2.431$.

Although our estimates from the time-dependent model are close to the actual estimates, the 95% credible intervals for p_{311} and p_{312} overlap, thereby making it difficult to predict Taft is the winner. Although the 95% credible intervals for our other models are shorter, the point estimates are not so good and they still overlap. One weakness in our analysis in Table 2 is that we have ignored the correlation between the two estimates (*i.e.*, we should really study the difference $p_{312} - p_{311}$, the margin of winning).

In Table 3 we present estimates of $\Lambda_e = p_{312} - p_{311}$ and $\Lambda_p = p_{412} - p_{411}$ at $\kappa_1 = \kappa_2 = 1$ for the three models. We have also included the numerical standard error (NSE) which is a measure of how well the numerical results can be reproduced; we have used the batch-means method to compute it. Small NSEs mean that if we repeat the entire computation the same way (*i.e.*, using another 1,000 iterates), we should see very little difference between the two sets of answers. In Table 3 the NSEs are small. The point estimators and predictors are all positive showing that Taft is the winner in both the October poll and the November election. However, the variability dwarfs this

result somewhat because the PSD are large, as expected even more so for prediction. This causes the 95% credible intervals for both parameters to contain 0. Thus, again when variability is considered, there is no difference between Taft and Fisher.

Table 3
Comparison of the three models for estimation and prediction using the posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals for Λ_e (Λ_p) and Δ_e (Δ_p)

	Model	PM	PSD	NSE	Interval
Λ_e	Ignorable	0.105	0.055	0.002	(-0.002, 0.209)
	Nonignorable	0.097	0.099	0.006	(-0.100, 0.280)
	Time-dependent	0.093	0.101	0.007	(-0.098, 0.276)
Λ_p	Ignorable	0.071	0.154	0.004	(-0.240, 0.362)
	Nonignorable	0.058	0.150	0.005	(-0.252, 0.369)
	Time-dependent	0.050	0.134	0.005	(-0.244, 0.314)
Δ_e	Ignorable	0.688	0.175	0.008	(0.295, 0.958)
	Nonignorable	0.663	0.200	0.012	(0.222, 0.959)
	Time-dependent	0.632	0.148	0.014	(0.336, 0.901)
Δ_p	Ignorable	0.688	0.175	0.008	(0.295, 0.960)
	Nonignorable	0.663	0.193	0.009	(0.253, 0.972)
	Time-dependent	0.648	0.155	0.011	(0.341, 0.923)

NOTE: See note to Table 2; $\Lambda_e = p_{312} - p_{311}$ (estimation, difference between Taft and Fisher for the October poll); $\Lambda_p = p_{412} - p_{411}$ (prediction, difference between Taft and Fisher for the November election); $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \boldsymbol{\alpha})$; and $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \boldsymbol{\alpha})$; see (26).

We seek an alternative parameter looking to help us predict the winner more convincingly. We pose the following question: "What is the probability that the proportion of Taft's voters in the October poll and the November election is larger than that of Fisher's voters?"

Thus, we consider the parameter $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \boldsymbol{\alpha})$ where $\alpha_{jk} = \mu_{jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, for the ignorable and nonignorable nonresponse models, and $\alpha_{jk} = p_{2jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, for the time-dependent model. In either case, letting $q_1 = p_{311}/p_{31\cdot}$, $q_2 = p_{312}/p_{31\cdot}$, and $q_3 = p_{313}/p_{31\cdot}$ with $p_{31\cdot} = \sum_{k=1}^{3} p_{31k}$ and $\sum_{k=1}^{3} q_k = 1$, it is easy to show that $(q_1, q_2, q_3) \sim \text{Dirichlet}(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)$, where $\tilde{\alpha}_1 = \alpha_{11}$, $\tilde{\alpha}_2 = \alpha_{12}$ and $\tilde{\alpha}_3 = \alpha_{13} + \sum_{k=1}^{c} \alpha_{2k}$. Therefore, we have

$$\Delta_e = \Pr(q_2 > q_1 | \boldsymbol{\alpha}) = \int_0^{1/2} \left\{ \int_{q_1}^{1-q_1} \frac{q_1^{\tilde{\alpha}_1^{-1}} q_2^{\tilde{\alpha}_2^{-1}} (1 - q_1 - q_2)^{\tilde{\alpha}_3^{-1}}}{D(\tilde{\alpha}_1, \tilde{\alpha}_2\ \tilde{\alpha}_3)} dq_2 \right\} dq_1.$$

Then, it is easy to show that

$$\Delta_e = 1 - F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(1/2) \int_0^{1/2} F_{\tilde{\alpha}_2, \tilde{\alpha}_3}\{q_1/(1-q_1)\} \left\{ \frac{q_1^{\tilde{\alpha}_1 - 1}(1-q_1)^{\tilde{\alpha}_2 + \tilde{\alpha}_3 - 1}}{F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(1/2) B(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)} \right\} dq_1, \quad (26)$$

where

$$F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_1 - 1}(1-x)^{\tilde{\alpha}_2 + \tilde{\alpha}_3 - 1}}{B(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)} dx$$

and

$$F_{\tilde{\alpha}_2, \tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_2 - 1}(1-x)^{\tilde{\alpha}_3 - 1}}{B(\tilde{\alpha}_2, \tilde{\alpha}_3)} dx.$$

We note that Δ_e is the probability that Talft received a higher proportion of the votes in the October poll, and Δ_p is the probability that Taft received a higher proportion of the votes in the November election. These parameters can be very useful for estimation (e) and prediction (p). Parameters like Δ_e or Δ_p are difficult to analyze in the non-Bayesian approach such as that of Chen and Stasny (2003); indeed this is a great strength of the Bayesian paradigm.

It is easy to compute (26) using Monte Carlo integration. For each $\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3$, Δ_e $q_1 \sim \text{Beta}(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)$ truncated to the $(0, 1/2)$ is used as an importance function. Thus, for each $\tilde{\boldsymbol{\alpha}}^{(h)}$, $h = 1, ..., M$, $M = 1{,}000$ from the Metropolis-Hastings sampler (or Gibbs sampler), we can compute $\Delta_e^{(h)}$. *A posteriori* inference about Δ_e is obtained in the standard empirical manner. For prediction, we have also considered $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \boldsymbol{\alpha})$, where $\alpha_{jk} = \mu_{jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, for the ignorable and nonignorable nonresponse models, and $\alpha_{jk} = p_{3jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, for the time-dependent model. Note that Δ_e and Δ_p are the same for the ignorable and nonignorable nonresponse models.

In Table 3 we also present estimates of Δ_e and Λ_p for the three models. First, note again that the NSEs are all small. The estimates of these parameters are similar for the three models, and larger than 0.60, but the 95% credible intervals contain 0.5. Thus, again the posterior means indicate that Taft is the winner, but variation is nullifying the effect of Taft being the winner. We note again that the time-dependent model provides sharper inference, not enough though. The parameters Δ_e and Δ_p are more sensible because they restrict inference to a smaller region by conditioning on $p_{311} + p_{312} + p_{313}$ and $p_{411} + p_{412} + p_{413}$, and from a probabilistic view these parameters are more appropriate.

Finally, we study sensitivity to inference about Λ_p and Δ_p for the nonignorable nonresponse model and the time-dependent model. We do not present results for Λ_e and Δ_e because they are similar to Λ_p and Δ_p. Also, we have dropped the ignorable model as well, and we do not present 95% credible intervals because the posterior densities are roughly symmetric. Our results are presented in Table 4 by model, κ_1 and κ_2. The posterior means of Λ_p and Δ_p are respectively very similar for different values of κ_1 and κ_2. Note that *a priori*

Table 4
Sensitivity of the posterior means (PM) and the posterior standard deviations (PSD) of Λ_p and Δ_p with respect to changes in κ_1 and κ_2 by model

		κ_2							
		1		5		25		50	
Model	κ_1	PM	PSD	PM	PSD	PM	PSD	PM	PSD
a. Λ_p									
Nonignorable	1	0.058	0.150	0.046	0.153	0.060	0.148	0.054	0.147
	2	0.051	0.153	0.046	0.146	0.062	0.151	0.054	0.145
	3	0.058	0.152	0.059	0.145	0.053	0.149	0.055	0.149
	4	0.055	0.151	0.057	0.148	0.049	0.148	0.043	0.154
Time-dependent	1	0.050	0.134	0.044	0.144	0.048	0.136	0.050	0.130
	2	0.049	0.136	0.052	0.140	0.056	0.129	0.047	0.137
	3	0.039	0.139	0.049	0.137	0.045	0.139	0.052	0.133
	4	0.037	0.138	0.042	0.138	0.041	0.141	0.051	0.129
b. Δ_p									
Nonignorable	1	0.663	0.200	0.650	0.194	0.666	0.186	0.670	0.182
	2	0.663	0.197	0.661	0.188	0.667	0.185	0.659	0.181
	3	0.663	0.199	0.647	0.196	0.666	0.184	0.669	0.180
	4	0.641	0.202	0.668	0.191	0.643	0.197	0.650	0.195
Time-dependent	1	0.648	0.155	0.642	0.123	0.657	0.099	0.661	0.095
	2	0.660	0.151	0.652	0.127	0.659	0.102	0.657	0.099
	3	0.622	0.153	0.636	0.137	0.649	0.120	0.648	0.115
	4	0.610	0.162	0.636	0.152	0.646	0.132	0.644	0.127

NOTE: We have taken $\tau_1 \sim \text{Gamma}\,(1/\kappa_1^2 c_0^2,\ 1/\kappa_2\mu_0\kappa_1^2 c_0^2)$ and we studied sensitivity with respect to κ_1 and κ_2. See note to Table 3.

$$\tau_1 \sim \text{Gamma}\left(\frac{1}{\kappa_1^2\, c_0^2},\ \frac{1}{\kappa_2\,\mu_0\,\kappa_1^2\, c_0^2}\right),$$

$E(\tau_1) = \kappa_2\,\mu_0$ and $\text{SD}(\tau_1) = \kappa_1\,\kappa_2\,c_0\,\mu_0$; so clearly, *a priori* $E(\tau_1)$ increases with κ_2 and $\text{SD}(\tau_1)$ increases with either κ_1 or κ_2, but not necessarily *a posteriori*. These changes do not have a lot of effect on inference *a posteriori*. For almost all combinations of κ_1 and κ_2, under the time-dependent model posterior standard deviations of Λ_p are smaller (but not substantially) than under the nonignorable nonresponse model. Under the time-dependent model posterior standard deviations of Δ_p are substantially smaller than under the nonignorable nonresponse model for all combinations of κ_1 and κ_2.

Concluding Remarks

The main contribution in this paper is the construction and analysis of a time-dependent nonignorable nonresponse model and its application to the Ohio polling data. We have done two additional things as well. First, we have compared the time-dependent model with an extended version (to include time) of the ignorable and nonignorable nonresponse models of Nandram, *et al.* (2005). Second, we have constructed a new parameter to help predict the winner; however, this parameter did not make an enormous difference partly because there are only three time points in the time-dependent model.

Our time-dependent model provides posterior inferences that are closer to the truth than the ignorable and nonignorable nonresponse models as well as those of Chen and Stasny (2003). It is natural for voters' preference to change as new information, detrimental or supportive, is revealed into the public place. Thus, our time-dependent model, which takes care of changes over time and provides improved precision, is to be preferred. The uncertainty in the prediction can be reduced in two ways. First, with an increased number of polls there will be increased precision in the parameters, which in turn, can lead to improved prediction. Second, with more prior information (*e.g.*, exit polling) about the November election, one can also improve the prediction.

Our 95% credible intervals can be shortened by using prior information on the proportion of voters going to Taft or Fisher. A referee suggested, "The major-party voting proportions are between 35% and 65% in general elections, and in specific states an objective political scientist could generally provide an even tighter prior." However, this is a complex problem because with truncated prior distributions on the $\boldsymbol{p}$s, there is a normalization constant which is a function of τ_2. Thus, when τ_2 is drawn from its conditional posterior density, we need to perform a Monte Carlo integration to compute the normalization constant at each iterate. While this will be a useful contribution, we prefer not to pursue this problem here.

The number of days to an election has an important impact on poll accuracy and that this effect can vary substantially across different campaign contexts (*e.g.*, DeSart and Holbrook 2003). Thus, it is really difficult to predict the outcome of an election weeks before it actually occurs, unless there exists an absolute margin. Someone who wishes to predict the outcome of an election must take into consideration additional information near the actual election. Our prediction assumes that there is no catastrophic change near the election; such an abrupt change in public opinion can occur. For example, in 1988 Dukakis lost the election against George Bush for various reasons: he spent the last week in Massachusetts, his cold personality, and Bush's attack on his liberal position. Also, an effective campaign can mobilize undecided voters near the election (*e.g.*, Truman and Dewey in 1948). One way to capture a possible catastrophe is to use mixture distributions or other heavy-tailed distributions (as researchers use Levy distributions in mathematical finance).

Acknowledgement

This work was done while Balgobin Nandram was on sabbatical leave at the National Center for Health Statistics, Hyattsville, Maryland, 2003-2004.

Appendix A

Time-dependent model: Conditional posterior densities of $\boldsymbol{p}_t, t = 0,\ldots,T$

We show how to draw a sample from the conditional posterior density of $\boldsymbol{p}_0$ in (17) using a grid method, and how to draw a sample from the conditional posterior densities of $\boldsymbol{p}_t, t = 1, ..., T$ in (18) using Metropolis steps, each with an independence chain.

First, we show how to draw a sample from the conditional posterior density of $\boldsymbol{p}_0$ in (17) using a grid method. Letting $(q_{01}, ..., q_{0L}) = (p_{01}, ..., p_{0rc})$ and $(q_{11}, ..., q_{1L}) = (p_{11}, ..., p_{1rc})$ where $L = rc$, with $\sum_{\ell=1}^{L-1} q_{0\ell} \le 1$, we have

$$\pi(q_{01},\ldots,q_{0L-1} \mid \text{else}, \boldsymbol{y}_1)$$

$$\propto \frac{q_{1L}^{(1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2-1}}{\Gamma((1-\sum_{\ell=1}^{L-1} q_{0\ell})\,\tau_2)} \prod_{\ell=1}^{L-1} \frac{q_{1\ell}^{q_{0\ell}\tau_2-1}}{\Gamma(q_{0\ell}\tau_2)}, 0 \le q_{0\ell} \le 1,\ \ell = 1,\ldots,L-1,$$

and it is easy to show that

$$\pi(q_{0\ell} \mid \text{else}, \boldsymbol{y}_1) \propto \frac{q_{1L}^{(1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2-1}}{\Gamma((1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2)} \frac{q_{1\ell}^{q_{0\ell}\tau_2-1}}{\Gamma(q_{0\ell}\,\tau_2)},$$

$$0 \le q_{0\ell} \le 1 - \sum_{\ell'=1,\ \ell' \ne \ell}^{L-1} q_{0\ell'},\ \ell = 1,\ldots,L-1.$$

For each ℓ we divide the range $0 \le q_{0\ell} \le 1 - \sum_{\ell'=1,\ \ell' \ne \ell}^{L-1} q_{0\ell'}$ into a number of subintervals. To obtain a random deviate $q_{0\ell}$ from its conditional posterior density, we select an interval proportional to its area, and draw a uniform random deviate from this interval.

Second, we show how to draw a sample from the conditional posterior densities of $\boldsymbol{p}_t, t = 1, ..., T$ in (18) using Metropolis steps, each with an independence chain. Consider $\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2, \boldsymbol{y}, t = 1,\ldots,T$. We use the candidate generating density

$$\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2, \boldsymbol{y} \sim \text{Dirichlet}(\boldsymbol{a}_t),$$

where

$$a_{tjk} = y_{t\cdot jk} + \tau_2 p_{t-1jk},\ t = 1,\ldots,T,\ j = 1,\ldots,r,\ k = 1,\ldots,c.$$

Then, the acceptance probability is $A_{s,s+1} = \min(1, \psi_{s+1}/\psi_s)$ where

$$\psi_s = \prod_{j=1}^{r} \prod_{k=1}^{c} p_{tijk}^{(s)\tau_2-1} / D(\boldsymbol{p}_t^{(s)}\tau_2).$$

References

Chang, L.C., and Krosnick, J.A. (2001). Improving election forecasting. *Technical Report*, Department of Psychology, The Ohio State University.

Chen, Q.L., and Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.

Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

DeSart, J., and Holbrook, T.M. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, Series B, 57, 45-97.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., Roberts, G.O. and Gilks, W.R. (1996), Efficient Metropolis jumping rules. In *Bayesian Statistics*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), Oxford, U.K.: Oxford University Press, 599-607.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. New York: Chapman & Hall.

Kalton, G., and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1-16.

Lavrakas, P.J. (1993). *Telephone Survey Methods*: *Sampling*, *Selection*, *and Supervision*. Newbury Park, CA: Sage Publications.

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

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

Nandram, B., and Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., and Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., Cox, L.H. and Choi, J.W. (2005). Bayesian analysis of nonignorable missing categorical data: An application to bone mineral density and family income. *Survey Methodology*, 31, 213-225.

Ritter, C., and Tanner, M.A. (1992). The Gibbs stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

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


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 51-64
Statistics Canada, Catalogue No. 12-001-X

A noninformative Bayesian approach to finite population sampling using auxiliary variables

Radu Lazar, Glen Meeden and David Nelson [1]

Abstract

In finite population sampling prior information is often available in the form of partial knowledge about an auxiliary variable, for example its mean may be known. In such cases, the ratio estimator and the regression estimator are often used for estimating the population mean of the characteristic of interest. The Polya posterior has been developed as a noninformative Bayesian approach to survey sampling. It is appropriate when little or no prior information about the population is available. Here we show that it can be extended to incorporate types of partial prior information about auxiliary variables. We will see that it typically yields procedures with good frequentist properties even in some problems where standard frequentist methods are difficult to apply.

Key Words: Sample survey; Noninformative Bayes; Auxiliary variable; Linear constraints and Polya posterior.

1. Introduction

Finite population sampling is one area of statistics where prior information is used routinely when making inferences. In most cases this prior information is not incorporated into the problem in a Bayesian manner. One reason for this is that the prior information usually does not lead, in a obvious way, to a sensible prior distribution. In the predictive approach (see Valliant, Dorfman and Royall 2000) a model is assumed and its unknown parameters are estimated after the sample has been observed. In the usual frequentist theory the prior information is encapsulated in the probability sampling plan or sample design. Basu showed that after the sample has been observed, the sampling design plays no role in the posterior distribution for a Bayesian. (For this fact and more of Basu's thoughts on finite population sampling see Ghosh (1988).) Although this radical conclusion has not been accepted by all Bayesians it is true that the usual frequentist theory and the Bayesian approach to survey sampling do not have a lot in common.

Traditional theory in survey sampling has emphasized estimation of the population mean. When the population mean of an auxiliary variable is known *a priori* the ratio estimator or the regression estimator is often employed. If one wishes to estimate something other than the mean, say a population quantile or the population distribution function, or if one has prior information about the auxiliary variable other than its mean then new methods need to be developed. Recent work along this line can be found in Chen and Qin (1993), Chen and Sitter (1999), Mak and Kuk (1993), Kuk and Mak (1989), Rao, Kovar and Mantel (1990) and Chambers and Dunstan (1986).

One advantage of a Bayesian approach is that a sensible posterior distribution for the population will incorporate the prior information into the estimation of several population parameters. Even if the posterior does not have a closed expression for a particular estimator for any given sample we can find its value approximately. This is done by sampling from the posterior distribution to simulate complete copies of the population and employing Monte Carlo estimation methods. If the posterior does not have a convenient form for sampling one should be able to use Markov Chain Monte Carlo methods to implement the simulation process. For each such simulated copy one computes the value of the parameter of interest. By simulating many such full copies of the population one can find, approximately, the corresponding Bayes point and interval estimates of the given population parameter. The problem then is to find a sensible Bayesian population model which utilizes the type of prior information available for the auxiliary variable.

Often, sensible Bayesian models can be based on the Polya posterior. The Polya posterior is a noninformative Bayesian approach to finite population sampling which uses little or no prior information about the population. A good source for more discussion on this approach is Ghosh and Meeden (1997). It is appropriate when a classical survey sampler would be willing to use simple random sampling as their sampling design. Here we show how it can be extended to cases where prior information about an auxiliary variable is present. For example the mean or median of an auxiliary variable might be known exactly or known to belong to some interval of possible values.

1. Radu Lazar, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: radu@stat.umn.edu; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: glen@stat.umn.edu; David Nelson, Center for Chronic Disease Outcomes Research, Minneapolis VA Medical Center, Minneapolis, MN 55417. E-mail: Dave.Nelson@med.va.gov.

The basic idea is to constrain or restrict the Polya posterior to put positive probability only on simulated populations which satisfy the constraints specified by the prior information for the auxiliary variables. This appropriately constrained Polya posterior can then be used to make inferences about the population parameters of interest. In Nelson and Meeden (1998) the authors considered several scenarios where a constrained Polya posterior yielded sensible frequentist results. There it was assumed that information about the population quantiles of the auxiliary variable was known *a priori*. Here we assume that we have more general forms of prior information about either the population mean or population quantiles for a set of auxiliary variables. These quantities may either be known exactly or known only to belong to some interval.

In section 2 we review the Polya posterior. In section 3 we introduce the constrained Polya posterior and discuss how to use Markov Chain Monte Carlo methods to find, approximately, the value of an estimate based on this distribution. In section 4 we apply the Polya posterior in a variety of situations and compare it to standard methods. In section 5 we discuss more formally how it relates to some standard frequentist methods. In section 6 we see that the constrained Polya posterior can used with designs other than simple random sampling. In section 7 we give a few concluding remarks. In the appendix we prove an admissibility result for the constrained Polya posterior which gives a theoretical justification for the methods presented here.

2. The Polya posterior

Consider a finite population consisting of N units. For unit i let y_i, a real number, be the unknown value of some characteristic of interest. We assume the unknown state of nature, $y = (y_1, ..., y_N)$, belongs so some known subset of N-dimensional Euclidian space. Suppose we wish to estimate some function $\gamma(y)$, of the unknown state of nature. The next step for a proper Bayesian analysis would be to specify a prior distribution over the parameter space. Then, given a sample generated by the sampling design, one would determine the posterior distribution of the unobserved members of the population conditioned on the values of the observed units in the sample. In most cases the posterior will not depend on the sampling design.

The Polya posterior can be used like a proper posterior distribution although it does not arise from a proper Bayesian model. It would be appropriate when there is little known about the population and the sample is assumed to be representative of the population. An example when it would be appropriate is when the sampling design is simple random sampling. Next, we briefly describe this distribution and outline its theoretical justification.

Given the data, the Polya posterior is a predictive joint distribution for the unobserved units in the population conditioned on the values in the sample. Given a sample we now show how to generate a set of possible values for the unobserved units from this distribution. Consider two urns where the first urn contains the n units in the sample along with their observed y values. The second urn contains the $N - n$ unsampled units. We begin by choosing one unit at random from each of the two urns. We then assign the observed y value of the unit selected from the first urn to the unit selected from the second urn and then place them both in the first urn. The urns now contain $n + 1$ and $N - n - 1$ balls respectively. This process is repeated until all the units have been moved from the second urn to the first and have been assigned a value. At each step in the process all the units in the first urn have the same probability of being selected. That is, the units which have been assigned a value are treated just like the ones that actually appeared in the sample. Once this is done, we have generated one complete realization of the population under the Polya posterior distribution. This simulated, completed copy contains the n observed values along with the $N - n$ simulated values for the unobserved members of the population. Hence, simple Polya sampling yields a predictive distribution for the unobserved given the observed. A good reference for Polya sampling is Feller (1968). The Polya posterior is related to the Bayesian bootstrap of Rubin (1981). See also Lo (1988) and Binder (1982).

This predictive distribution often generates estimators similar to standard frequentist estimators under simple random sampling. Consider, for example, estimation of the population mean. Before continuing we need a bit more notation.

Let s denote a possible sample of size $n(s)$. It is a subset of $\{1, 2, ..., N\}$, the set of labels for the finite population. If $s = \{i_1, ..., i_{n(s)}\}$ then $y_s = \{y_{i_1}, ..., y_{i_{n(s)}}\}$ is the set of observed values for y, the characteristic of interest. We let $z = (s, y_s)$ denote a typical observed sample. Then given $z = (s, y_s)$ we have

$$\bar{z}_s = \sum_{j=1}^{n(s)} y_{i_j}/n(s)$$

and

$$\mathrm{Var}(z) = \sum_{j=1}^{n(s)} (y_{i_j} - \bar{z}_s)^2/(n(s) - 1)$$

are the sample mean and sample variance. Let $\gamma_{mn}(y) = \sum_{i=1}^{N} y_i/N$ be the population mean. Under the Polya posterior distribution,

$$E(\gamma_{mn}(y) \mid z) = \bar{z}_s$$

and

$$\mathrm{Var}(\gamma_{mn}(y) \mid z) = (1 - f)\frac{\mathrm{Var}(z)}{n(s)}\frac{n(s) - 1}{n(s) + 1}$$

where $f = n(s)/N$. Note that, except for the last factor in the posterior variance, these two terms are just the sample mean and its variance under simple random sampling. The design probabilities play no explicit role in these calculations. Nonetheless, for the Polya posterior to be appropriate, in the judgment of the survey sampler, the values for the characteristic of interest for the observed and unobserved units need to be roughly exchangeable. It is in such situations that simple random sampling without replacement is used.

Under the Polya posterior the Bayesian credible interval for the population mean or point and interval estimates of other population quantities cannot always be found explicitly. In such cases it is easy to find these estimates approximately by repeatedly simulating completed copies of the population. For each simulated copy we calculate the population parameter of interest. Experience has shown that 500 to 1,000 simulated values will usually give good results. The mean of these computed values will be our point estimate and the 0.025 and 0.975 quantiles of these computed values will be our interval estimate.

Since under the Polya posterior the only y values that appear in a simulated completed copy of the population are those that appeared in the sample the Polya posterior is just a way to assign random weights, *i.e.*, probabilities, to the units in the sample. Under the Polya posterior the average weight assigned to each unit in the sample is $1/n(s)$ so, as we have seen, its estimate of the population mean is just the sample mean. It is this relationship and the Bayes like character of the Polya posterior which allows one to prove the admissibility of the sample mean for estimating the population mean under squared error loss. This suggests that inferential procedures based on the Polya posterior will tend to agree with frequentist procedures and will have good frequentist properties.

As further documentation of this point we note that recently two of the authors (Nelson and Meeden 2006) demonstrated that Bayesian credible intervals based on the Polya posterior for the population median agree asymptotically with the standard Woodruff interval (Woodruff 1952). For another example consider estimating either the mean or the total of a subpopulation or domain when a simple random sample from the entire population is used. Here the number of units in the sample which belong to the domain is a random variable. Hence the mean of the units in the sample which fall into the domain is the ratio of two random variables. This estimate is more complicated than the mean of all the units in the sample. To get an estimate of variance for this estimator the usual frequentist method conditions on the number of units in the sample that are in the domain. However when estimating the domain total this conditional argument does not work and an unconditional method is used to get an estimate of variance. See for example Cochran (1976). Recently one of the authors (Meeden 2005) showed that inferences based on the Polya posterior agree with the usual frequentist answers. Hence the Polya posterior handles both situations with one simple theory. It is important to remember that conditioning in the frequentist approach can be done under simple random sampling but for more complex designs, conditioning is not generally feasible since the conditional randomization distribution is unknown. As a final example note that the usual frequentist two stage cluster sampling procedures can be been justified from an extension of the Polya posterior (Meeden 1999).

The Polya posterior is similar in spirit to bootstrap methods for finite population sampling. Both methods use a type of exchangeability argument to generate pseudo-versions of the population. The basic idea for the bootstrap is found in Gross (1980). Suppose we have a simple random sample of size $n(s)$ from the population and suppose $N/n(s) = m$ is an integer. Given the sample we create a good guess for the population by combining m replicates of the sample. We then take repeated random samples of size $n(s)$ from this created population to study the behavior of the estimator of interest. The asymptotic properties of estimators can also be studied (see Booth, Bulter and Hall 1994 for details). This is in contrast to the Polya posterior which for a fixed sample generates complete versions of the population and examines the distribution of the parameter of interest in the population rather than properties of the estimator for the parameter. For the given population quantity of interest the properties of its estimator derive directly from this predictive distribution for the population values.

The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. The original Bayesian bootstrap applies to a random sample from an infinite population. Rubin showed that the bootstrap and Bayesian bootstrap are operationally very similar. The same type of analogy holds for the finite population setup. To study the variability of an estimator each repeatedly assigns random weights to the units in the sample. The logic for assigning the weights are different in the two cases as well as their theoretical justifications. The bootstrap has an asymptotic justification under repeated random sampling. The Polya posterior has a decision theoretic justification based on its stepwise Bayes nature (Ghosh and Meeden 1997).

Rather than generating a complete copy of the population it is often more efficient to use a well known approximation

to the Polya posterior. Assume that the sampling fraction f is small. For $j = 1, ..., n(s)$ let p_j be the proportion of units in a complete simulated copy of the entire population which take on the value y_j. Then, under the Polya posterior, $p = (p_1, ..., p_{n(s)})$ has approximately a Dirichlet distribution with a parameter vector of all ones, *i.e.*, it is uniform on the $n(s) - 1$ dimensional simplex, where $\sum_{j=1}^{n(s)} p_j = 1$. This approach will be very useful when we consider the constrained Polya posterior.

3. The constrained Polya posterior

3.1 The basic idea

In many situations, in addition to the variable of interest, y, the sampler has in hand auxiliary variables, x, for which prior information is available. For example, the population mean, μ_x, of x could be known. Given a unit in a random sample we observe its pair of values (y, x). Following our earlier notation we denote the sample by

$$z = (s, (y, x)_s) = (s, \{(y_{i_1}, x_{i_1}), ..., (y_{i_{n(s)}}, x_{i_{n(s)}})\}).$$

In this situation the regression estimator is often used when estimating the population mean. How should the Polya posterior be adjusted to take into account the fact that the population mean of x is known? The simple answer is to constrain the predictive distribution to put mass only on populations consistent with the prior information. In practice, we would only generate completed copies of the population consistent with the known prior information. To see how this can be done we consider the approximate form of the Polya posterior described at the end of the previous section.

For $j = 1, ..., n(s)$ let p_j be the proportion of units in a completed copy of the population that have the value (y_{i_j}, x_{i_j}). Rather than using the uniform distribution for $p = (p_1, ..., p_{n(s)})$ over the simplex to generate simulated copies of the population we should use the uniform distribution restricted to the subset of the simplex satisfying

$$\sum_{j=1}^{n(s)} p_j x_{i_j} = \mu_x. \qquad (1)$$

Before describing how we can generate vectors of p from this constrained Polya posterior we consider how the resulting estimator is related to the regression estimator.

Numerous simulation results (not presented here) show that the constrained Polya posterior behaves very much like the regression estimator under simple random sampling. The following simple argument shows why these two point estimates should often agree even though the Polya posterior makes no assumptions about the relationship between y and x.

Suppose in the population $y_i = a + bx_i + \varepsilon_i$ where ε_i is a random error with expectation zero. Let $\bar{X}$ be the known population mean of x. Then given a sample and p_i's satisfying $\sum_{i \in s} p_i x_i = \bar{X}$ we have

$$\begin{aligned} E\left(\sum_{i \in s} p_i y_i\right) &= aE\left(\sum_{i \in s} p_i\right) + bE\left(\sum_{i \in s} p_i x_i\right) + \sum_{i \in s} E(p_i e_i) \\ &\doteq a + b\bar{X} \\ &\doteq \bar{y}_s - \hat{b}\bar{x}_s + \hat{b}\bar{X} \\ &= \bar{y}_s + \hat{b}(\bar{X} - \bar{x}_s) \end{aligned}$$

where $\hat{b}$ is the least squares estimate of b. Here the sample values are fixed and the p_i's and e_i's are random and the expectation of the p_i's is with respect to the constrained Polya posterior. The first approximation follows since under simple random sampling we expect to see balanced samples on the average and the p_i's and e_i's to be roughly independent.

3.2 Linear constraints and the Polya posterior

Prior information involving auxiliary variables can arise in many ways. We have already discussed the case where the population mean of an auxiliary variable is known. Another case is knowing a population median. More generally one might only know that a population mean or median belongs to some interval of real numbers. Although such cases are little discussed in the usual design based literature they seem quite realistic. Another case is where a pair of auxiliary variables describe a two way table where each unit must belong to one of the cells and the population row and column totals for the numbers falling into each cell are known. Before describing the constrained Polya posterior approach to such problems we need to mention a minor technical point.

Suppose the population mean of the auxiliary variable x is known to equal $\mu(x)$. There will be samples where the value of x is less than $\mu(x)$ for each unit in the sample. In such cases it would be impossible to use the constrained Polya posterior. But as a practical matter this will hardly ever happen. We will always assume that the sample we are considering is "consistent" with the prior information. This is explained in more detail just below. In our simulation studies we always reject a sample which is not consistent and select another. Again, in most cases, the probability of having to reject a sample is very small.

Each of our examples of prior information can be represented by one or more linear equality or inequality constraints. We have seen that knowing the population mean yields one linear equality constraint. If one knows that the population mean falls in some interval this yields two

linear inequality constraints. We next develop some notation that will allow us to consider a variety of situations where prior information can be described using linear equality and inequality constraints.

We assume that in addition to the characteristic of interest y the population has a set of auxiliary variables $x^1, x^2, ..., x^m$. For unit i let

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, ..., x_i^m)$$

be the vector of values for y and the auxiliary variables. We assume that for any unit in the sample this vector of values is observed. We assume the prior information about the population can be expressed through a set of weighted linear equality and inequality constraints on the distinct auxiliary values in the population with weights corresponding to the proportions of the population taking these individual distinct values. We illustrate this issue more precisely by explaining how we translate this prior information about the population to the observed sample values so that we can construct pseudo-versions of the population consistent with the prior information.

Let s be a sample and, for $j = 1, 2, ..., n(s)$, let (y_{i_j}, x_{i_j}) be the observed values which for simplicity we assume are distinct. Let $p = (p_1, ..., p_{n(s)})$ be the proportion of units which are assigned the value (y_{i_j}, x_{i_j}) in a simulated complete copy of the population. Any linear constraint on the population values of an auxiliary variable will translate in an obvious way to a linear constraint on these simulated values. For example, if the population mean of x^1 is known to be less than or equal to some value, say b_1, then for the simulated population this becomes the constraint

$$\sum_{j=1}^{n(s)} p_j x_{i_j}^1 \le b_1.$$

If the population median of x^2 is known to be equal to b_2 then the constraint for the simulated population becomes

$$\sum_{j=1}^{n(s)} p_j w_j = 0.5$$

where $w_j = 1$ if $x_{i_j}^2 \le b_2$ and it is zero otherwise. If the population mean of x^2 is less than or equal to the population mean of x^3 then the simulated population constraint becomes

$$\sum_{j=1}^{n(s)} p_j (x_{i_j}^2 - x_{i_j}^3) \le 0.$$

Hence, given a family of population constraints based on prior information and a sample we will be able to represent the corresponding constraints on the simulated p by two systems of equations

$$A_{1,s}\, p = b_1 \tag{2}$$

$$A_{2,s}\, p \le b_2 \tag{3}$$

where $A_{1,s}$ and $A_{2,s}$ are $m_1 \times n(s)$ and $m_2 \times n(s)$ matrices and b_1 and b_2 are vectors of the appropriate dimensions. This generalizes the argument leading to equation 1.

We assume the sample is such that the subset of the simplex it defines by equations 2 and 3 is non-empty. For such a sample the asymptotic approximation to the constrained Polya posterior puts a uniform distribution over this subset of the simplex. Before addressing the issue of simulation from this distribution we note that it has a theoretical justification. It can be given a stepwise Bayes justification which guarantees that it will yield admissible procedures. Details are given in the appendix.

3.3 Computation

Let P denote the subset of the simplex which is defined by equations 2 and 3. P is a non-full dimension polytope. We would like to generate independent observations from the uniform distribution over P. Unfortunately we do not know how to do this. Instead, we use Markov chain Monte Carlo (MCMC) methods to generate dependent samples.

In particular we will use the Metropolis-Hastings algorithm which depends on using a Markov chain to generate a dependent sequence of random values for $p \in P$. The process works as follows. We begin by finding a starting point in p_0 in the relative interior of P. This is Step 1 below. Next we choose a random direction d in P. This is a bit tricky because the dimension of P is strictly less than $n(s) - 1$. This is accomplished in Steps 2 and 3 below. Next we find the line segment which is the intersection of the line passing through p_0 in direction d with P. This is Step 4 below. Next we choose a point at random from the uniform distribution over this line segment. This is the first observation in our Markov chain. We then repeat the process with this point playing the role of p_0 to get a second random point. Letting this second random point play the role of p_0 we get a third and so on. More formally our algorithm is:

Step 1. Choose an initial positive probability vector p_0 such that $A_{1,s}\, p_0 = b_1$ and $A_{2,s}\, p_0 < b_2$ and set $i = 0$.

Step 2. Generate a random direction d_i uniformly distributed over the unit sphere in R^n.

Step 3. Let d_i^* be the normalized projection of d_i onto the null space of $A_{1,s}$.

Step 4. Find the line segment $L_i = \{\alpha \in R \mid p_i + \alpha d_i^* \in P\}$ and generate α_i uniformly over the line segment.

Step 5. Set $p_{i+1} = p_i + \alpha_i d_i^*$ and $i = i + 1$ and go back to step 2.

At first glance it might not be clear what role the constraints are playing in this process. They are there however through the definition of P. The Markov chain generated in this way converges in distribution to the uniform distribution over the polytope. The convergence result of such mixing algorithms was proven by Smith (1984). If we wish to approximate the expected value of some function defined on P then the average of the function computed at the simulated values converges to its actual value. This allows one to compute point estimates of population parameters. Finding the 0.95 Bayesian credible interval approximately is more difficult.

One possibility is to run the chain for a long time; for example, we may generate 4.1 million values, throw away the first 100,000 values, and find the 0.025 and 0.975 quantiles of the remaining values. These two numbers will form our approximate 0.95 credible interval. In this manuscript we will only consider sample sizes of less than 100. For such sample sizes we have found that chains of a few million suffice.

How fast a chain mixes can depend on the constraints and the parameter being estimated. It seems to take longer to get good mixing when estimating the median than when estimating the mean. This is not surprising when one recalls that in standard bootstrap methods many more bootstrap samples are required when estimating quantiles rather than means. See for example Efron and Tibshirani (1993).

Another approach which can work well is to run the chain for a long time and then just use every m^{th} point where m is a large integer. Although this is inefficient it can give good answers when finding a 0.95 credible interval for the median.

4. Applications

In this section we show how various types of partial information about auxiliary variables can be incorporated in the estimation of the parameters when the constrained Polya posterior is employed. In many instances, the prior information used in the constrained Polya posterior estimation cannot be utilized by the standard frequentist methods.

4.1 Stratification

Stratification is a type of prior information which is commonly used in finite population sampling. We note that the usual stratified estimator can be thought of as arising from independent Polya posteriors within each stratum. Details can be found in Vardeman and Meeden (1984). When, in addition to stratification, an auxiliary variable is present a good estimate of the population mean can be found by combining the estimates obtained from the regression estimator within each stratum. For details, see Cochran (1976). If only the population mean of the auxiliary variable is known then under standard approaches it is difficult to combine this information with stratification unless a common model is assumed across strata. The constrained Polya can incorporate both types of information which can lead to improved estimates yet it does not require the common model assumption.

To demonstrate, we constructed a stratified population of size 900 consisting of three strata. The strata sizes were 300, 200 and 400. There were two auxiliary variables, say x_1 and x_2. In stratum one the $x_{1,i}$'s were a random sample from a gamma$(10,1)$ distribution and the $x_{2,i}$'s were a random sample from a gamma$(2,1)$ distribution. In the second stratum the $x_{1,i}$'s and the $x_{2,i}$'s were generated by the gamma$(15,1)$ and the gamma$(7,1)$ distributions respectively. In the third stratum the $x_{1,i}$'s and the $x_{2,i}$'s were generate by the gamma$(5,1)$ and the gamma$(3,1)$ distributions respectively. The characteristic of interest for the population was generated as follows:

stratum 1: $y_i = 1 + x_{1i}\, x_{2i} + \varepsilon_i$

stratum 2: $y_i = 3 + x_{1i} + x_{1i}\, x_{2i} + \varepsilon_i$

stratum 3: $y_i = 2 + x_{2i} + x_{1i}\, x_{2i} + \varepsilon_i$

where in stratum one the ε_i's were normal$(0,1)$, while in stratum two they were normal$(0,1.5^2)$, and in stratum three they were normal$(0,3.5^2)$. All the ε_i's were independent.

In addition to the strata sizes we assumed that the population median of x_1 and the population mean of x_2 were known. We generated 500 random samples according to our sampling plan drawing 75 units such that 25 units were in the first stratum, 20 units were in the second stratum and 35 units were in the third stratum. For each sample we computed the sample mean, the usual stratified estimate which is the sum of the sample means within each stratum adjusted for the size of all strata, the constrained Polya estimate, and the corresponding 95% confidence intervals and 0.95 credible intervals for these estimates.

The results of the simulations are given in Table 1. From the table, we see that the constrained Polya estimator on average agrees with the usual stratified estimator and is essentially unbiased. But its average absolute error is much smaller than the average errors of the other two. This is to be expected since the more information an estimator uses the better it should perform and the constrained Polya estimates are using information from the auxiliary variables that is ignored by the estimates which just use stratification. Note that the constrained Polya made no assumptions about how y and x_1 and x_2 were related. Furthermore it is not clear

how standard methods could make use of knowing the population median of x_1 and the population mean of x_2. If just information about means is available then the empirical likelihood based methods of Chen and Sitter (1999) and Zhong and Rao (2000) could be used. The results clearly show that the constrained Polya posterior is utilizing this additional information in a sensible manner.

Table 1
Simulation results for the stratified example where the median of the first auxiliary variable and the mean of the second are known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	47.978	4.821	36.44	23.09	1.000
Strataest	43.395	2.072	38.22	10.35	0.942
Polyaest	43.355	1.516	40.19	6.75	0.936

In this example the constrained Polya estimates were obtained using Markov chains of length 4,000,000 after the initial 100,000 points were discarded.

4.2 Categorical auxiliary variables

Assume that the elements of a population of known size N are associated with the elements of k categorical auxiliary variables. For simplicity, we consider $k = 2$ but the theory applies to more than two categorical variables. If one auxiliary variable takes on r distinct values and the other takes on c distinct values they allow the elements of the population to be classified into a two-way table with $r \times c$ cells. Let N_{ij} be the number of elements in the population that belong to the ij-cell, for i in {1, ..., r} and j in {1, ..., c}, then $\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij} = N$. If the N_{ij}'s are known and s is a random sample with n_{ij} elements from the ij-cell then a good estimate of the population mean is given by

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{N}\,\bar{y}_{ij}^{s},$$

where $\bar{y}_{ij}^{s}$ is the mean of the n_{ij} elements from the ij-cell in the sample. This is the usual stratified estimator where the cells in the table are consider the strata.

A harder problem is the estimation of the population mean when the counts, N_{ij}'s, are not known but the marginal counts are known. Let $N_{i\cdot} = \sum_{j=1}^{c} N_{ij}$ denote the marginal row counts, for i in {1, ..., r} and $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$ denote the marginal column counts, for j in {1, ..., c}. In such cases, one way of estimating the population mean is the frequentist procedure called calibration or raking. In this procedure, given a sample s, the estimator is given by $\sum_{k\in s} \hat{w}_k y_k$, where the $\hat{w}_k$'s are not the design weights but are new weights assigned to the units in the sample. A good set of weights needs to satisfy two conditions. The first is that the weights must preserve the known marginal counts, for example, $\sum_{k\in s(\cdot,j)} \hat{w}_k = N_{\cdot j}$ where $s(\cdot, j)$ is the portion of the sample falling in the j^{th} column of the two-way table. The second is that the weights should in some sense be close to the sampling design weights $1/\pi_k$, where $\pi_k = P(k \in s)$. Depending on the function used to measure the distance different calibration estimators can be obtained. Although this is a sensible idea, selecting the right distance measure and then getting a sensible estimate of variance for the resulting estimator has no standard frequentist answer. For details, see Deville and Särndal (1992).

The Polya posterior gives an alternative approach to this problem since the information provided by the known marginal totals determines a set of linear constraints on the random weights it assigns to the units in the sample. If there are continuous auxiliary variables for which we have prior information then additional constraints can be added. To see how this could work in practice we considered a simple example with two dichotomous variables so each unit can be classified into a cell of a 2×2 table, together with a third continuous auxiliary characteristic. We assumed four different levels of prior information.

1. The marginal counts for the 2-way table are known.
2. The marginal counts and the mean of the continuous auxiliary variable are known.
3. The marginal counts and the median of the continuous auxiliary variable are known.
4. The marginal counts are known and the mean of the continuous auxiliary variable is known to lie between two bounds. We chose the 45th and 65th quantiles of its population of values to specify these bounds.

For each case we formed a population using the following model where all the random variables are independent.

Cell 1,1 $x_i \sim \text{gamma}(8,1)$, $\varepsilon_i \sim \text{normal}(0, 7^2)$ and $y_i = 25 + 3x_i + \varepsilon_i$ for i in {1, ..., 150}.

Cell 1,2 $x_i \sim \text{gamma}(10,1)$, $\varepsilon_i \sim \text{normal}(0, 7^2)$ and $y_i = 25 + 3x_i + \varepsilon_i$ for i in {1, ..., 350}.

Cell 2,1 $x_i \sim \text{gamma}(6,1)$, $\varepsilon_i \sim \text{normal}(0, 4^2)$ and $y_i = 25 + 2x_i + \varepsilon_i$ for i in {1, ..., 250}.

Cell 2,2 $x_i \sim \text{gamma}(4,1)$, $\varepsilon_i \sim \text{normal}(0, 4^2)$ and $y_i = 25 + 2x_i + \varepsilon_i$ for i in {1, ..., 250}.

For each of the cases we generated a population and took 500 random samples of size 80 with 20 units from each cell. For each sample we computed the sample mean and the stratification estimate, assuming that the true population cell counts were known, and their corresponding 95% confidence intervals. We also computed the constrained Polya

estimates along with their 0.95 credible intervals. The constrained Polya estimates were obtained from the last 4,000,000 points of a Markov chain of size 4,100,000. The results of the simulations are given in Tables 2 through 5. The results in the tables show that the constrained Polya estimates based on known marginal counts and a known mean, median or known interval about the mean are better than the strata estimates based on known cell counts. The stratified estimates are better than the constrained Polya posterior only when the constrained Polya posterior only makes use of the known marginal counts.

Table 2
Simulation results for the categorical example when just the marginal cell counts are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.805	0.919	41.107	5.396	0.976
Strataest	44.355	0.846	42.259	4.191	0.940
Polyaest	43.909	0.896	41.863	4.197	0.922

Table 3
Simulation results for the categorical example when the marginal cell counts and the mean of the auxiliary variable are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.804	0.922	41.063	5.482	0.964
Strataest	44.399	0.862	42.272	4.256	0.942
Polyaest	44.506	0.510	43.257	2.497	0.960

Table 4
Simulation results for the categorical example when the marginal cell counts and the mean of the auxiliary variable are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.439	0.877	40.783	5.312	0.986
Strataest	43.927	0.884	41.804	4.244	0.940
Polyaest	43.784	0.785	42.032	3.640	0.920

Table 5
Simulation results for the categorical example when the marginal cell counts are assumed known and the mean of the auxiliary variable is known to lie between its known 45^{th} and 65^{th} quantiles

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.463	0.840	40.789	5.348	0.978
Strataest	43.948	0.865	41.825	4.245	0.948
Polyaest	43.519	0.829	41.555	4.029	0.938

4.3 An example

In this section we consider data from the Veterans Health Administration. In 1998 the VA Upper Midwest Health Care Network administered a functional status survey of the veteran users of the VA facilities within the network (Singh, Borowsky, Nugent, Murdoch, Zhao, Nelson, Petzel and Nichol 2005). Veterans eligible for this survey were those with any outpatient encounter or inpatient stay between October 1997 and March 1998 at any one of the five VA facilities in the network. In addition to basic demographic measures, such as age and sex, the primary component of the survey was the SF36-V (Kazis, Miller, Clark, Skinner, Lee, Rogers, Spiro, Payne, Fincke, Selim and Linzer 1998). This health-related quality of life survey instrument consists of eight sub-scales of physical functioning, role limitations due to physical problems, bodily pain, general health, energy/vitality, social functioning, role limitations due to emotional problems, and mental health. These scales are combined to form physical (PCS) and mental (MCS) component summary scores. Larger scores represent better health status. VHA administrative data measuring major comorbid conditions present in the year before the survey were also collected.

From the population of one of the five facilities we selected a subpopulation comprising all of the women and a random subset of the men to form a population of 2,500 individuals. For purposes of this example the number of comorbidities was categorized into three categories to represent measures of good, average and poor health. We then selected 200 stratified random samples of size 100 from the population. The strata sizes along with the sample sizes are given in table 6. Our sampling plan over sampled the women. Such unbalanced sampling plans can often occur in practice.

Table 6
The strata sizes along with the sample sizes for the Veterans Administration data

	Good	Average	Poor
F	353(20)	155(10)	117(10)
M	890(30)	493(20)	492(10)

We compared three different estimators of the mean PCS score for this population of 2,500; the sample mean which ignores the stratification, the usual stratified estimator which assume the strata sizes are known, and a constrained Polya posterior estimator which assumes that the marginal row and column totals of table 6 are known along with the average age of the individuals in the population. The population correlation between PCS and age is -0.22. The correlations of PCS with gender and with categorized comorbidity-based state of health are -0.13 and -0.28. From

the results in table 7 we see that the constrained Polya estimator performs about the same as the stratified estimator and both are a bit better than the sample mean. To compute the constrained Polya estimator we generated Markov Chains of length 7,000,000.

Table 7
Results for estimating *PCS* in the Veterans Administration data. The constrained Polya estimator assumes the row and column totals are know along with the average age of the individuals in the population

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	37.235	1.040	34.907	4.650	0.938
Strataest	36.648	0.925	34.322	4.651	0.948
Polyaest	36.644	0.925	34.344	4.605	0.958

5. Relation to empirical likelihood methods

In this section we review some frequentist methods for problems where constraints are involved and discuss their relationship to the constrained Polya posterior.

Chen and Qin (1993) considered an empirical likelihood approach to estimation in survey sampling when prior information about an auxiliary characteristic is available. To construct estimators after the sample has been observed the units in the sample are weighted to reflect the prior information. For example, suppose that the sample mean is less than the known population mean of the x values. Then positive weights, which sum to one, are selected for the sampled units such that the mean of the x_s values under the probability distribution given by the weights satisfies the known constraint. Although these weights can not be found explicitly they are easy to compute. When estimating the population mean of y the resulting estimator was first noted in Hartley and Rao (1968) and shown to be asymptotically equivalent to the regression estimator. If the population median of x is known then the units in the sample less than the known population median are given equal weights which sum to 0.5 and similarly for the sampled units with x values larger than the known population median. When estimating the population median the resulting estimator is one proposed by Kuk and Mak (1989).

An advantage of the constrained Polya posterior, and more generally of a Bayesian approach, is that it is straightforward to estimate many population quantities besides the mean without developing any new theory or methods. Given a simulated copy of the entire population which satisfies the constraints one just calculates the population parameter of interest. Then one uses such simulated values just as when one is estimating the mean.

To compare the Chen and Qin estimator of the population median of y with the constrained Polya posterior estimator when the population mean of x is known eight different populations were constructed. In half of the populations one would expect the regression estimator to do well in estimating the population mean while the remaining half did not satisfy the usual super-population model assumptions associated with the regression estimator. For each population 500 random samples of sizes 30 and 50 were taken, subject to satisfying the constraint that the sample contained values for x greater and less than the known mean. In all cases the two estimators using the prior information performed better than the sample median. These results were consistent with the simulation results of Chen and Qin. We calculated the average absolute error for the two estimators using the mean constraint. In each of the 16 different sets of simulations we then calculated the ratio of the constrained Polya posterior absolute error to that of the estimator of Chen and Qin. The range of these 16 values was 0.85 to 1.00 with a mean of 0.91. So in terms of absolute error, the constrained Polya posterior performed about 10% better, on average, than the estimator of Chen and Qin.

Suppose now that the population median of x is known. To simplify matters suppose that none of the actual values are equal to the population median of x. Let n_l be the number of units in the sample whose x values are less than the known population median of x. Then $n_u = n(s) - n_l$ is the number of units in the sample which are on the other side of the known median. Let $p_l = (p_1, ..., p_{n_l})$ and $p_u = (p_1, ..., p_{n_u})$ be two probability vectors. Intuitively, a sensible posterior distribution given the sample and the know population median would be for p_l and p_u to be independent Dirichlet distributions with all parameters equal to one with each of them assigned a weight of one half so that their total sum is one. It follows from the Theorem proved in the appendix that under our sampling plan these posteriors are stepwise Bayes. Note that under these posteriors the expected values of the proportions assigned to each unit in the sample are the weights assigned to the sample by Chen and Qin. This proves the admissibility of their estimator of the population median and consequently of Kuk and Mak's. Simulation results show that this constrained Polya posterior's 0.95 credible intervals cover approximately 95% of the time except in one special case. If the sample size is small and y and x are highly correlated then the medians for the simulated populations under the constrained Polya posterior do not vary enough and the resulting intervals are too short and their coverage frequency may be considerable less than 95%.

This close relation between the empirical likelihood approach and the Polya posterior is not surprising when one

notes that in the unconstrained case the sequence of priors leading to the Polya posterior can be used to prove the admissibility of the maximum likelihood estimator for the probability vector of a multinomial distribution.

6. Other sampling designs

All the simulation results presented thus far have used (stratified) simple random sampling without replacement (SRS) as the sampling design. In an earlier version of the manuscript a referee wanted to know how much the behavior of estimators based on the constrained Polya posterior depended on using this design. The answer is there is some dependence but not as much as you might initially believe.

We have seen under SRS that the constrained Polya posterior (CPP) estimator behaves much like the regression estimator (REG). Formally, the regression estimator depends only on knowing the population mean of the auxiliary variable. Its properties are usually studied under simple random sampling and the estimator of its variance is only valid for large samples.

For a general design the Horvitz-Thompson estimator (HT) is often used. It is unbiased but computing the exact inclusion probabilities can be difficult. This is true, for example, if the sampling is done with selection probability proportional to the size for an auxiliary variable x, say PPS(x). In practice one simply assumes that the inclusion probability of a unit is proportional to its value of x and the resulting estimator will be approximately unbiased.

We implemented several simulation studies comparing these three methods for estimating a population total. For brevity, we present the results of two of the studies. In these studies we constructed two populations of size 500. The auxiliary variable is the same in both populations and is a random sample from a gamma distribution with shape parameter 5 and scale parameter 1. Plots of the two populations are given in figure 1. We are not suggesting that in practice one would be likely to use the regression estimator in the second population. It is presented here simply to illustrate what can happen.

For each population we took 400 random samples of size 30 and 60 under two different sampling designs. They were PPS(x) and PPS($1/(x+5)$). We assumed that the population mean of x was known. For each sample we calculated the three estimates of the population total. The results from the first design are given in table 8. We see that CPP is the clear winner. The HT interval estimator's observed frequency of coverage is closest to the nominal level of 0.95. But the interval length is ridiculously long. This occurs because the reciprocals of the inclusion probabilities vary greatly. For the first population, which is roughly linear, REG and CPP behave similarly. However, for the second population, CPP performs better than REG. It's only shortcoming here is that it under covers with the smaller sample size. Under PPS($1/(x+5)$) the story is much the same although the difference between REG and CPP is much smaller for the second population. For example, when the sample size is 30 the average absolute error and frequency of coverage for REG is 131.9 and 0.875 compared to 124.3 and 0.908 for CPP. When the sample size is 60 the numbers for REG are 88.4 and 0.905 compared to 90.1 and 0.958 for CPP. The average length of their intervals are 384 and 560 respectively.

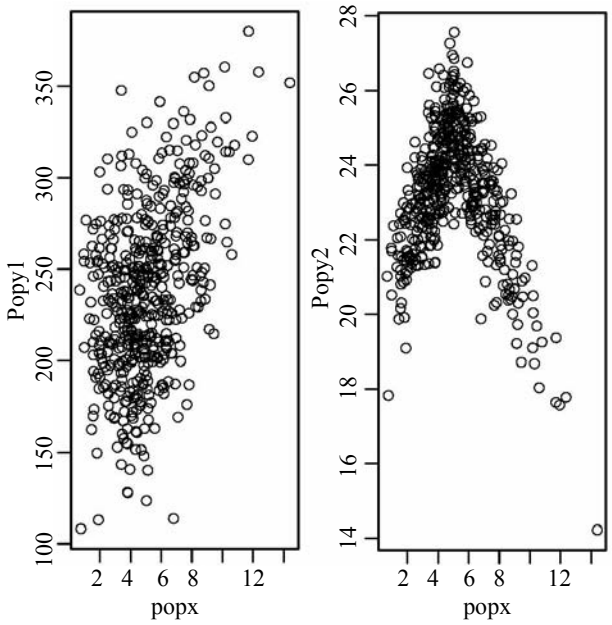

Figure 1 Plots of the two populations used in the simulations in table 8. The correlations for the two populations are 0.47 and -0.22 and their totals are 118,210 and 11,648.7

For the second population we did a second set of simulations using the PPS(x) design for sample sizes of 30 and 60. This time we assumed that the population means of x and x^2 are both known. We then compared the CPP estimator which incorporates constraints on both x and x^2 with the regression estimator which assumes a quadratic function of x as the model. These estimators are denoted by CPP2 and REG2 in the table. At first glance it might seem surprising that the results for CPP and CPP2 are essentially the same. But upon reflection it is what one should expect. The constrained Polya is simulating full copies of the population that are "balanced" with respect to x, that is agree with its known population mean. The additional constraint that a simulated copy of the population must be "balanced" with respect to x^2 as well adds little

information. On the other hand with a sensible model the regression estimator can exploit the additional information. This results in an improved point estimator but its interval estimates still under cover.

Table 8
Simulating results for the two populations in figure 1 when estimating the population total. In each case they are based on 400 samples which were select using PPS(x) as the design. Note abserr is the absolute value of the difference between the estimate and the true population total

Method	Ave. of estimates	Ave. of abserr	Ave of length	Freq. of coverage
For population 1 with total = 118,210.2				
for a sample size of 30				
HT	118,803.1	8,095.3	38,696.6	0.905
REG	116,838.1	3,355.3	14,136.4	0.905
CPP	117,515.7	3,277.3	14,330.7	0.905
for a sample size of 60				
HT	119,139.2	5,395.6	28,233.3	0.952
REG	117,041.4	2,213.2	9,561.0	0.910
CPP	118,041.4	2,195.3	11,836.5	0.938
For population 2 with total = 11,648.7				
for a sample size of 30				
HT	11,737.2	783.5	4,012.0	0.945
REG	11,800.3	179.7	533.0	0.745
CPP	11,689.9	122.4	535.4	0.900
REG2	11,660.0	97.2	382.3	0.862
CPP2	11,689.9	122.4	535.4	0.900
CPPbd	11,683.2	116.5	537.0	0.918
for a sample size of 60				
HT	11,774.2	564.8	2,908.2	0.955
REG	11,795.8	155.2	373.1	0.635
CPP	11,647.9	80.4	524.4	0.978
REG2	11,663.1	66.7	266.2	0.895
CPP2	11,651.2	88.4	523.6	0.962
CPPbd	11,644.6	83.9	552.1	0.978

For the second population we did a third set of simulations using PPS(x) as the design for sample sizes of 30 and 60. In this case we assumed that the population mean of x was contained in the interval (4.45, 5.53). These are the 0.45 and 0.65 quantiles of the x population. The mean of this population is 5.02. The results are in table 8 under the label CPPbd. We see that the results are very similar to those where the population mean of x was assumed to be known.

All three estimators are using the information contained in the auxiliary variable x but the HT estimator is the only one that depends on knowing the sampling design. As we have noted, it is well known that Bayesian estimators do not use the design probabilities in their computation. In these examples we see that CPP is making effective use of the information contained in the auxiliary variable. In general, the Polya posterior and variations on the Polya posterior, like the Constrained Polya posterior, do not rely directly upon simple random sampling, stratified random sampling, or any other design. Their suitableness and their performance are dependent upon the agreement of the structure underlying the population and the structure specified in the chosen predictive distribution.

The basic idea underlying the CPP is that one should use the sample and the available auxiliary information to simulate complete representative copies of the population. In simple examples like those given above we see that its point estimator should have excellent frequentist properties for a wide class of designs and the performance of its interval estimator will be adequate if the sample size is not to small. Does this mean that it can automatically adjust "bad" samples to get good estimates? Not really since with a very bad sample, one that agrees poorly with the known prior information, two bad things can happen. First, extremely unbalance or biased samples will introduce some bias into the point estimate. Second, they will severely constrain the possible values of p under the CPP and result in a posterior variance that is too small, which will lead to interval estimates that are too short and under cover the quantity of interest. In more complicated situations further study needs to done to discovery when the CPP can profitably be employed.

7. Final remarks

One problem with standard frequentist methods is that each different problem demands its own solution. Estimating the population median of y when the population mean of x is known is a different problem than estimating the mean of y when the mean of x is known. Also, if the population mean of x is not known exactly but is only known to belong to some interval of values then the standard frequentist methods cannot make use of this information. A strength of a Bayesian approach is that once you have a posterior distribution which sensibly combines the sample with the prior information inference can be done for many population parameters of interest simply by simulating completed copies of the population.

Here we have argued that the constrained Polya posterior is a sensible method of introducing objective prior information about auxiliary variables into a noninformative Bayesian approach to finite population sampling. The resulting point estimators have a stepwise Bayes justification which guarantees their admissibility. Their 0.95 credible intervals will usually be approximate 95% confidence intervals and they give sensible answers for problems where there are no standard frequentist procedures available. This demonstrates an important strength of the Polya posterior. Once you can simulate sensible copies of the

entire population inference for a variety of problems becomes straightforward. On the downside, one needs to use MCMC methods for their calculation. All our computations were done in R (R Development Core Team 2005). Two of the authors have recently released an R package polyapost which makes it easy for others to use our methods. Here we have restricted ourselves to samples of less than 100. This was just a matter of convenience so we could do our simulations in a reasonable amount of time. In practice for a larger specific sample one just needs to run a longer chain. Then one can use some of the standard diagnostics to decide whether or not it seems to have converged.

Appendix

An admissibility proof

The basic theoretical justification for point estimators arising from the Polya posterior is that are admissible. The proofs of admissibility use the stepwise Bayes nature of the Polya posterior. This section presents a proof for point estimators based on the constrained Polya posterior.

In these stepwise Bayes arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order is important. A different prior distribution is defined on each of the subsets. Then, the Bayes procedure is found for each sample point that receives positive probability under the first prior. Next, the Bayes procedure is found for each sample point which receives positive probability under the second prior and which was not considered under the first prior. Then, the third prior is considered and so on. For a particular sample point the value of the stepwise Bayes estimate is the value of the Bayes procedure from the step at which it was considered. It is the stepwise Bayes nature of the Polya posterior that explains its somewhat paradoxical properties. Given a sample, it behaves just like a proper Bayesian posterior but one never has to explicitly specify a prior distribution. For more details and discussion on these points see Ghosh and Meeden (1997).

To prove the admissibility of the estimators arising from the Polya posterior for the parameter space $[0, \infty)^N$ the main part of the stepwise Bayes argument first assumes that the parameter space is Λ^N, where Λ is an arbitrary finite set of positive real numbers. Once admissibility has been demonstrated for such general Λ, admissibility for the parameter space $[0, \infty)^N$ follows easily. A similar argument will be used for the constrained Polya posterior.

Dealing with constraints on finite populations introduces some technical problems which are difficult to handle. For this reason, we will suppose that the population is large enough compared to the sample size that the approximate form of the Polya posterior involving the Dirichlet distribution is appropriate. For simplicity we assume that the population U is infinite.

We assume that for all j in U, $(y_j, X_j) = a_i$ for some i in $\{1, ..., k\}$, where $a_i = (a_{i1}, ..., a_{i(m+1)})$ are distinct vectors in R^{m+1} and where k can be very large. That is, the vectors (y_j, X_j) can take on only a finite number of values. If p_i is the proportion of (y_j, X_j)'s in the population which are equal to a_i, , for i in $\{1, ..., k\}$, then the population mean of Y is $\sum_{i=1}^{k} p_i a_{i1}$.

We assume that there is prior information available about the auxiliary variables $X^i := \{x_j^i \mid j \in U\}$ for i in $\{1, ..., m\}$, which gives rise to linear equalities and inequalities involving the proportions p of the form

$$A_1 p = b_1 \tag{4}$$

$$A_2 p \leq b_2 \tag{5}$$

where A_1, A_2 are $m_1 \times k$ and $m_2 \times k$ matrices and b_1, b_2 vectors of appropriate dimensions. In this setting, for instance, we may want to estimate

$$\mu(p) = \sum_{i=1}^{k} p_i a_{i1}$$

subject to the constraints in equations 4 and 5 and where $\sum_{i=1}^{i=k} p_i = 1$ with $p_i \geq 0$, for all i in $\{1, ..., k\}$.

Consider a sample s of size n which for notational convenience we will assume consists of n distinct a_i's. Let a_s denote this set of values. We then let $A_{1,s}$ and $A_{2,s}$ be the $m_1 \times n$ and $m_2 \times n$ matrices which are just A_1 and A_2 restricted to the columns corresponding to the members of a_s. Let p_s be p restricted to the members of a_s. Then the constraints on the population given in equations 4 and 5 translate into the following constraints

$$A_{1,s} p_s = b_1 \tag{6}$$

$$A_{2,s} p_s \leq b_2 \tag{7}$$

for the random weights assigned to members of the sample. That is, given a sample the constrained Polya posterior is just the uniform distribution over the subset of the simplex defined by equations 6 and 7.

A technical difficulty when proving admissible under constraints is that even when the population satisfies the stated constraints it is always possible to get a sample which fails to satisfy them. There are several ways one can handle such cases. One possibility is to assume that the constraints are wrong and just ignore them. This tactic was used in Nelson and Meeden (1998). Another possibility is to use prior information to augment the sample so that it satisfies the constraints. This can be messy and your answer can depend strongly on how you adjust the sample. We will take a third approach here.

We will assume the sampling design is simple random sampling and that the our prior information must be correct. In such a situation it might make sense to reject any sample which does not satisfy the constraints since it is clearly an unrepresentative sample. More specifically, suppose we take a simple random sample of size n from the population and observe all x_j^i's in the sample. Let $p^s = (p_1^s, ..., p_k^s)$ be the proportions of the possible vectors for the x_j that are observed in the sample. The element p_i^s is zero whenever the vector $a_i = (a_{i2}, ..., a_{i(m+1)})$ does not appear in the sample. If p^s satisfies equations 6 and 7 we keep the sample, if not we discard it and try again. We will call this sampling plan constraint restricted random sampling. In practice, for typical constraints, it will almost never be necessary to discard a sample. Although this is a sampling plan that would never be used it is not a bad approximation to what is actually done.

More formally, let Z_i be the number of (y_j, X_j)'s in the sample that equal a_i, for i in $\{1, ..., k\}$, then $(Z_1, ..., Z_k)$ is Multinomial$(n, p_1, ..., p_k)$ where the parameter values belong to

$$P := \left\{ (p_1, ..., p_k) \middle| A_1 p = b_1, A_2 p \le b_2, \sum_{i=1}^{k} p_i = 1, \text{ and } p_i \ge 0 \forall i \in \{1, ..., k\} \right\}. \quad (8)$$

For a given sample $s = (z_1, ..., z_k)$ let

$$P^s := \{p \mid p \in P \text{ and } p_i = 0 \text{ whenever } p_i^s = 0 \text{ for } i = 1, ..., k\}. \quad (9)$$

We see that we keep a sample if and only if P^s is not empty.

Denote the $k-1$ dimensional simplex by

$$F := \left\{ (p_1, .., p_k) \middle| \sum_{i=1}^{k} p_i = 1, p_i \ge 0 \forall i \in \{1, ..., k\} \right\}.$$

For $i = 1, ..., k$ let e_i denote the vertices of F. The e_i's are the unit vectors whose i^{th} value is 1 and is 0 elsewhere.

Now P is a convex polytope which is the intersection of F with the space

$$G := \{(p_1, ..., p_k) \mid A_1 p = b_1, A_2 p \le b_2\}.$$

A partition of the parameter space P can be found in the following way. Let F_j denote the set of faces of dimension j of the simplex F, $j = 0, 1..., k-1$. Then F_0 is the set of its vertices, F_j is the collection of the convex hulls of all combinations of $j+1$ vertices, for $j = 1, ..., k-2$ and F_{k-1} is the simplex F. If $\text{int}(F_j)$ is the set of the interiors of the faces of dimension j, for $j = 1, ..., k-1$, then $\{F_0, \text{int}(F_1), ..., \text{int}(F_{k-1})\}$ determines a partition of the simplex F. If $G_0 := F_0 \bigcap G$ and $G_j := \text{int}(F_j) \bigcap G$ for $j = 1, ..., k-1$ then $\{G_0, G_1, ..., G_{k-1}\}$ is a partition of the parameter space P. Note that some of G_j's might be empty. The stages of the stepwise Bayes argument follow the nonempty members of the G_j's.

If Z is the sample space of the counts $(Z_1, ..., Z_k)$ then for $p \in P$ the distribution of the counts, say $f_P(z \mid p)$ is Multinomial$(n, p_1, ..., p_k)$ when the sample size is n. Let P_F be the restriction of the parameter space P to F, where F is any subset of P and Z_{P_F} be the restriction of the sample space Z determined by P_F.

We are now ready to prove the admissibility of the constrained Polya posterior estimator of $\mu(p)$ over P. Suppose we are at the stage where we are considering G_j, for some $j \ge 0$. Assume $G_j = G \bigcap F$ for some subset is nonempty. There are two possible cases.

Case 1. If the dimension of G_j is zero, *i.e.*, it consists of one vector, say p^0, then we take the prior that puts unit mass on this vector. The posterior also then puts unit mass on this vector and if z is the unique member of Z_{P_F} then the Bayes estimator is $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z) = \mu(p^0)$.

Case 2. If the dimension of G_j is greater than zero then the distribution of $(Z_1, ..., Z_k)$ restricted to Z_{P_F} is

$$f_{P_F}(z \mid p) = \frac{f_P(z \mid p)}{\sum_{z \in Z_{P_F}} f_P(z \mid p)}.$$

The prior we consider on P_F is

$$\pi_{P_F}(p) \frac{\sum_{z \in Z_{P_F}} f_P(z \mid p)}{\prod_{\{i \mid p \in P_F, p_i > 0\}} p_i},$$

which can be normalized to be a proper prior since $\sum_{z \in Z_{P_F}} f_P(z \mid p)$ can be written as $g(p) \prod_{\{i \mid p \in P_F p_i > 0\}} p_i$ where $g(p)$ is a bounded function of p. With this prior, the posterior distribution is the Dirichlet density kernel restricted to P_F,

$$f_{P_F}(p \mid z) \propto f_{P_F}(z \mid p) \pi_{P_F}(p) \propto \prod_{\{i \mid p \in P_F, p_i > 0\}} p_i^{z_i - 1}.$$

The Bayes estimator of $\mu(p)$ against π_{P_F}, where p belongs to P_F, is $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z)$ for all z in Z_{P_F}. Hence, if we use the sequence of priors

$$\{\{\pi_{P_F \mid F \in G_0}\}, \{\pi_{P_F \mid F \in G_1}\}, \{\pi_{P_F \mid F \in G_2}\}, ..., \{\pi_{P_F \mid F \in G_{\gamma-1}}\}\},$$

ignoring the empty sets at each step, then the estimator $\delta(z)$ defined by

$$\delta(z) = \delta_{\pi_{P_F}}(z) \text{ for } z \in Z_{P_F}, F \in G_i, i = 1, ..., \gamma - 1, \quad (10)$$

where $\gamma = k$ if $k < n$ and n if $k \ge n$, is an admissible estimator for $\mu(p)$. This concludes the proof of the following theorem.

Theorem 1. *Under the constraint restricted random sampling plan defined by equations 2 and 3 with parameter*

space defined in equation 8 the constrained Polya posterior estimator given in equation 10 for estimating the population mean is stepwise Bayes and hence admissible under squared error loss.

Acknowledgements

The research of Glen Meeden and Radu Lazar was supported in part by NSF Grant DMS 0406169. The authors would like to thank the referees and associate editor for several helpful comments.

References

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society*, Series B, 44, 388-393.

Booth, J.G., Bulter, R.W. and Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., and Sitter, S.S. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Cochran, W. (1976). *Sampling Techniques*. New York: John Wiley & Sons, Inc., 3rd Edition.

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

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

Feller, W. (1968). *An Introduction of Probability Theory and its Applications*. New York: John Wiley & Sons, Inc., Volume I.

Ghosh, J.K. (1988). *Statistical Information and Likelihood*: *A Collection of Critical Essays*. By (Ed. D. Basu). New York: Springer-Verlag.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman and Hall.

Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the Section of Section on Survey Research Methods*, American Statistical Association, 181-184.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.

Kazis, L.E., Miller, D.R., Clark, J., Skinner, K., Lee, A., Rogers, W., Spiro, A. 3rd, Payne, S., Fincke, G., Selim, A. and Linzer, M. (1998). Health related quality of life in patients served by the department of veterans affairs. *Archives of Internal Medicine*, 158, 626-632.

Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society*, Series B, 51, 261-269.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Mak, T.K., and Kuk, A. (1993). A new method for estimating finitepopulation quantiles using auxiliary information. *Canadian Journal of Statistics*, 21, 29-38.

Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā*, Series A, 61, 133-144.

Meeden, G. (2005). A noninformative bayesian approach to domain estimation. *Journal of Statistical Planning and Inference*, 129, 85-92.

Nelson, D., and Meeden, G. (1998). Using prior information about population quantiles in finite population sampling. *Sankhyā*, Series A, 60, 426-445.

Nelson, D., and Meeden, G. (2006). Noninformative nonparametric finite population quantile estimation. *Journal of Statistical planning and Inference*, 136, 53-67.

R Development Core Team (2005). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Singh, J.A., Borowsky, S.J., Nugent, S., Murdoch, M., Zhao, Y., Nelson, D., Petzel, R. and Nichol, K.L. (2005). Health related quality of life, functional impairment, and health care utilization by veterans: Veterans' quality of life study. *Journal of American Geriatric Society*, 53, 108-113.

Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32, 1296-1308.

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

Vardeman, S., and Meeden, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.

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

Zhong, B., and Rao, J.N.K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population infomation. *Biometrika*, 87, 920-938.

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 65-78
Statistics Canada, Catalogue No. 12-001-X

Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates

Alan M. Zaslavsky, Hui Zheng and John Adams [1]

Abstract

We consider optimal sampling rates in element-sampling designs when the anticipated analysis is survey-weighted linear regression and the estimands of interest are linear combinations of regression coefficients from one or more models. Methods are first developed assuming that exact design information is available in the sampling frame and then generalized to situations in which some design variables are available only as aggregates for groups of potential subjects, or from inaccurate or old data. We also consider design for estimation of combinations of coefficients from more than one model. A further generalization allows for flexible combinations of coefficients chosen to improve estimation of one effect while controlling for another. Potential applications include estimation of means for several sets of overlapping domains, or improving estimates for subpopulations such as minority races by disproportionate sampling of geographic areas. In the motivating problem of designing a survey on care received by cancer patients (the CanCORS study), potential design information included block-level census data on race/ethnicity and poverty as well as individual-level data. In one study site, an unequal-probability sampling design using the subjectss residential addresses and census data would have reduced the variance of the estimator of an income effect by 25%, or by 38% if the subjects' races were also known. With flexible weighting of the income contrasts by race, the variance of the estimator would be reduced by 26% using residential addresses alone and by 52% using addresses and races. Our methods would be useful in studies in which geographic oversampling by race-ethnicity or socioeconomic characteristics is considered, or in any study in which characteristics available in sampling frames are measured with error.

Key Words: Descriptive population quantity; Measurement error; Neyman allocation; Regression models; Sample design; Surveys.

1. Introduction

A sample survey is to be designed to obtain data that will be used to estimate coefficients of one or more regression models. Information about the population distribution of the covariates is available, and also some covariate information is available in the sampling frame. How can this information be used to make the survey design more efficient? How much can variance be reduced with such a design, relative to simple random sampling, and how is that answer affected if the frame only provides covariate distributions aggregated over groups, but not for individual subjects?

These questions were motivated by design of a survey of health care processes (such as provision of chemotherapy when appropriate) and outcomes (such as quality of life after treatment) for a large sample of cancer patients at seven sites in the United States, conducted as part of the CanCORS (Cancer Care Outcomes Research and Surveillance) study (Ayanian, Chrischilles, Wallace, Fletcher, Fouad, Kiefe, Harrington, Weeks, Kahn, Malin, Lipscomb, Potosky, Provenzale, Sandler, Vanryn and West 2004). Among the primary objectives of this study was to estimate joint effects of race and income on these measures, using regression models that include both of these patient characteristics. However, only limited data were available when patients were sampled for enrollment in the study. Prior experience suggested that race and residential address might be determined with reasonable accuracy at the time cases were ascertained for possible study recruitment, but income could not be determined until the subject was recruited and interviewed, and could not practically be collected in a screening interview. We undertook the research reported here to determine how the available patient data could be combined with census data on race-income distributions in census blocks to sample patients disproportionately and thereby improve estimates of race and income effects.

Such concerns arise frequently when survey data will be used to estimate coefficients of one or more regression models. For example, the National Health Interview Survey (NHIS) uses geographical oversampling together with a screening interview to oversample Black and Hispanic respondents for improved domain estimation (Botman, Moore, Moriarity and Parsons 2000, page 12); NHIS data have been used extensively in regression analyses, of which domain estimation is a special case. Sastry, Ghosh-Dastidar, Adams and Pebley (2005, pages 1013-1014) oversampled census tracts by minority composition, using simulations to evaluate the power of various designs for regression analyses of interest. The Youth Risk Behavior Surveillance

1. Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.; Hui Zheng, Biostatistics Center, Massachusetts General Hospital, 50 Staniford Street, Boston, MA 02114, U.S.A.; John Adams, RAND Corporation, 1776 Main Street, Santa Monica, CA 90401.

System oversamples *schools* in high-minority PSUs to improve precision of estimates for minority racial/ethnic groups (Eaton, Kann, Kinchen, Ross, Hawkins, Harris, Lowry, McManus, Chyen, Shanklin, Lim, Grunbaum and Wechsler 2006, pages 2-3).

The literature on optimal design of experiments is extensive. Design objectives for surveys, however, differ in important ways from those for experiments, in which the researcher can arbitrarily assign *a priori* identical units to treatments. A strongly model-based approach to estimation of regression coefficients would suggest selection of a suitable set of high-leverage observations, much as in design of an experiment (Royall 1970), but the application of these principles to survey design is controversial (see Sec. 4). The design-based approach requires a sample that is representative, through a known probability mechanism, of a defined population; intermediate positions are also possible (Sec. 2.5, 3.4). From this perspective, the sampler is not free to select, for example, 100 white respondents from a convenient primarily white neighborhood and 100 Hispanic respondents from a convenient primarily Hispanic neighborhood and call the sample "representative" for estimating differences between whites and Hispanics. The objects of design-based inference are quantities that describe the population; in the case of regression we will refer to "descriptive population quantity" or DPQ regressions (Pfeffermann 1993, pages 319-321).

Since Neyman (1934), the extensive literature on optimal design of surveys (reviewed in standard texts like Cochran 1977 or Särndal, Swensson and Wretman 1992) has primarily focused on estimation of simple quantities such as a mean or ratio, or of several such quantities (Kish 1974; Bellhouse 1984; Chromy 1987). Although variance estimation for design-based estimates of regression coefficients has received considerable attention (Fuller 1975; Fuller 1984; Binder 1981; Binder 1983), relatively little attention has been given to the corresponding optimal sample designs. (Regression-*assisted* estimation of a mean (Cassel, Särndal and Wretman 1976; Särndal, Swensson and Wretman 1992, Sec. 12.2) is a distinct problem.)

Furthermore, characteristics that might be used to define an unequal-probability sampling scheme are likely to be recorded with error in sampling frames, because they are based on aggregated data or because the characteristics associated with a unit (such as an address or a household) change over time. Such errors can greatly affect the efficiency of a putatively optimal sampling scheme; see Morris, Newhouse and Archibald (1979, Sec. III) on stratified sampling for domain estimation and Thomsen, Tesfu and Binder (1986) on probability-proportional-to-size sampling. Waksberg (1973, 1995) considers stratification by census blocks on a single aggregated characteristic for estimation of means for domains such as racial/ethnic groups or the poor, with or without a subsequent screening interview.

Our objective in this article is to describe optimal designs for samples that will be used in DPQ (design-weighted) regression analysis, in the sense of minimizing the weighted sum of variances of some preselected linear combinations of regression coefficients. We also consider some classes of estimands and corresponding estimators that depart from the DPQ approach to improve efficiency. In Section 2, we establish notation and derive optimal sampling rates for DPQ regression under scenarios representative of the individual and area-level information that might be encountered in population surveys with imperfect frames. We first assume that exact design information is available in the sampling frame and then generalize to situations in which some variables are available only as aggregates for subdomains or from inaccurate data. We next consider optimal estimation of combinations of coefficients from more than one model and of flexible combinations of coefficients. In Section 3 we estimate the potential benefits of these methods for a survey in the CanCORS study sites, using block-level census data on race/ethnicity and poverty. Finally, in Section 4 we consider the relevance of the DPQ approach and possible extensions of the methodology.

2. Optimal design calculations

2.1 Notation

Suppose that the target population is divided into cells indexed by $b = 1, 2, ..., B,$ with elements indexed by $k = 1, 2, ..., K_b$ in cell b. With each element is associated a covariate vector $\mathbf{x}_{bk}$ with $\mathbf{x}'_{bk} = (\mathbf{u}'_{bk}, \mathbf{t}'_{bk})$, where $\mathbf{u}_{bk}$ is the component observed for identifiable individuals. The distribution of $\mathbf{t}_{bk}$ in each cell is known but the values for individuals are not observed; thus the cell is the unit of aggregation for some or all of the design variables. Hence we know the finite population values $\mathbf{T}_b = (\mathbf{t}_{b1}, \mathbf{t}_{b2}, ..., \mathbf{t}_{bK_b})'$ but cannot identify the rows with individuals. Define $\bar{\mathbf{t}}_b = 1'\mathbf{T}_b / K_b,$ the mean of $\mathbf{t}$ in cell b.

Associated with sampling each element is a cost c_{bk}. A sampling plan is defined by assigning a probability of selection π_{bk} to each element. Assume a constraint on expected cost,

$$\sum_{b,k} c_{bk}\,\pi_{bk} \le C. \qquad (1)$$

To simplify the presentation, we also assume that the sampling rate is low and potential benefits of stratification are minimal, so the design can be described approximately as unstratified unequal-probability sampling with replacement. We also assume single-stage element sampling. The

population is $U = \{(b, k): b = 1, 2, ..., B; k = 1, 2, ..., K_b\}$ and a sample is $S \subset U$.

The population-descriptive ordinary least squares (OLS) regression coefficient, corresponding to the model $y_{bk} = \boldsymbol{\beta}'\mathbf{x}_{bk} + \varepsilon_{bk}$ with $\varepsilon_{bk} \sim [0, \sigma^2]$, is $\boldsymbol{\beta}_U = (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{X}_U' \mathbf{y}_U$, where subscript U signifies matrices or vectors corresponding to the entire population. (Here $[0, \sigma^2]$ signifies a distribution with mean 0 and variance σ^2, but unspecified form.) Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_S' \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}_S' \mathbf{W}_S \mathbf{y}_S \qquad (2)$$

is the usual design-based estimator of $\boldsymbol{\beta}$, where S signifies that only the rows corresponding to the sample are included, and $\mathbf{W}$ is the diagonal matrix of weights $1/\pi_{bk}$.

To *design* the survey, we must make some assumptions about the distribution of outcomes y_{bk}, even if we would not rely on the same assumptions in *analysis* of the data. Specifically, we assume that the outcomes are generated by a model $\xi: y_{bk} = \mathbf{x}_{bk}'\boldsymbol{\beta} + \varepsilon_{bk}$, with independent $\varepsilon_{bk} \sim [0, \sigma_{bk}^2]$ and known σ_{bk}^2 (up to a constant factor). Note that the distributions of the design variables $\mathbf{x}_{bk}$ and the residuals are relevant to optimization of the design, but the value of $\boldsymbol{\beta}$ is not since it does not affect the variance of the regression estimators. Furthermore, the assumption of independent residuals from a regression model might be more reasonable then independence of data values. We allow for heteroscedasticity, even when fitting an OLS model. OLS coefficients (including special cases such as the overall mean or domain means) are often useful descriptive statistics even if the OLS model does not actually hold, but if information about heteroscedasticity is available it can be used to make the design more efficient.

2.2 Optimal DPQ regression design with individual-level variables only

Consider first the case in which $\mathbf{t}$ is empty, so $\mathbf{x}_{bk} = \mathbf{u}_{bk}$, reflecting a scenario in which all relevant design variables (race and income in our CanCORS design) are available to the researcher before sampling. Since the cells now consist of single cases we drop the subscript b, writing $\hat{\boldsymbol{\beta}} = (\sum_S w_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_S w_k \mathbf{x}_k y_k)$. Then for any fixed linear combination of coefficients with weights $\mathbf{a}$, assuming that the first factor is a design-consistent estimator (after scaling) of $N(\mathbf{X}_U' \mathbf{X}_U)^{-1}$, we have the expectation under sampling of the model-based variance (White 1980) of the estimator,

$$\begin{aligned} V_a &= E_\pi \operatorname{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}} \\ &\approx \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(E_\pi \operatorname{Var}_\xi \sum_{k\in S} \mathbf{x}_k y_k/\pi_k\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a} \\ &= \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(\sum_{k\in U} (\sigma_k^2/\pi_k)\mathbf{x}_k \mathbf{x}_k'\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}. \end{aligned} \qquad (3)$$

For design-based inference, the relevant measure is the average variance under the sampling design over possible populations obtained under the model ξ, $E_\xi \operatorname{Var}_\pi \mathbf{a}'\hat{\boldsymbol{\beta}}$ (the "anticipated variance" of Isaki and Fuller 1982; see also Bellhouse 1984, sec. 1); this quantity is approximately equal to the expected model-based variance (see Appendix for proof and asymptotic conditions).

By the typical Lagrange multiplier argument for optimal allocation problems (*e.g.*, Valliant, Dorfman and Royall 2000, pages 169-170), V_a is minimized subject to the expected cost constraint (1) when $\partial V_a / \partial \pi_k = c_k \lambda$ for some constant λ and all k, so $\pi_k \propto \sigma_k(\mathbf{a}'\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{x}_k)/\sqrt{c_k}$. Thus the optimal sampling rate is higher for cases with greater model variance and lower cost (as in the usual case of estimation of a mean) and also for cases with greater leverage in the regression. This result differs from the standard model-based calculations for optimal experimental design, which would allocate the entire sample to a few high-leverage design points. The design-consistent estimator of the DPQ regression does not assume the correctness of the model and therefore requires that every case have a positive probability of selection. Thus, for estimation of a ratio β under a homoscedastic model $y_k = \beta x_k + \varepsilon_k$, model-based estimation would suggest selection of the units with the largest values of x, but our probabilities of selection are proportional to x.

Typically, more than one estimand will be of interest in a study; CanCORS is intended to estimate both race and income effects. We generalize (3) to simultaneous estimation of several linear combinations of coefficients by optimizing a weighted sum of variances $V = \sum_i d_i V_{\mathbf{a}_i}$, where i indexes the estimands. By the same arguments the optimal sampling probabilities for this objective are

$$\pi_k \propto \sigma_k \left(\sum_i d_i (\mathbf{a}_i'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{x}_k)^2 / c_k\right)^{1/2}. \qquad (4)$$

With some choices of the $\{\mathbf{a}_i\}$, strict adherence to (4) could lead to arbitrarily small π_{bk} (and hence arbitrarily large weights) for cases with leverage approaching zero. To prevent this, we could set a positive floor on the π_k. Alternatively, by making estimation of the population mean one of the objectives (Section 2.4), we guarantee sampling with positive probability over the entire population. Either method makes the design more robust against error in the approximate calculation of leverage and better prepared for possible post hoc decisions to estimate quantities not foreseen in the original design plan (Section 2.6). Furthermore, reasonably good estimation of means is needed to guarantee design-consistency of the first factor of (2).

2.3 Optimal design with individual- and aggregate-level variables

Now suppose that the covariate vector $\mathbf{t}_{bk}$ is nonempty and $\mathbf{u}_{bk}$ is constant in each cell, as when aggregated design information is available for cells corresponding to covariate classes of $\mathbf{u}$ within blocks. In CanCORS, if race ($\mathbf{u}$) but not income ($\mathbf{t}$) is known for individual subjects, and income distributions are available for each race in each census block, we would define cells to consist of people of a single race in a single census block.

Since cases in the same cell cannot be distinguished on covariates we further assume that $\sigma_{bk} = \sigma_b$ and $c_{bk} = c_b$ are *a priori* constant across the cell, so the optimal design also makes $\pi_{bk} = \pi_b$ constant in each cell.

We can now rewrite (3) as

$$V_a \approx \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(E_\pi \mathrm{Var}_\xi \sum_S \mathbf{x}_{bk} y_{bk}/\pi_b\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}$$

$$= \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(\sum_{b,k}(\sigma_b^2/\pi_b)\mathbf{S}_b\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}, \qquad (5)$$

where

$$\mathbf{S}_b = \begin{pmatrix} \mathbf{u}_b\mathbf{u}_b' & \mathbf{u}_b\bar{\mathbf{t}}_b' \\ \bar{\mathbf{t}}_b\mathbf{u}_b' & \mathbf{S}_{Tb} \end{pmatrix}$$

is the matrix of mean squares and crossproducts in cell b, with $\mathbf{S}_{Tb} = \mathbf{T}_b'\mathbf{T}_b/K_b$. The optimal sampling probabilities corresponding to (4) are then

$$\pi_b \propto \sigma_b\left(\sum_i d_i\,\mathbf{a}_i'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{S}_b(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}_i/K_b c_b\right)^{1/2}. \qquad (6)$$

If $\mathbf{t}$ is measured through a census of each cell, then $\bar{\mathbf{t}}_b$ and S_{Tb} are known exactly. The same principles apply, however, if S_{Tb} is not directly observed but instead is estimated under a model ζ. We then replace $\bar{\mathbf{t}}_b$ and $\mathbf{S}_b$ in (5) with predictive expectations $\tilde{\bar{\mathbf{t}}}_b = E_\zeta\,\bar{\mathbf{t}}_b$ and $\tilde{\mathbf{S}}_b = E_\zeta\,\mathbf{S}_b$. Examples might include the following situations: (1) data for each cell are only available for a sample, (2) design data are old and the distribution of design variables in the cell may have changed over time, or (3) data on individual elements are measured with error. Similarly, the distribution of $\mathbf{t}$ might be available only for a supercell that contains multiple values of $\mathbf{u}$ (for example, race and census block of residence are known for each individual, but the income distribution is known for the block as a whole but not for each race within the block), so $\bar{\mathbf{t}}_b$ and $\mathbf{S}_{Tb}$ must be estimated under a model.

2.4 More than one model

The preceding development assumes that all estimands of interest are combinations of parameters of a single model. More generally, the contemplated analyses might involve fitting several models, and V might sum the variances of combinations of parameters from these models. An obvious special case is estimation of a population mean (as suggested in Section 2.2), the coefficient of the model $y_k = \beta_0 \cdot 1 + \varepsilon_{bk}$, together with some regression coefficients. Another simple example is estimation of the means of variously defined domains, that is the coefficients of models of the form $y_k = \boldsymbol{\beta}'_{(m)}\mathbf{x}_{(m)bk} + \varepsilon_{(m)bk}$ where $\mathbf{x}_{(m)bk}$ is a vector of domain membership indicators with alternative domain definitions indexed by $m = 1, \ldots, M$, or contrasts of these means. For example, we might be interested in estimating mean outcomes both by race and by age.

If each of the combinations of interest only includes parameters of a single model, then each combination has its own design matrix, so the model index m can be identified with the estimand index i. Thus in (5) and (6) we replace $\mathbf{X}_U$ with $\mathbf{X}_{U(i)}$ and replace $\mathbf{S}_b$ with $\mathbf{S}_{(i)b}$.

If some estimands combine parameters from different models, we stack the estimators $\hat{\boldsymbol{\beta}}_{(m)}$ for the different models. Then in (5) and (6) we replace $\mathbf{X}_U'\mathbf{X}_U$ with $\mathrm{diag}(\mathbf{X}_{U(m)}'\mathbf{X}_{U(m)}, m = 1, \ldots, M)$ and redefine $\mathbf{S}_b$ as the combined sums of squares and crossproducts matrix for all of the models, with blocks

$$\mathbf{S}_{b(m,m')} = \begin{pmatrix} \mathbf{u}_{b(m)}\mathbf{u}_{b(m')}' & \mathbf{u}_{b(m)}\bar{\mathbf{t}}_{b(m')}' \\ \bar{\mathbf{t}}_{b(m')}\mathbf{u}_{b(m)}' & \mathbf{S}_{Tb(m,m')} \end{pmatrix}.$$

The remainder of the optimization is unchanged from Section 2.3.

2.5 Flexible contrast weights

In CanCORS, we are interested in the income effect controlling for race and averaged across races. It is less important to us how the races are weighted in that average, since the study areas are not representative of national proportions by race. Then we might estimate the poverty/nonpoverty income effect for each race and combine them with weights chosen to minimize the variance of the estimator of the weighted average of within-race income effects.

In general, we consider situations in which scientific interest is directed at estimating or testing *any* combination $\mathbf{a}_i = \mathbf{A}_i\mathbf{f}_i$ where $\mathbf{A}_i$ is fixed and each $\mathbf{f}_i$ is arbitrary (and not necessarily all of the same dimension) subject to the constraints $\mathbf{1}'\mathbf{f}_i = 1$, $f_{ij} \geq 0$. In our motivating example, the underlying model includes eight indicator variables for each of the groups defined by four race groups crossed with dichotomous poverty level, and $\mathbf{A}_1$ is an 8×4 matrix in which each column contains a 1 and -1 for the contrast between poor and nonpoor within one race. Then $\mathbf{f}_1$ contains the weights given to the contrast in each race, and $\mathbf{a}_1' = (f_{11}, -f_{11}, f_{12}, -f_{12}, f_{13}, -f_{13}, f_{14}, -f_{14})$ is the weighted contrast of the eight indicator coefficients.

Substituting into (5)-(6), we optimize over both sampling probabilities $\boldsymbol{\pi} = \{\pi_k\}$ and combining weights $\mathbf{f} = \{\mathbf{f}_i\}$. With multiple models, we use either of the formulations of Section 2.4, depending on whether the combinations of

interest include coefficients of one or several models. The definition of $\mathbf{a}_i$ is thus determined in part by scientific considerations and in part by the information available from the population at hand.

A natural approach to jointly optimizing $\boldsymbol{\pi}$ and $\mathbf{f}$ is alternately to minimize V with respect to $\boldsymbol{\pi}$ using the modified (6) and with respect to $\mathbf{f}$, observing the constraints on $\mathbf{f}$. In the optimization, $\mathbf{f}_i$ appears in an expression of the form $\mathbf{f}_i'\mathbf{D}_i(\pi)\mathbf{f}_i$. Minimizing subject to the constraint $\mathbf{f}_i' \cdot \mathbf{1} = 1$ using Lagrange multipliers, we obtain $\hat{\mathbf{f}}_i = \mathbf{D}_i^{-1}(\pi)\mathbf{1}/(\mathbf{1}'\mathbf{D}_i^{-1}(\pi)\mathbf{1})$ as long as $\pi_{bk} > 0$ and the nonnegativity constraints are not binding. If the nonnegativity constraints are binding, quadratic programming methods can be used.

2.6 Precision of unanticipated analyses

A design that is intended to be optimal for one regression coefficient might be very inefficient for other regression coefficients in the same or different models. Making the population mean one of the estimands helps to control this risk. We illustrate this by an example with design variables x_k, z_k with joint distribution

$$\zeta\colon (X, Z) \sim N\left(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right)$$

fully observed for individuals (indexed by k as in Section 2.2) and the following constant and univariate regression models:

Model 0: $y_k = \alpha_0 + \varepsilon_k^0$, $\varepsilon_k^0 \sim [0, \sigma_0^2]$

Model 1: $y_k = \alpha_X + \beta_X x_k + \varepsilon_k^X$, $\varepsilon_k^X \sim [0, \sigma_X^2]$

Model 2: $y_k = \alpha_Z + \beta_Z z_k + \varepsilon_k^Z$, $\varepsilon_k^Z \sim [0, \sigma_Z^2]$. To simplify notation we assume $\sigma_0^2 \approx \sigma_X^2 \approx \sigma_Z^2 \approx 1$, $\bar{x}_U = \bar{z}_U = 0$ and costs c_k are constant.

Consider the sample design optimized for $V = dV(\hat{\alpha}_0) + V(\hat{\beta}_X), d \geq 0$. By (4), the optimal design has $\pi_k \propto \sqrt{d + x_k^2}$. Under this design, the anticipated variance is approximated by $V(\hat{\beta}_Z) \approx n^{-1}\sigma_Z^2(Z_U'Z_U)^{-1}(Z_U'W_UZ_U)(Z_U'Z_U)^{-1}$ where $Z_U = (z_1, \ldots, z_N)'$ and $W_U = \text{diag}(\pi_1^{-1}, \ldots, \pi_N^{-1})$. Then $E_\zeta nV(\hat{\beta}_Z) \approx c_0 E_\zeta(Z^2/\sqrt{d + X^2})$ where c_0 depends only on d so $E_\zeta nV(\hat{\beta}_Z)$ depends on ρ and d. If $d = 0$ (no weight is attached to the estimation of mean), $E_\zeta nV(\hat{\beta}_Z) = \infty$ unless $\rho = \pm 1$. Thus unless the objective gives some weight to the variance of the mean estimator, the design is potentially very poor for the coefficients attached to covariates that are not in the span of variables of the optimized models. But if $d > 0$ we can decompose Z into components parallel and orthogonal to X, $Z = Z_1 + Z_2$ where $Z_1 = \rho X$ and $Z_2 = Z - \rho X$, so $Z_1 \perp Z_2$, $Z_2 \perp X$ and $E_\zeta Z_2 = 0$. Then $E_\zeta nV(\hat{\beta}_Z) = \rho^2 c_0 E_\zeta(X^2/\sqrt{d + X^2}) + (1-\rho^2) c_0 E_\zeta(1/\sqrt{d + X^2}) = \rho^2 E_\zeta V_{\text{opt}}(\hat{\beta}_X) + (1-\rho^2) E_\zeta V_{\text{opt}}(\hat{\alpha}_0)$.

In words, the variance of the coefficient of the new model is a combination of the two variances that were controlled in the optimization. This suggests that a design that includes estimation of the overall population mean in the optimization gives some protection against extreme inefficiency for other models with variables that were not considered in the original design, although the simple results given here do not necessarily generalize to cover every case.

3. Application: Regressions on race and poverty status

3.1 Description of sites and data

The CanCORS project (Ayanian *et al.* 2004) consists of five geographically-defined sites (northern California, Los Angeles, Alabama/Georgia, North Carolina, and Iowa) and two organizationally-based sites. The northern California site consists of 9 counties extending from the San Francisco Bay area to semirural Placer County on the Nevada border. This site is ethnically diverse and geographically varied and therefore best illustrates the methods. We describe results for this site in detail and then summarize results for other sites.

Our data were based on the 2000 U. S. Decennial Census "long form" sample and were extracted for the 9 counties of our target area (Alameda, Contra Costa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara, and Solano) from SF-3, Tables 159a-159i, "Poverty Status in 1999 by Age." We cross-tabulated the sampled residents at least 65 years old of each census block group (a small contiguous area roughly equivalent to several city blocks, henceforth referred to as a block) by race/ethnicity and income, using census sampling weights. The age restriction roughly corresponds to the ages of most incident cancer cases eligible for the study. Household income was dichotomized as exceeding or falling below the standard poverty line. The census included separate items on Hispanic ethnicity and race; we classified the population as Hispanic or as non-Hispanic white, Black, or Asian-American. A heterogeneous "Rest" category constitutes the remaining 3% of the elderly population. (For conciseness we henceforth refer to these as "race" categories.) The study site contained 844,560 over-65 individuals in 5,098 block groups, or an average of 166 per block group.

Table 1 summarizes the distribution of race and income in the northern California site. Blacks have the highest overall poverty rate and are also the most segregated (largest coefficient of variation of percent Black by block), consistent with national patterns of residential segregation (Denton and Massey 1993). Hispanics have the most relative geographical variation in poverty rates (largest coefficient of variation of poverty rates by block).

Table 1
Distributions of race and poverty for those with age ≥ 65 years, by census block group in the northern California site. (CV = coefficient of variation)

	White	Black	Asian	Hispanic	Rest	Total
Percent of population	65.70	6.40	16.80	8.20	3.00	100.00
Percent poor	5.20	14.20	10.10	10.60	11.30	7.20
CV block percent in race	0.46	2.94	1.21	1.73	2.28	-
CV block percent poor	1.53	1.37	1.58	1.89	2.30	1.16

3.2 Design conditions: Available information and design objectives

We calculated the efficiency relative to simple random sampling (SRS) of the optimal design for scenarios defined by two conditions: (1) the choice of objective function, and (2) the assumptions about the information available for determining sampling probabilities.

We considered six possible assumptions about available information for race (unavailable, or available at the individual level) and income (unavailable, only available by block, or available at the individual level). Because race is more often recorded in hospital records than income, we excluded the case where individual income group is known but race is only known by block group. Each assumption corresponds to a definition of the cell for the development of Sections 2.3 and a corresponding definition of variables $\mathbf{t}$ and $\mathbf{u}$:

1. No design information available: the cell is the entire population and $\mathbf{u}$ includes race and income. (Columns headed "SRS" in Table 2.)
2. Race alone: the cell is a race category, $\mathbf{u}$ contains race variables, and $\mathbf{t}$ is income. (Columns headed "Race.")
3. Block-aggregated data alone: the cell is a census block group, $\mathbf{u}$ is empty and $\mathbf{t}$ includes race and income. (Columns headed "Block.")
4. Individual race, block-aggregated income data by race: the cell is the population of one race in a block group, $\mathbf{u}$ is race, and $\mathbf{t}$ is income. (Columns headed "Race+Block.")
5. Individual income, no race data: the cell is an income group, $\mathbf{u}$ is income and $\mathbf{t}$ is empty. (Columns headed "Income.")
6. Race and income both available for each individual: the cell is a race by income category, $\mathbf{u}$ includes race and income, and $\mathbf{t}$ is empty. (Columns headed "Race+Income.")

We calculated optimal sampling rates under each assumption about available information, with a variety of objective functions. Each of the objective functions we considered weights together variances of coefficient estimates in some or all of four regression models: (1) the "intercept only" model whose single parameter is the population mean, (2) a race model parametrized as a white mean and contrasts for differences between whites and each of the other major race groups (Blacks, Hispanics, and Asians), (3) an income model parametrized as a nonpoor mean and a contrast between poor and nonpoor, and (4) an additive joint model including race and income effects. Every objective includes weight $d_{\text{mean}} > 0$, which guarantees that all $\pi_{bk} > 0$, avoiding numerical problems in the optimization. Thus, at least two models are represented in each objective (Section 2.4). When the objective weights both income and race effects, the single income effect is given weight $d_{\text{income}} = 3$ to match the three race effects with weights of 1.

We explored a selection of objective weights that emphasized estimation of race effects, income effects, or both. Each panel of Table 2 represents a single choice of objective weights d_i (third column) for the contrast coefficients $\mathbf{a}_i$ (second column) of a series of models (first columns). The fourth column shows the variance (normalized to unit sample size) $nV_{\mathbf{a}_i}$ for estimation of that coefficient under SRS assuming residual variance $\sigma^2 = 1$. The remaining columns present design effects, the ratios of the normalized variance $nV_{\mathbf{a}_i}$ for the optimized design with various assumptions about available design information to the variance under SRS. Rows with objective weight $d_i = 0$ do not affect the optimization but are included to illustrate the effect of each design on efficiency for estimating a coefficient that is not included in the objective function. The final row summarizes the weighted design effect correspondding to the loss function, that is, the weighted combination of variances.

3.3 Efficiency with fixed models

The first two objective functions optimize for estimation of race contrasts and the overall mean. Using individual race greatly improves efficiency for estimating Black and Hispanic effects. The greatest gains are for the Black effect (the smallest of the three major racial minorities), whose variance is reduced to 43% of its value under SRS. Conversely there is no gain for Asian-Americans, whose population representation is close to the optimal sampling rate. With this objective, once race is available, additional design information (block or individual income) is irrelevant to the optimization. If individual race is unknown, using block of residence can help with oversampling of Blacks (the most segregated group residentially), reducing the variance of the estimated Black effect to about 65% of that under SRS, but oversampling by block only slightly reduces the variance of the estimated Hispanic effect. Knowing income by itself is of little use to improve sampling for estimation of race effects.

Table 2
Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions

Objective 1: Optimized for race effects

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.1	1.0	181	119	181	100	181
Race	Black	1	17.2	43	65	43	99	43
	Asian	1	7.5	100	106	100	100	100
	Hispanic	1	13.7	55	90	55	100	55
Income	Poor	0	15.0	182	104	182	81	182
Race+Income	Black	0	17.4	44	65	44	99	44
	Asian	0	7.5	100	106	100	100	100
	Hispanic	0	13.8	55	90	55	100	55
	Poor	0	15.2	182	104	182	81	182
Total = $nV = n\sum d_i V_{a_i}$			38.6	59	82	59	99	59

Objective 2: Optimized for race effects and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	136	115	136	100	136
Race	Black	1	17.2	44	66	44	99	44
	Asian	1	7.5	100	104	100	100	100
	Hispanic	1	13.7	56	90	56	100	56
Income	Poor	0	15.0	121	101	121	82	121
Race+Income	Black	0	17.4	45	66	45	99	45
	Asian	0	7.5	100	104	100	100	100
	Hispanic	0	13.8	56	90	56	100	56
	Poor	0	15.2	122	102	122	82	122
Total = $nV = n\sum d_i V_{a_i}$			41.5	65	84	65	100	65

Objective 3: Optimized for income effect

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.001	1.0	103	154	173	173	173
Race	Black	0	17.2	75	119	152	163	163
	Asian	0	7.5	90	144	173	170	170
	Hispanic	0	13.7	86	142	196	168	168
Income	Poor	3	15.0	97	74	60	27	27
Race+Income	Black	0	17.4	75	119	153	164	164
	Asian	0	7.5	90	144	174	171	171
	Hispanic	0	13.8	86	143	197	169	169
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			45.0	97	74	60	27	27

Objective 4: Optimized for income effect and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	134	147	151	151
Race	Black	0	17.2	76	107	128	142	142
	Asian	0	7.5	91	127	145	148	148
	Hispanic	0	13.7	86	125	161	147	147
Income	Poor	3	15.0	97	75	61	27	27
Race+Income	Black	0	17.4	76	107	129	143	143
	Asian	0	7.5	91	127	146	149	149
	Hispanic	0	13.8	86	125	162	147	147
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			48.0	97	79	66	35	35

Table 2 (continued)
Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions

Objective 5: Optimized for separate race effects, income effect and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	117	135	114	150
Race	Black	1	17.2	54	74	55	109	53
	Asian	1	7.5	95	106	109	112	116
	Hispanic	1	13.7	67	96	69	112	67
Income	Poor	3	15.0	101	82	72	38	37
Race+Income	Black	0	17.4	55	74	55	109	52
	Asian	0	7.5	95	106	109	113	115
	Hispanic	0	13.8	67	96	69	112	66
	Poor	0	15.2	101	82	72	39	35
Total = $nV = n\sum d_i V_{a_i}$		0	86.4	86	86	74	73	56

Objective 6: Optimized for race effects and income effect in two-factor model and for overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	119	138	114	156
Race	Black	0	17.2	55	74	55	108	53
	Asian	0	7.5	95	107	109	112	114
	Hispanic	0	13.7	67	96	69	111	67
Income	Poor	0	15.0	101	82	72	38	37
Race+Income	Black	1	17.4	55	74	55	109	52
	Asian	1	7.5	95	107	109	113	113
	Hispanic	1	13.8	67	96	69	112	66
	Poor	3	15.2	101	81	71	39	35
Total = $nV = n\sum d_i V_{a_i}$			87.2	86	86	73	73	54

Objective 7: Optimized for income effect in two-factor model and for overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	135	149	148	156
Race	Black	0	17.2	77	100	98	139	97
	Asian	0	7.5	91	124	132	145	132
	Hispanic	0	13.7	86	122	135	144	122
Income	Poor	0	15.0	97	75	62	28	28
Race+Income	Black	0	17.4	77	100	98	140	96
	Asian	0	7.5	91	124	132	146	132
	Hispanic	0	13.8	86	122	135	144	121
	Poor	3	15.2	97	75	62	29	27
Total = $nV = n\sum d_i V_{a_i}$			48.6	97	79	67	36	35

Disproportionate sampling, tuned to optimize for estimation of race effects, inflates the variances of the other parameter estimators. When minimal weight is given to the mean in the optimization objective (Objective 1), this inflation can be quite large: a factor of 181% for the mean and income effects. Giving more weight to the mean (Objective 2) moderates this effect, reducing the variance inflation to 136% for the mean and 121% for the income effect, while only slightly increasing variances for the race effects.

The minimum possible normalized variance for estimation of the income effect (Objective 3) is 4 (27% of the variance under SRS), attained when income is known for individuals under a design that divides the sample equally between poor and nonpoor. With block-level information, variance can be reduced to 74% of that under SRS. Although knowing race alone has little benefit for this objective, adding individual race to block-level information further reduces the variance of the estimated income effect to 60% of that under SRS. Variances of estimates of the mean and of race effects are substantially increased under

these designs, but increasing the weight of the mean (Objective 4) substantially ameliorates the variance inflation for the mean and race effects, only slightly increasing the variance of the estimated income effect.

Including both race and income effects in Objective 5 yields designs that are not quite as good as the optimal designs for either alone, but still much better than SRS. For example, variances of the race effects with race and block of residence known are 10% to 24% higher than with the designs using the same design information but separately optimized for race or income. When only individual race or only individual income is known, the design essentially optimizes for the effects corresponding to the available variable, inflating the variance of estimated effects of the other variable.

The design optimized for joint race and income effects in the two-factor additive model (Objective 6) is quite close to that optimizing for race and income effects in separate marginal models (Objective 5). When optimizing for separate effects, variances of these effects are slightly smaller than those of the corresponding effects in the two-factor model. When optimizing for effects in the joint model, their variances are reduced although in most cases still slightly larger than those of the corresponding effects in marginal race and income models, due to the partial confounding of race and income effects.

Likewise, optimization for the income effect in the two-factor model (Objective 7) is fairly similar to optimization for the univariate income effect (Objective 4) when no race data are available. Making race data available together with either block or individual income, however, considerably reduces variances for race effects under the design for the two-factor model. Because of the partial confounding of race and income effects under this model, this design adapts to estimate the former more efficiently, accumulating more data at the design points that are critical to unconfounding these effects.

3.4 Efficiency with flexible contrast weights

We next consider the potential benefits of estimating income effects under a flexible weighting scheme (Table 3). The objective function considers coefficients of two models, the constant model whose parameter is the population mean, and a model with indicator variables for each race-by-income cell. The income effect within each race is estimated as the difference of the coefficients for poor and nonpoor within that race, and these estimates are combined with flexible weights to estimate an overall income effect. This strategy is most nearly parallel to Objective 7, which also estimates income effects controlling for race. The flexible-contrast analysis is less model-dependent than the two-factor model in that it does not rely on that model's additivity assumption. On the other hand, the way the races are combined does not necessarily reflect population proportions. The weights given to the income contrast in each race, estimated as described in Section 2.5, are presented in the lower panel of Table 3 to demonstrate how this approach allows us to modify the estimand to exploit available design information. (The alternating-optimization algorithm converged to adequate accuracy within 7 iterations.)

Under SRS the variance of the income effect under the flexible-weights model is slightly larger than in the two-factor model (15.91 versus 14.99). The weight given to the white contrast under this design (51%) is less than the white proportion of the population (66%) because relatively few whites are poor and therefore the income contrast among whites is relatively imprecise. Conversely, the weight for the Black income contrast (12%) is almost twice that group's share of the population, because of the disproportionately high poverty rates in that group.

Using individual race in the design accentuates this disproportion: more sample, and much more weight (75%), is given to the Black group, with the highest percentage in poverty. Thus flexible weighting makes possible a large reduction in the variance of the estimated income effect (to 63% of that under SRS) using only race, which was not possible under the more restrictive two-factor DPQ model.

Block-level information is slightly less useful for this design than race information. The combination of block and race information, however, is very powerful, reducing the variance of the income effect to 48% of that under SRS. Under this design, much more weight (46%) is given to the Hispanic income contrast, which can be estimated efficiently because of the greater income segregation among Hispanics (Table 1). When individual income is available (with or without race), the contrasts weights approximate the proportions by race, since efficient income contrasts can be obtained within any race and the inclusion of the overall mean in the objective pulls the design toward proportionate sampling. Thus, the design is dramatically different under alternative assumptions about availability of design information.

3.5 Comparisons across sites

Table 4 compares the gains for disproportionate sampling at four CanCORS sites, excluding the nongeographical sites and one site (Iowa) that was almost all white. At each site we optimized for unit ($d_i = 1$) weighting of variances of overall mean and the income effect in the two-way model (proportional to Objective 7), under alternative assumptions about available design information. The theoretical minimum for this objective with a balanced population is 5 ($V_{\text{mean}} = 1$, $V_{\text{income}} = 4$). SRS is inefficient at every site,

especially in Alabama and northern California, and race information alone would be of little help. Conversely, the best variance attainable using full race and income information on individuals is between 5.60 and 5.72 at each site. Oversampling based on block-level income information would substantially reduce variances, with substantially greater gains in Alabama and northern California than in the other sites.

4. Discussion

To develop design alternatives for a health services study, we extended previous methods for optimal design in domain estimation to show how an optimal unequal-probability sampling scheme can be designed for estimation of regression coefficients in one or more models. In our application, substantial reductions in variance were possible even if some variables were only available for geographical aggregates. Particularly large gains were possible for categorical regressors (poverty status, race) with very imbalanced distributions.

In essence, our approach to survey design with imprecisely measured design variables uses the predictive distribution of the design variables for each sampled unit, specifically the expectations of the variables and of their squares and cross-products. This concept unites design using cell aggregates (estimated from census or sample data), using variables measured with error, or using a sampling frame whose units might have changed their characteristics over time.

Table 3
Normalized variances and contrast weights for optimal DPQ designs with flexible weighting of income contrasts by race. Lines for fixed contrasts are included to demonstrate the effect of various choices of flexible weights, for comparison to fixed-weight objective scenarios

Variances for "flexible-weight" estimate and for contrasts represented in Table 2								
				Variance as percent of variance under SRS (by available design information)				
Model	Effect	Weight (d_i)	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	233	139	206	152	152
Flexible contrast	Income	3	15.9	63	74	48	28	26
Race	Black	0	17.2	34	90	61	143	139
	Asian	0	7.5	197	124	172	149	146
	Hispanic	0	13.7	172	124	61	148	144
Income	Poor	0	15.0	209	77	124	27	39
Race+Income	Black	0	17.4	35	90	61	144	139
	Asian	0	7.5	197	124	173	150	147
	Hispanic	0	13.8	172	124	61	148	144
	Poor	0	15.2	210	77	125	29	39
Total = $\sum d_i V_{a_i}$			50.7	73	78	57	35	33

Optimum weights (as percent, 100% × f_i) of each within-race income contrast in calculation of the combined estimate of the income effect, under each design information assumption. (Columns may not sum to 100% due to roundoff error.)

	Design information assumptions					
Contrast	SRS	Race	Block	Race+Block	Income	Race+Income
Black	12	75	17	25	9	6
Asian	24	9	26	16	21	17
Hispanic	12	5	13	46	11	8
White	51	12	45	13	59	68

Table 4
Normalized objective function for optimal DPQ designs under equal (d_i = 1) weighting of variances of the overall mean and the income effect in the two-way model, at four CanCORS sites

	Variance under	Variance as percent of variance under SRS				
Site location	SRS	Race	Block	Race+Block	Income	Race+Income
Alabama	16.2	97	79	67	36	35
Los Angeles	11.8	98	85	76	49	47
North Carolina	10.2	97	89	86	59	55
Northern California	16.2	97	79	67	36	35

The methods described here for optimizing element sampling probabilities can be combined with stratification and cluster or multistage sampling. (Neither of these design features appeared in the CanCORS study which motivated our research. Stratification was inconvenient given the sequential identification of subjects and there was little prior information to guide construction of homogeneous strata. Telephone interviewing made it operationally unnecessary to cluster our subjects.) Because these design features can affect the sampling distributions of both the design and outcome variables, and the design objectives involve both the posited population model and the scientific model of interest, the number of possible combinations is even larger than in design for estimation of a population mean. We therefore limit ourselves to suggesting a few ideas to be followed up in future research.

Stratification can improve a design for a regression analysis in at least three ways: (1) to implement disproportionate sampling (using probabilities equal or close to those derived under our methodology), (2) to control the distribution of design variables to be closer to the optimal design than in an unstratified unequal-probability design, and (3) to reduce the within-stratum variation of the case influence statistics and thereby reduce the variance of coefficient estimates (Fuller 1975). Since the efficiency of the design is insensitive to small deviations around the optimum, some stratified designs with equal probabilities within strata might approach the efficiency of the optimal design. *Ad hoc* stratifications might have poorer efficiency, even with optimal allocation to strata. For example, stratifying blocks by the least prevalent race-income group represented yielded a design with about half the efficiency gain of our design using aggregated block composition.

With regard to the last point, note that designing homogeneous strata for estimation of regression coefficients is likely to be more difficult than for estimation of a mean. The influence of an observation depends on its residual from the regression model, not its raw value, so to reduce homogeneity the stratification would have to involve predictive variables not included in the model. Influence also depends on the observation's leverage for each coefficient, a possibly complex function of the covariates.

For cluster sampling, the equivalence of $E_\pi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ and $E_\xi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ might not hold except under restrictive assumptions such as independent residuals; thus the terms of the middle factor of (5) would take a more complex form. There are several possible cases for cluster sampling depending on the relationship between the cells and the clusters, which should be elaborated on further research.

Another natural extension is to nonlinear regression models and other estimands defined by estimating equations. The weighted least squares formulation of the Newton-Raphson step (McCullagh and Nelder 1989, sec. 2.5) for a generalized linear model can be applied by suitably defining σ^2_{bk} in (3) and hence in (4)-(6); a similar procedure can be applied for other estimating equations (Binder 1981; Binder 1983; Morel 1989). Because the variances are functions of the model predictions, implementing this modification requires design assumptions about the fitted model as well as about the distribution of the covariates.

Every optimization has its costs, which for our methods can be both practical and statistical.

In the CanCORS study, incident cases of the cancers under study were identified in real time through a field operation ("rapid case ascertainment"); patients then had to be contacted on a very tight schedule to start contacting them for interviews within the desired interval (3 months from their dates of diagnosis). Thus, the practical issues of survey implementation were exacerbated. Among the concerns that ultimately led us not to implement the DPQ design were (1) the difficulty of accurately geocoding patients within the time frame allowed; (2) incomplete and inaccurate race identification in the case ascertainment data, and (3) lower-than-expected participation rates, which made any sampling problematical.

Such issues are less problematic in surveys with a static sampling frame that can be processed on a less stringent timeline, particularly in large-scale and/or repeated surveys in which even modest variance reductions justify some added complexity. They could be used, for example, to evaluate the potential gains through geographically-based oversampling in surveys for which national estimates by race are required.

Statistical concerns about our design strategy arise because optimization for one set of predetermined statistical objectives is likely to reduce efficiency for others. It is difficult in any but the most tightly focused study to anticipate all potential analyses. Simultaneous optimization for a reasonably comprehensive collection of analyses, and investigation of sensitivity of the design to varying the relative weights of the various objectives, should give some protection against an overspecialized design. However, this approach can only be used with variables for which there are some data prior to the study. The results in Section 2.6 suggest that monitoring the effect of disproportionate sampling on the precision of the population mean gives some protection against designs that are excessively inefficient for unanticipated analyses and variables, although the bounds there are not very general.

More broadly, we might ask when the DPQ analysis is the scientifically relevant estimand. Regression models are often used in analyses intended to be generalizable to broader populations, rather than to describe the finite

population at hand, just as the CanCORS sites were selected purposively to study patterns and variations in care that might reflect broader national patterns. While using sampling weights in enumerative studies is relatively uncontroversial, there has been a lively debate about the use of weights in analytic studies (Hansen, Madow and Tepping 1983 and discussion; DuMouchel and Duncan 1983; Bellhouse 1984; Pfeffermann 1993, Fuller 2002, sec. 5). A population-descriptive analysis offers some robustness against the possibility that the sample will be selected in way that distorts typical relationships. Thus, even where a pure DPQ analysis cannot be justified on grounds of enumerative representativeness, a sample drawn to optimize unweighted estimation of regression coefficients might have limited scientific value. For example, suppose that the CanCORS data would be analyzed with an *unweighted* regression to estimate a simple income effect (a contrast of means), using block level design information from the census. Optimally the sample would draw from a collection of blocks which, taken together, have about half their residents in poverty. Since poverty rates are rarely that high, this effectively requires sampling only from the blocks with the highest poverty rates. Such a sample would be unrepresentative of either of the income groups. Similarly, a sample that overrepresented Black residents by sampling from mostly Black blocks would (if analyzed without weights) be unrepresentative of the Black population in general, because the services available in highly segregated areas are likely to differ from those in more mixed areas.

More general formulations are needed, with clearly stated assumptions and objectives, that "consider[s] the model parameters as the ultimate target parameters but at the same time focuses on the DPQ's as a way to secure the robustness of the inference" (Pfeffermann 1993), taking into account the scientific objectives of the study. Previous proposals include testing the null hypothesis that the weights have no effect on the regression (DuMouchel and Duncan 1983; Fuller 1984), or including design variables (Nathan and Holt 1980; Little 1991) or the weights themselves (Rubin 1985) as control variables in the regression. These approaches are problematical, however, when the weights are functions of the covariates of primary scientific interest. We have attempted through flexible contrast weighting (Section 2.5) to take a step toward such a general formulation, extending the DPQ approach to allow a focus on a range of valid inferences for particular scientific objectives rather than exclusively on inference for finite populations. From this range, the investigator can select an inferential objective and sample design adapted to the structure of the population and the practicalities of study design.

Appendix

Equivalence of sampling and model variances

We show that $E_\xi \mathrm{Var}_\pi \hat{\boldsymbol{\beta}} \approx E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}$ under the following conditions:

1. $1/N\ \mathbf{X}_U'\mathbf{X}_U \to \Sigma$ for some positive definite Σ. This minimal condition relates the hypothetical sequence of populations.
2. The design-based regression estimator can be written as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_U + R_{n,S}$ where $\mathrm{Var}_\xi E_\pi R_{n,S} = o(n^{-1})$ and $\mathrm{Var}_\pi E_\xi R_{n,S} = o(n^{-1})$. Note that $\hat{\boldsymbol{\beta}}$ cannot strictly be defined as in (2), because the matrix inverse is undefined when the sample values of x do not span the design space and hence its expectation and variance are also undefined. A scalar ratio estimator likewise might be undefined with nonzero but $o(n^{-1})$ probability because the sample might have only 0 values for the denominator variable. Assigning some arbitrary value in that event, the estimator nonetheless could have good asymptotic properties. A similar argument lets us assume that a suitable $\hat{\boldsymbol{\beta}}$ can be defined. We do not specify how (2) must be modified to technically satisfy the condition since this depends on the specifics of ξ and the sequence of designs.
3. $\max(\pi_i) = O(n/N)$ and $n = o(N)$, essentially our assumption that finite population corrections can be ignored.
4. Homoscedasticity, $\mathrm{Var}_\xi y_k = \sigma^2$; this is not restrictive since it can always be made true by a suitable transformation of x and y.

Under these conditions,

$$\begin{aligned}\mathrm{Var}_{\pi\xi}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}_\pi E_\xi \hat{\boldsymbol{\beta}} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= \mathrm{Var}_\pi \boldsymbol{\beta} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= o(n^{-1}) + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}\end{aligned}$$

On the other hand

$$\mathrm{Var}_{\xi\pi} \hat{\boldsymbol{\beta}} = \mathrm{Var}_\xi E_\pi \hat{\boldsymbol{\beta}} + E_\xi \mathrm{Var}_\pi \hat{\boldsymbol{\beta}}$$

The first term in the above equation is

$$\begin{aligned}\mathrm{Var}_\xi E_\pi(\boldsymbol{\beta}_U + R_{n,S}) &= \mathrm{Var}_\xi(\boldsymbol{\beta}_U + E_\pi R_{n,S}) \\ &= (\mathbf{X}_U'\mathbf{X}_U)^{-1}\sigma^2 + o(n^{-1}) + o(N^{-1/2}n^{-1/2}) \\ &= O(N^{-1}) + o(n^{-1}) + o(n^{-1/2}N^{-1/2}) = o(n^{-1})\end{aligned}$$

This proof is an elaboration of one by Isaki and Fuller (1982), summarized in Pfeffermann (1993, page 321).

Acknowledgements

This research was funded by grants U01-CA93344 (Zaslavsky and Zheng), U01-CA93324 (Zaslavsky), and U01-CA093348 (Adams) from the National Cancer Institute. The authors thank Nat Schenker and Van Parsons for useful comments on an earlier draft, and the associate editor and two referees for thoughtful comments.

References

Ayanian, J.Z., Chrischilles, E.A., Wallace, R.B., Fletcher, R.H., Fouad, M., Kiefe, C.I., Harrington, D.P., Weeks, J.C., Kahn, K.L., Malin, J.L., Lipscomb, J., Potosky, A.L., Provenzale, D.T., Sandler, R.S., Vanryn, M. and West, D.W. (2004). Understanding cancer treatment and outcomes: The Cancer Care Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 2, 2992-2996.

Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.

Binder, D.A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7, 157-170.

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

Botman, S.L., Moore, T.F., Moriarity, C.N. and Parsons, V.L. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004*. Vital and Health Statistics, 2(130). Washington, DC: National Center for Health Statistics.

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

Chromy, J.R. (1987). Design optimization with multiple objectives. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA. 194-199.

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

Denton, N.A., and Massey, D.S. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard University Press.

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

Eaton, D.K., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Harris, W.A., Lowry, R., Mcmanus, T., Chyen, D., Shanklin, S., Lim, C., Grunbaum, J.A. and Wechsler, H. (2006). Youth risk behavior surveillance - United States, 2005. *Morbidity and Mortality Weekly Report*, 55(SS-5), 1-108.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.

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

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

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

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-97.

Kish, L. (1974). Optimal and proximal multipurpose allocation. Alexandria, VA. In *Proceedings of the Social Statistics Section*, American Statistical Association, 111-118.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (Second Edition). London: Chapman & Hall Ltd.

Morel, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, 15, 203-223.

Morris, C., Newhouse, J.P. and Archibald, R. (1979). On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments. *Research in Experimental Economics*, 1, 199-220.

Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, Series B, *Methodological*, 42, 377-386.

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

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

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

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith). Amsterdam: Elsevier/North-Holland, 463-472.

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

Sastry, N., Ghosh-Dastidar, B., Adams, J. and Pebley, A.R. (2005). The design of a multilevel survey of children, families, and communities: The Los Angeles Family and Neighborhood Survey. *Social Science Research*, 35, 1000-1024.

Thomsen, I., Tesfu, D. and Binder, D.A. (1986). Estimation of design effects and intraclass correlations when using outdated measures of size. *International Statistical Review*, 54, 343-349.

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

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. In *ASA Proceedings of the Social Statistics Section*, American Statistical Association. Alexandria, VA.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. In *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, 497-502.

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

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 79-89
Statistics Canada, Catalogue No. 12-001-X

Generalized regression estimators of a finite population total using the Box-Cox technique

Yan Li [1]

Abstract

A new generalized regression estimator of a finite population total based on the Box-Cox transformation technique and its variance estimator are proposed under a general unequal probability sampling design. By being design consistent, the proposed estimator maintains the robustness property of the GREG estimator even if the underlying model fails. Furthermore, the Box-Cox technique automatically finds a reasonable transformation for the dependent variable using the data. The robustness and efficiency of the new estimator are evaluated analytically and via Monte Carlo simulation studies.

Key Words: Generalized regression (GREG) estimator; Box-Cox technique; Pseudo-maximum likelihood (PML).

1. Introduction

Generalized regression (GREG) estimators for finite population totals and means are derived using suitable regression models. Although models are employed to construct such estimators, randomization must be used to select the samples and to evaluate the statistical properties of the resulting estimation strategies. Examples may be found in Särndal, Swensson and Wretman (1992), Estevao, Hidiroglou and Särndal (1995), Fuller, Loughin and Baker (1994), and Jayasuriya and Valliant (1996). A good model is crucial in limiting the variability of a *model-assisted* estimator like the GREG. If the assumed model describes the finite population well, the GREG estimator can potentially bring about a large variance reduction when used in place of the Horvitz-Thompson estimator (Horvitz and Thompson 1952). A general discussion of regression estimation can be found in Fuller (2002). Särndal *et al*. (1992) provide a comprehensive description of the model-assisted framework for constructing survey estimators.

Studies on the GREG estimator have mostly been conducted in the context of linear regression modeling. The GREG essentially incorporates relevant auxiliary variables through their known population control totals even when the auxiliary variables are known for every unit in the population (Cassel, Särndal and Wretman 1976; Särndal 1980; Deville and Särndal 1992; Särndal *et al*. 1992; Jiang and Lahiri 2006). The availability of complete auxiliary information is fairly common these days: census data, administrative registers, remote sensing data and previous surveys provide a wealth of valuable information that can be used to increase the precision of the estimation procedure (Montanari and Ranalli 2003). As a result, complex models and flexible techniques making use of complete auxiliary information have been introduced into survey sampling in recent years. Penalized spline techniques have been adapted to construct model-based (Zheng and Little 2004) and model-assisted (Breidt, Claeskens and Opsomer 2005) estimators for a finite population total based on complex survey data. Breidt and Opsomer (2000) considered a nonparametric, model-assisted regression estimator using local polynomial regression and showed that nonparametric regression can significantly improve the efficiency of estimators when parametric models are misspecified. Their work was further extended from the single-covariate model to the case of the semiparametric additive model. Wu and Sitter (2001) fit a general working model, which could have both linear and nonlinear components, and then calibrated on the resulting fitted values using simple linear regression. Montanari and Ranalli (2005) combined model calibration estimation and nonparametric methods and proposed nonparametric model-assisted estimators for a finite population mean.

In mainstream statistics, a suitable transformation on the dependent variable in the assumed model is often taken to achieve normality, linearity, and homoscedasticity (Carroll and Ruppert 1988), but the literature on transformations in finite population inference is not very rich. There is, however, a growing interest in developing methods that use an appropriate transformation with survey data. In some survey applications, especially in business and establishment surveys, it is common to have highly skewed continuous and positive survey variables (*e.g.*, income). To estimate the finite population total of the survey variable, a linear model may not be appropriate for a study variable, but may be reasonable for a strictly monotonic transformation of the study variable. Chen and Chen (1996) considered transformed survey data in order to improve on the precision of the normal approximation. Korn and Graubard (1998) compared different confidence intervals, including intervals

1. Yan Li, University of Maryland at College Park. Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS; Executive Plaza South; Room 8014, 6120 Executive Boulevard, MSC7244, Bethesda, MD 20892, U.S.A. E-mail: lisherry@mail.nih.gov.

based on a logit-transformation, for proportions with small expected number of positive counts. Karlberg (2000) proposed an estimator based on a lognormal-logistic super-population model to predict the finite population total of a highly skewed survey variable. The simulation results indicated that the lognormal-logistic model estimator offers a sensible alternative to other estimators, especially when the sample size is small. Chambers and Dorfman (2003) discussed the estimation of a finite population mean under certain general but known transformation on the continuous data.

Researchers find the transformation technique useful in analyzing survey data. The key step is the identification of an appropriate transformation that fits the survey data well. In many applications, the form of transformation is determined subjectively. Unfortunately, prior knowledge or theory may not suggest the transformation to be used. In such situations, it would be convenient to determine the transformation adaptively using the survey data.

The work of Box and Cox (1964) has led to the development of "data-decide-transformation" methods for constructing models with independently and identically distributed errors. Their paper and other papers on the subject, including Tukey (1957), John and Draper (1980), and Bickel and Doksum (1981), have inspired a large volume of applied research. Spitzer (1976) examined the relationship between the demand for money and the liquidity trap with a generalized Box-Cox model. In the context of research related to malaria, Newman (1977) concluded that the Box-Cox functional specification was superior to earlier specifications. Miner (1982) and Davison, Arnade and Hallahan (1989) considered modeling of soybean yield functions and the U.S. soybean export respectively. They concluded that the Box-Cox transformation provides approximately normally distributed error terms. A bibliography related to the Box-Cox transformation can be found in a review paper by Sakia (1992). For an application of the Box-Cox methodology to a mixed linear model, see Gurka (2004, 2006).

Li and Lahiri (2007) used the Box-Cox transformation on the study variable to generate robust model-based predictors of a finite population total. Model-assisted estimators were also mentioned in a sub-section (Section 2.6), but the properties of the proposed estimators were not investigated. This article provides that analysis. The Box-Cox technique is employed to fit a regression between the study variable and a set of auxiliary variables. The fitted regression is then used to predict the values of study variable for the unobserved units of the finite population which, in turn, provide an adaptive regression type estimator within the model-assisted framework.

The article is organized as follows. In Section 2, a new estimator is proposed and the analytical properties of the proposed estimator with respect to the sampling design are investigated. To better assess the robustness and the efficiency of the proposed estimator, we compare it to GREG estimators based on the underlying linear and log-linear working models via Monte Carlo simulations in Section 3. Section 4 discusses the results. Finally, we offer some concluding remarks about areas for potential future research in Section 5.

2. A new estimator of the finite population total

Suppose that the quantity of interest is the finite population total

$$T = \sum_{i \in U} y_i,$$

where $U = \{1, ..., N\}$ denotes a finite population of known size N, and $y_i > 0$ is the value of the study variable associated with unit i. Write $\mathbf{Y} = (y_1, ..., y_N)'$. To estimate T, a sample s of size n is drawn from the finite population using a probability sampling scheme. Let w_i be the sampling weight for unit i. The sampling weight is simply the inverse of the inclusion probability for the unit i, denoted by $\pi_i = P(i \in s)$ $(i = 1, ..., N)$. We assume that we have information on $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N)'$, where $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})'$ is a column vector of k known auxiliary variables for the unit i. For any sample s of size n, we redefine $\mathbf{Y}$ and $\mathbf{X}$ so that the first n rows of $\mathbf{Y}$ and $\mathbf{X}$ correspond to those in the sample. Write

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

where

- $\mathbf{y}_s$ is a $n \times 1$ column vector of observed study variable;
- $\mathbf{y}_r$ is a $(N - n) \times 1$ column vector of unobserved study variable;
- $\mathbf{X}_s$ is a $n \times (k + 1)$ matrix of known auxiliary variables in the sample;
- $\mathbf{X}_r$ is a $(N - n) \times (k + 1)$ matrix of known auxiliary variables outside the sample.

Throughout the paper, we use E_d and V_d to denote the expected value and variance with respect to the sampling design.

2.1 GREG estimators of finite population totals

The GREG estimator is defined here as

$$\hat{T}_G = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} (y_i - \hat{y}_{i,w}) / \pi_i,$$

where $\hat{y}_{i,w}$ is the predictor of y_i based on a model. Regardless of how well the underlying model describes the population, GREG estimators of the finite population total are design-consistent under mild conditions (Särndal *et al.* 1992). The most commonly used model is the standard linear regression model, given by

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$, a N-variate probability distribution with the mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{I}$, and $\mathbf{I}$ is the $N \times N$ identity matrix (nothing would be lost in this context by replacing $\sigma^2\mathbf{I}$ with a more general positive definite matrix). In this equation, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ column vector of regression coefficients. Both σ^2 and $\boldsymbol{\beta}$ are unknown superpopulation parameters. An unbiased predictor for the i^{th} unit is

$$\hat{y}_{i,w} = \mathbf{x}_\mathbf{i}'\hat{\boldsymbol{\beta}}_w, \qquad (1)$$

where $\hat{\boldsymbol{\beta}}_w$ is the weighted least square (WLS) estimator of $\boldsymbol{\beta}$ under $\mathbf{M}_1$ and

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i\in s} w_i \mathbf{x}_\mathbf{i}\mathbf{x}_\mathbf{i}'\right)^{-1}\left(\sum_{i\in s} w_i \mathbf{x}_i y_i\right).$$

In some applications, especially in business and agricultural surveys, a linear model may not be appropriate for y, but may be reasonable for a strictly monotonic transformation of y. For the data set given in Royall and Cumberland (1981), Chen and Chen (1996) observed that the finite population distribution was severely skewed and that the log-transformation helped achieving symmetry. The need and the benefit of taking the log-transformation were obvious. Therefore, we consider the log-linear regression model where the log-transformation is used on the dependent variable

$$\mathbf{M}_2 : \log\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. An obvious predictor for the i^{th} unit is given by

$$\hat{y}_{i,w} = e^{\mathbf{x}_\mathbf{i}'\hat{\boldsymbol{\beta}}_{lw}}, \qquad (2)$$

where

$$\hat{\boldsymbol{\beta}}_{lw} = \left(\sum_{i\in s} w_i \mathbf{x}_\mathbf{i}\mathbf{x}_\mathbf{i}'\right)^{-1}\left(\sum_{i\in s} w_i \mathbf{x}_i \log y_i\right).$$

Model $\mathbf{M}_2$ requires a subjective specification of the transformation applied to the study variable. This may be reasonable in situations where we know the appropriate transformation from prior empirical evidence or from the theory. In absence of any prior knowledge about the transformation, it is prudent to choose the transformation from among a flexible family of transformations using the data.

Tukey (1957) considered the following family of power transformations:

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

where $y > 0$. In order to remove the discontinuity at $\lambda = 0$, Box and Cox (1964) proposed the following family of transformations:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

where $y > 0$. The parameter λ determines the nature of transformation. For example, $\lambda = 1, 0, 0.5, -1$ correspond to no transformation, log-transformation, square root transformation, and reciprocal transformation, respectively. The transformation parameter λ is estimated by the data. The Box-Cox analysis may lead to a log-transformation, but may equally lead to some other transformation in the above family - it depends on the actual data observed.

We consider the following superpopulation model for the transformed study variable:

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ are approximately normal with mean $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{I}$.

Schlesselman (1971) showed that the maximum likelyhood estimators of the Box-Cox model parameters are scale-invariant so that rescaling the original y-variable leads to the same log-likelihood function as long as the regression model contains an intercept term. Following his study and most papers on the Box-Cox models, we include an intercept term in the model.

Under $\mathbf{M}_3$, the predictor for the i^{th} unobserved unit is obtained by a simple back-transformation from the Box-Cox transformation:

$$\hat{y}_{i,w} = g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) = (\hat{\lambda}_w \mathbf{x}_\mathbf{i}'\hat{\boldsymbol{\beta}}_w + 1)^{1/\hat{\lambda}_w}, \qquad (3)$$

where $\hat{\boldsymbol{\beta}}_w$ and $\hat{\lambda}_w$ are estimators of the model parameters. The estimation method is explained in next subsection. Equations (2) and (3) do not provide unbiased predictors for y_i under the respective underlying models. Li and Lahiri (2007) showed that if the error variance is small, that is, the model fits the data very well, the right-hand sides of equations (2) and (3) are good alternatives to the unbiased predictors. For simplicity and to reduce the computational burden, we will treat the right-hand sides of equations (2) and (3) as appropriate alternatives for the unbiased predictors. Recall that our purpose is to describe model-assisted estimators for finite population totals. The

underlying model is only used to suggest an estimator, which will be evaluated under the randomization framework. Even though the predictors of the individual y_i are biased, we can still construct design-consistent estimators for finite population totals, unlike the strictly model-based estimators proposed by Li and Lahiri (2007).

We denote GREG estimators under the three models $\mathbf{M_1} - \mathbf{M_3}$ by $\hat{T}_{G\_L}$, $\hat{T}_{G\_\mathrm{LOG}L}$, and $\hat{T}_{G\_BC}$, respectively. The $\hat{T}_{G\_BC}$ estimator is different from the $\hat{T}_{G\_L}$ and $\hat{T}_{G\_\mathrm{LOG}L}$ because the data dictates the transformation to be used.

It is possible to incorporate Box-Cox transformations on both the y-variable and x-variables. In the past, different functional forms of the Box-Cox model have been investigated. Khan and Ross (1977), Spitzer (1976), Zarembka (1968), Boylan, Cuddy and O'Muircheartaigh (1980), and others, considered a particular case of the general Box-Cox model when a common transformation parameter is assumed for the y-variable and x-variables. Gemmill (1980), Boylan, Cuddy and O'Muircheartaigh (1982), and others, applied the general Box-Cox model with different transformation parameters on y-variable and x-variables. Li and Lahiri (2007) also used the general Box-Cox model to predict the finite population total under a model-based framework. In future, this method can be extended to different functional forms of the Box-Cox transformation.

We only discuss a Box-Cox transformation of the y-variable here. This allows for a fairer comparison among the three GREG estimators.

2.2 Estimation of model and transformation parameters $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \lambda, \sigma^2)'$ using the pseudo-maximum likelihood (PML) method

In order to ease the estimation of λ using existing computational procedures, one must replace $\mathbf{Y}^{(\lambda)}$ in the model $\mathbf{M_3}$ by a scaled transformation $\mathbf{Y}^{*(\lambda)}$. For the i^{th} unit,

$$y_i^{*(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda \tilde{y}^{\lambda-1} & \lambda \neq 0, \\ \tilde{y}\log(y_i) & \lambda = 0, \end{cases}$$

where $\tilde{y}$ is the geometric mean of y's. The following calculation will be based on the new scaled model:

$$\mathbf{M}_4 : \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{\varepsilon}^*$ are approximately normal with mean $\mathbf{0}$ and variance matrix $\sigma_e^{*2}\mathbf{I}$. Let $\boldsymbol{\varphi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma_e^{*2})'$.

The maximum likelihood estimator (MLE) of $\boldsymbol{\varphi}^*$ maximizes the log-likelihood

$$l(\boldsymbol{\varphi}^*) = \sum_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}),$$

where

$$f(y_i; \boldsymbol{\varphi}^*, \tilde{y}) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-(2\sigma_e^{*2})^{-1}(y_i^{*(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta}^*)^2\right\} \cdot (y_i/\tilde{y})^{\lambda-1}.$$

Skinner, Holt and Smith (1989) redefines $\boldsymbol{\varphi}^*$ as the value of $\tilde{\boldsymbol{\varphi}}^*$ which maximizes

$$l(\tilde{\boldsymbol{\varphi}}^*) = \sum_{i\in U} \log f(y_i, \tilde{\boldsymbol{\varphi}}^*),$$

the sum being taken over all units in the finite population. Thus, among all possible models $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, the one which "best fits" the finite population is chosen. If we choose the $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$ family poorly, this best fit will still be poor, but our inference treats it as the target we are trying to hit with our sample data. Thus, it is important to select appropriate choices for $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$.

For the finite population, $\boldsymbol{\varphi}^*$ satisfies

$$\dot{l}_U(\boldsymbol{\varphi}^*) = \sum_{i\in U}\left[\partial \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y})/\partial\boldsymbol{\varphi}^*\right] = 0,$$

where

$$\tilde{y} = \prod_{i=1}^{N} y_i^{1/N}.$$

For given $\boldsymbol{\varphi}^*$, let $\dot{l}_U(\boldsymbol{\varphi}^*)$, summation of the first derivative of the log-likelihood with respect to $\boldsymbol{\varphi}^*$, be a finite population parameter. We take a sample, and, by approximating $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y})$ for each unit i in the sample by $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w)$, we estimate the population total, $\dot{l}_U(\boldsymbol{\varphi}^*)$, by $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PML}})$:

$$\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PML}}) = \sum_{i\in s} w_i\left[\partial \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w)/\partial\boldsymbol{\varphi}^*\right]_{\boldsymbol{\varphi}^*=\hat{\boldsymbol{\varphi}}^*_{\mathrm{PML}}},$$

where

$$\tilde{y}_w = \prod_{i\in s} y_i^{w_i/\sum_{i\in s} w_i},$$

the weighted geometric mean of y's in the sample and $\hat{\boldsymbol{\varphi}}^*_{\mathbf{PML}}$ is the pseudo maximum likelihood estimator of $\boldsymbol{\varphi}^*$, satisfying $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PML}}) = 0$ (Wu and Sitter 2001). The PML estimator, $\hat{\boldsymbol{\varphi}}^*_{\mathbf{PML}} = (\hat{\boldsymbol{\beta}}^*_w, \hat{\lambda}_w, \hat{\sigma}^{*2}_{e,w})'$, can be obtained by a grid search method. That is, calculating and plotting the weighted log likelihood values,

$$\log L(\boldsymbol{\varphi}^*) = \sum_{i\in s} w_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w) \qquad (4)$$

against the set of values for λ will locate the PML estimate, $\hat{\lambda}_w$, of the transformation parameter. When we evaluate the log-likelihood function at each fixed value of λ in the sampling context, $\boldsymbol{\beta}^*$ and σ_e^{2*} are estimated by incorporating the sampling weights as:

$$\hat{\boldsymbol{\beta}}_w^{\,*} = \left(\sum_{i\in s} w_i \mathbf{x_i}\mathbf{x_i'}\right)^{-1}\left(\sum_{i\in s} w_i \mathbf{x}_i y_i^{*(\lambda)}\right),$$

$$\hat{\sigma}^{*2}_{e,w} = \sum_{i\in s} w_i (y_i^{*(\lambda)} - \mathbf{x_i'}\hat{\boldsymbol{\beta}}^*_w)^2 \Big/ \sum_{i\in s} w_i .$$

Since model $\mathbf{M_3}$ is of the interest, converting $\hat{\boldsymbol{\varphi}}^*_{\mathbf{PML}}$ that maximizes (4) back to $\hat{\boldsymbol{\varphi}}_{\mathbf{PML}}$ in the model $\mathbf{M_3}$ is necessary and $\hat{\boldsymbol{\beta}}_w = \tilde{y}_w^{\hat{\lambda}_w - 1}\hat{\boldsymbol{\beta}}^*_w$, $\hat{\sigma}^2_{e,w} = \tilde{y}_w^{2(\hat{\lambda}_w - 1)}\hat{\sigma}^{*2}_{e,w}$.

Since we plan to use a design-consistent estimator for the population total, it is reasonable to question why we need the sampling weights in estimating the model parameters. Sverchkov and Pfeffermann (2004) argue that incorporating weights can produce better estimates of model parameters when the model is correct in the population but wrong in the sample. That can happen when the probabilities or selections are correlated within the ε_i.

Let B^* and Λ be the first two components of the full-sample solution to the maximization of $\sum_U \log f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, and $\mathbf{B} = \tilde{y}^{\Lambda-1}\mathbf{B}^*$. This allows us to define $\boldsymbol{\theta}_N = (\mathbf{B}, \Lambda)$ as finite population parameters, irrespective of the validity of the model.

It can be shown that $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_w, \hat{\lambda}_w)$ is a design-consistent estimator of $\boldsymbol{\theta}_N$, that is,

$$\hat{\boldsymbol{\beta}}_{\mathbf{w}} \to \mathbf{B} \text{ in probability and } \hat{\lambda}_w \to \Lambda \text{ in probability}$$

under certain regularity conditions using arguments similar to Binder (1983), Wu (1999) and Wu and Sitter (2001). Here the probabilistic convergence is with respect to the sampling design.

2.3 Consistency property of $\hat{T}_{G\_BC}$ estimator

It is well-known that the $\hat{T}_{G\_L}$ estimator has the desirable property of design-consistency under mild conditions (Särndal *et al*. 1992). This means that the relative difference between the estimator and what it estimates will converge in probability to 0 as the sample grows arbitrarily large whether or not the working model on which it is based holds. That property can be maintained for the $\hat{T}_{G\_BC}$ estimator.

Define

$$\hat{T}_{D\_G\_BC} = \sum_{i\in U} g_i(\boldsymbol{\theta}_N) + \sum_{i\in s}(y_i - g_i(\boldsymbol{\theta}_N))/\pi_i,$$

and

$$\hat{T}_{G\_BC} = \sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) + \sum_{i\in s}(y_i - g_i(\hat{\boldsymbol{\theta}}_w))/\pi_i.$$

Theorem: Under the following assumptions, the Box-Cox-based GREG estimator $\hat{T}_{G\_BC}$ is design consistent for T, in the sense that $N^{-1}(\hat{T}_{G\_BC} - T) = O_p(1/\sqrt{n})$. Furthermore, the asymptotic variance of $\hat{T}_{G\_BC}$ is given by

$$AV_d(\hat{T}_{G\_BC}) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)(y_i - g_i(\boldsymbol{\theta}_N))(y_j - g_j(\boldsymbol{\theta}_N))/(\pi_i\pi_j),$$

which can be estimated by

$$\hat{V}_d(\hat{T}_{G\_BC}) = \sum_{i\in s}\sum_{j\in s}((\pi_{ij} - \pi_i\pi_j)/\pi_{ij})(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))/(\pi_i\pi_j). \quad (5)$$

Assumption 1: $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$;

Assumption 2: For each $\mathbf{x}_i$, $\partial g_i(\mathbf{t})/\partial \mathbf{t}$ is continuous in $\mathbf{t}$ and $|\partial g_i(\mathbf{t})/\partial \mathbf{t}| \le h(\boldsymbol{\theta})$ for $\mathbf{t}$ in a neighbourhood of $\boldsymbol{\theta}$;

Assumption 3: The Horvitz-Thompson estimators with the basic design weights for certain population totals are asymptotically normally distributed.

Assumption 4: For each $\mathbf{x}_i$, the second derivative of $g_i(\mathbf{t})$ with respect to $\mathbf{t}$ is continuous and bounded in the neighborhood of $\boldsymbol{\theta}$.

Proof: (see Appendix).

The proposed variance estimator in equation (5) is based on large sample approximations. For a given nominal level $1 - \alpha$, the usual confidence interval based on the normal approximation for the variance estimator gives approximately $100(1 - \alpha)\%$ coverage rate in repeated large samples. Unfortunately, in some cases, it has been observed that the coverage properties of this type of variance estimator can be poor for some choices of the assisted model for the $\hat{T}_{G\_L}$ estimators (Särndal 1982; Särndal, Swensson and Wretman 1989; *etc*). Theoretical and empirical studies on the coverage property of the proposed variance estimator need further investigated.

The $\hat{T}_{G\_BC}$ estimator is design-consistent for the finite population total T under the randomization approach, and the Box-Cox technique allows a reasonable transformation on the dependent variable to be automatically determined by the data from a large family of functions, and hence increased efficiency can be achieved.

3. A simulation study

The purpose of this simulation study is to evaluate the performance of different GREG estimators for a finite population total. In this simulation exercise, a finite population from the Australian Agricultural and Grazing Industries Survey (AAGIS) is generated. This survey data contains information on the number of cattle (y) and farm area (x) for each of the 431 farms.

We consider a finite population of size $N = 4{,}000$, generated from the following model:

$$\mathbf{M_5}: y_i^{(\lambda)} = (y_i^{\lambda} - 1)/\lambda = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where ε_i's are independent with approximate $N(0, \sigma^2)$, and x_i is the logarithm of a value generated from an exponential distribution with mean μ_x and standard error σ_x. In order to mimic a true situation, we choose $\lambda = 0.1$, $\beta_0 = 4.20$, and $\beta_1 = 2.66$ which are the estimates obtained by fitting the real survey data to the model $\mathbf{M_3}$. We set $\mu_x = 1{,}040$, $\sigma_x = 1{,}000$ to ensure $y_i > 0$ for almost all unit i. Strictly speaking, we have a truncated normal distribution of y, since all negative values of y generated are discarded. The effect of this is negligible since less than 0.1% of the generated y values need to be discarded. The same phenomenon was found by Taylor (1986).

Simulation is based on repeated sampling from the generated finite population. Two sampling designs are investigated: simple random sampling (SRS) and stratified SRS (STSRS). When a sample is selected by STSRS, unequal selection probabilities among different strata are applied. We define two strata using the boundary value: median of y values in the finite population. For stratum h of size N_h, a simple random sample of size n_h is selected. Define p_1 and p_2 selection probabilities for stratum 1 and stratum 2, respectively. We specify $p_1 = 2 \times p_2$. For fixed sample size n, $n_1 = N_1 \times p_1$, and $n_2 = N_2 \times p_2$.

We are interested in estimating the finite population total

$$T = \sum_{i \in U} y_i.$$

In this simulation study, we study the performances of $\hat{T}_{G\_BC}$ estimator, $\hat{T}_{G\_L}$, $\hat{T}_{G\_\mathrm{LOG}L}$, along with the design-based Horvitz-Thompson estimator ($\hat{T}_D$), where the subscripts "-L", "-LOGL", and "-BC" denote the underlying linear model, log-linear model, and Box-Cox model, respectively.

One thousand samples are selected from the simulated finite population for each of the sample size $n \in (30, 80, 150)$. Four estimators are produced for each selected sample. Estimator of the finite population transformation parameter Λ is also produced for each sample. For the purpose of comparison, two methods are used to estimate Λ. Let $\hat{\lambda}(\hat{\lambda}_w)$ be the OLS/ML (PML) estimators of Λ. Over all the 1,000 samples, we compute the empirical percentage relative biases (RelBias) and root mean square errors (rmse) to evaluate these estimators using the following formulae:

$$\mathrm{RelBias} = B^{-1} \sum_{b=1}^{B} (\hat{\varpi}_b - \varpi)/\varpi,$$

and

$$\mathrm{rmse} = \sqrt{B^{-1} \sum_{b=1}^{B} (\hat{\varpi}_b - \varpi)^2},$$

where B is the number of the replications in the Monte Carlo simulation and $\hat{\varpi}$ represents an arbitrary estimates of the finite population parameter ϖ.

4. Results

In Table 1 we present the RelBias and rmse of four estimators using different sampling designs with varying sample sizes when $\sigma = 0.5$. All four estimators give RelBias close to zero [maximum of the absolute values of RelBias (|RelBias|) in Table 1 is less than 0.01]. Among them, $\hat{T}_{G\_BC}$ has the smallest |RelBias| and rmse over different sampling sizes and sampling designs. Therefore, the Box-Cox technique protects $\hat{T}_{G\_BC}$, which achieves improvement in efficiency compared to other GREG estimators.

The RelBias and rmse under the same conditions when $\sigma = 1$ and $\sigma = 2$ are also investigated. Figure 1 presents the rmse for the three GREG estimators ($\hat{T}_{G\_L}$, $\hat{T}_{G\_\mathrm{LOG}L}$, and $\hat{T}_{G\_BC}$) using different sampling designs. We note that $\hat{T}_{G\_BC}$ consistently has the smallest rmse when $\sigma = 0.5$ and 1. Thus, a robust model chosen by the Box-Cox method reduces the rmse, especially when the model is appropriate with small σ.

Table 1
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes ($\lambda = 0.1$)

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_\mathrm{LOG}L}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	4.37	-9.82	5.46	3.62
	$n = 80$	1.43	-3.67	1.24	0.65
	$n = 150$	-2.60	-1.24	1.22	0.53
rmse($\times 10^{7}$)	$n = 30$	7.17	3.54	2.02	1.94
	$n = 80$	4.26	2.09	1.18	1.09
	$n = 150$	3.20	1.58	0.88	0.79
Stratified simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	6.01	-5.82	3.84	1.92
	$n = 80$	9.93	-1.01	3.04	0.98
	$n = 150$	1.75	-1.85	1.06	0.42
rmse($\times 10^{7}$)	$n = 30$	5.63	3.67	2.29	2.11
	$n = 80$	3.51	2.31	1.43	1.28
	$n = 150$	2.49	1.59	1.01	0.90

[1] $\hat{T}_D$: the design-based Horvitz-Thompson estimator;
[2] $\hat{T}_{G\_L}$: the GREG estimator with the underlying linear model;
[3] $\hat{T}_{G\_\mathrm{LOG}L}$: the GREG estimator with the underlying log-linear model;
[4] $\hat{T}_{G\_BC}$: the GREG estimator with the underlying Box-Cox model.

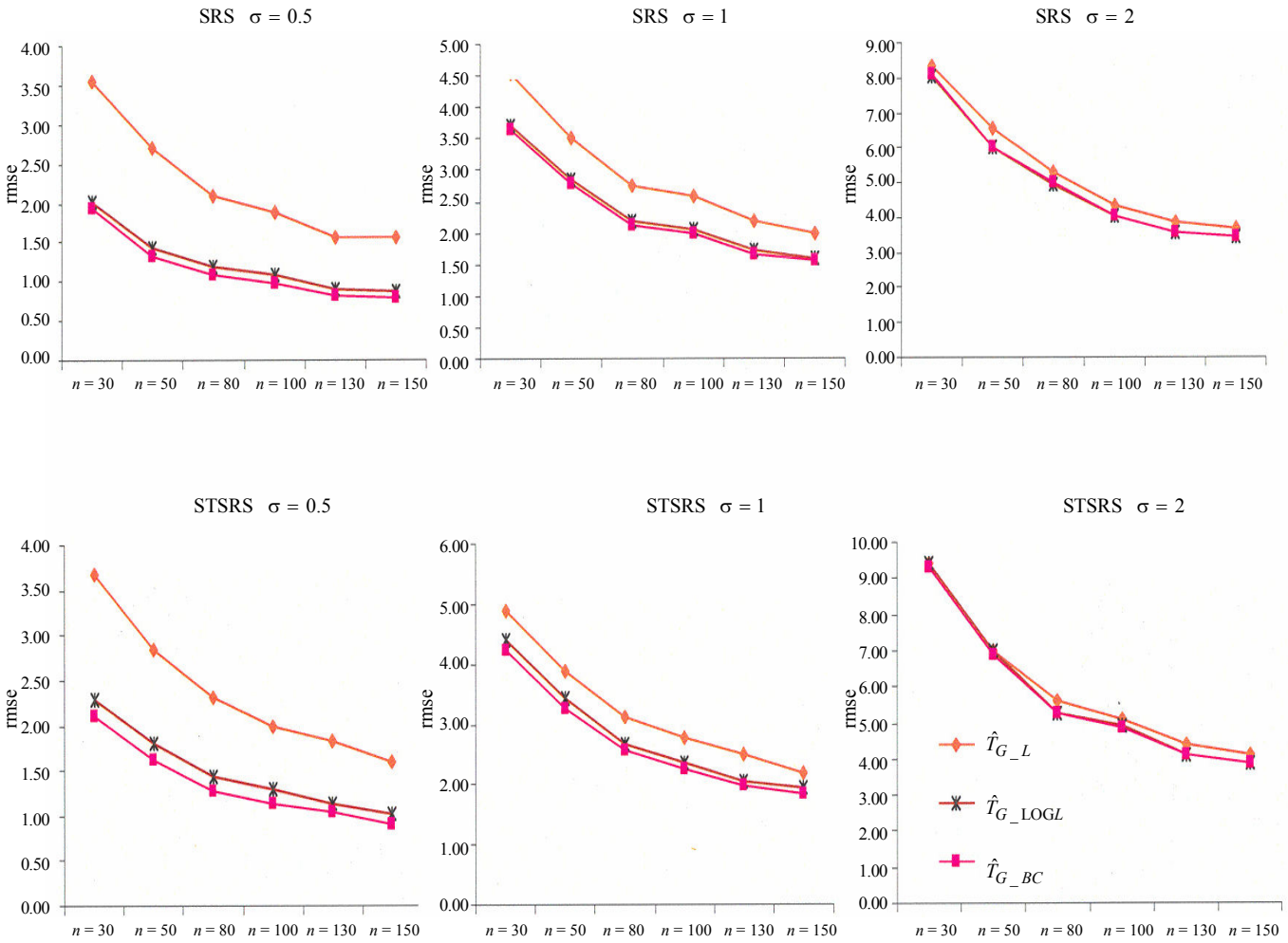

Figure 1 Comparison of root mean square error of $\hat{T}_{G\_L}$, $\hat{T}_{G\_LOGL}$, and $\hat{T}_{G\_BC}$ with varying sampling designs, sample sizes, and standard deviations

We may also be interested in how $\hat{T}_{G\_L}$ performs compared to $\hat{T}_{G\_BC}$ when $\lambda = 1$, the situation favoring $\hat{T}_{G\_L}$. From Table 2, we can see that $\hat{T}_{G\_BC}$ is quite comparable to $\hat{T}_{G\_L}$, especially when sample size is large. This result implies we don't have much loss in RelBias and rmse by using $\hat{T}_{G\_BC}$ when $\hat{T}_{G\_L}$ should be used.

In order to better assess the robustness and the improvement in efficiency of the $\hat{T}_{G\_BC}$ versus $\hat{T}_{G\_L}$ and $\hat{T}_{G\_LOGL}$, a finite population of size $N = 4{,}000$ is generated from the model not in agreement with the Box-Cox model $\mathbf{M_5}$, but

$$\mathbf{M_6}: y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + z_i + \varepsilon_i,$$

where $z_i = x_i^2$. The same x_i's and parameter values as specified in Section 3 are used to generate y_i's. The same four estimators are studied based on the new finite population $\{(x_i, y_i)$'s for $i = 1, \ldots, 4{,}000\}$. This is the situation not ideal for any of the GREG estimators. The results are shown in Table 3. The advantage of using $\hat{T}_{G\_BC}$ is obvious in terms of the Relbias and rmse.

Table 4 presents the RelBias and rmse of $\hat{\lambda}$ and $\hat{\lambda}_w$ for STSRS sampling with varying sample sizes and σ's based on population values generated from $\mathbf{M_5}$. Since we stratified on the y-variable, using the weights should have an impact, at least on the bias of the parameter estimate. When σ is small, that is, when the simulated data are well fitted to the assumed model, $\hat{\lambda}_w$ gives RelBias closer to zero, but $\hat{\lambda}_w$ and $\hat{\lambda}$ perform equally well in terms of the rmse. When σ is large, however, $\hat{\lambda}_w$ consistently gives smaller absolute values of RelBias and rmse, as compared to $\hat{\lambda}$, although the rmse's remain close. Indeed, when it come to estimating T, neither approach has the advantage in terms of empirical bias or root mean squared error no matter the sample size or the setting for σ (not shown). This may be because the estimator for T is model-biased no matter how well λ is estimated.

Table 2
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes ($\lambda = 1$)

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_LOGL}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	0.31	0.16	0.63	0.24
	$n = 80$	-0.37	-0.09	0.03	-0.09
	$n = 150$	-0.25	0.03	0.10	0.02
rmse($\times 10^{7}$)	$n = 30$	20.91	3.51	3.98	3.63
	$n = 80$	12.22	2.12	2.36	2.13
	$n = 150$	8.98	1.57	1.75	1.57
Stratified simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	0.51	0.27	0.56	0.22
	$n = 80$	3.97	-0.11	-0.11	-0.13
	$n = 150$	-0.23	0.04	0.06	0.04
rmse($\times 10^{7}$)	$n = 30$	12.54	3.79	4.20	3.91
	$n = 80$	8.39	2.27	2.61	2.29
	$n = 150$	5.48	1.67	1.90	1.67

[1] $\hat{T}_D$: the design-based Horvitz-Thompson estimator;
[2] $\hat{T}_{G\_L}$: the GREG estimator with the underlying linear model;
[3] $\hat{T}_{G\_LOGL}$: the GREG estimator with the underlying log-linear model;
[4] $\hat{T}_{G\_BC}$: the GREG estimator with the underlying Box-Cox model.

Table 3
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes (y values generated from a model $\mathbf{M_6}$)

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_LOGL}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	-16.89	-54.56	28.81	-2.17
	$n = 80$	-5.65	-23.15	11.76	-1.43
	$n = 150$	-13.78	-13.78	10.73	-0.76
rmse($\times 10^{11}$)	$n = 30$	30.08	24.68	12.11	2.87
	$n = 80$	17.95	13.85	7.25	1.78
	$n = 150$	13.60	10.29	5.50	1.33
Stratified simple random sampling					
RelBias($\times 10^{-3}$)	$n = 30$	1.11	-36.13	31.95	-7.87
	$n = 80$	5.59	-18.13	13.11	-3.26
	$n = 150$	-2.79	-7.30	7.33	-1.43
rmse($\times 10^{11}$)	$n = 30$	34.10	27.15	14.19	4.37
	$n = 80$	19.61	15.93	8.49	2.61
	$n = 150$	14.60	12.19	6.62	2.02

[1] $\hat{T}_D$: the design-based Horvitz-Thompson estimator;
[2] $\hat{T}_{G\_L}$: the GREG estimator with the underlying linear model;
[3] $\hat{T}_{G\_LOGL}$: the GREG estimator with the underlying log-linear model;
[4] $\hat{T}_{G\_BC}$: the GREG estimator with the underlying Box-Cox model.

Table 4
Relative biases and root mean square error of $\hat{\lambda}^1$ and $\hat{\lambda}_w^2$ for STSRS sampling with varying sample sizes and standard deviations.

	$\sigma = 2$		$\sigma = 1$		$\sigma = 0.5$	
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$
Relative biases						
$n = 30$	-0.58	0.13	-0.28	0.10	-0.16	-0.01
$n = 80$	-0.42	0.14	-0.19	0.11	-0.13	-0.02
$n = 150$	-0.39	0.10	-0.16	0.09	-0.10	-0.01
Root mean square error						
$n = 30$	0.14	0.12	0.11	0.11	0.07	0.07
$n = 80$	0.08	0.07	0.06	0.06	0.04	0.04
$n = 150$	0.06	0.05	0.05	0.04	0.03	0.03

[1] Estimator $\hat{\lambda}$ is obtained using ordinary least square method/maximum likelihood method;
[2] Estimator $\hat{\lambda}_w$ is obtained by pseudo-maximum likelihood method.

5. Concluding remarks

In this article, we have proposed a generalized regression estimator of a finite population total based on the Box-Cox transformation technique under a general unequal probability sampling design. The proposed estimator, being design-consistent, maintains the robustness property of GREG even if the underlying model fails. In many situations, some version of the model in $\mathbf{M_3}$ will at least provide a useful approximation of dependent-variable behavior. The Box-Cox technique allows a reasonable-fitting transformation on the dependent variable to be automatically determined by the data. The robustness and efficiency of the proposed estimator were evaluated analytically and via Monte Carlo simulations.

When comparing a GREG based on an underlying linear model ($\hat{T}_{G\_L}$) to one based on a Box-Cox model ($\hat{T}_{G\_BC}$), we should remember that $\hat{T}_{G\_L}$ doesn't require complete auxiliary information. Moreover, it can produce a single set of weights usable for all variables of interest, unlike the $\hat{T}_{G\_BC}$. To achieve higher efficiency, however, both estimators usually require different weights for different variables of interest, because each study variable is best fit by its own working model. The $\hat{T}_{G\_BC}$ can provide even more efficiency than the $\hat{T}_{G\_L}$ but at the cost of requiring complete information about the x-variables. Such information, although rarely available in North American household surveys, is often available in business surveys.

Surveys are rarely conducted to measure a single variable of interest. The question is how to estimate the finite totals for mutually exclusive and exhaustive subpopulations such that those estimates will add up to the estimate of the finite total for the entire population. We need to take special care of this problem since Box-Cox estimators are not linear in y. Such estimates can be obtained using a standard

benchmarking tool outlined in Li and Lahiri (2007). Another approach is to use model calibration (Wu and Sitter 2001). Treating the predictions $\hat{y}_{i,w}$ from equation (3) as the auxiliary variable into a design-consistent linear regression estimator, such as

$$\hat{T}_{G\_BC} = \sum_{i\in s} y_i/\pi_i + \left(\sum_{i\in U}\mathbf{z}_i' - \sum_{i\in s}\mathbf{z}_i'/\pi_i\right)\mathbf{d},$$

where

$$\mathbf{d} = \left(\sum_{i\in s} w_i\,\mathbf{z}_i\,\mathbf{z}_i'\right)^{-1}\left(\sum_{i\in s} w_i\,\mathbf{z}_i\,y_i\right), \text{ and } \mathbf{z}_i = (1\ \ \hat{y}_{i,w})',$$

will produce a set of calibration weights

$$\left(w_{i,c} = \left[1 + \left(\sum_U \mathbf{z}_i' - \sum_S \mathbf{z}_i'/\pi_i\right)\left(\sum_S w_j\,\mathbf{z}_j\,\mathbf{z}_j'\right)^{-1}\mathbf{z}_i\right]w_i\right)$$

that can be used generally. Moreover, we can, in principle, incorporate more than one set of predictors as auxiliary variables, either for different variables of interest or the same variable of interest broken into several subpopulations. This is a fruitful area for future research.

Several other extensions of our current method merit further exploration. We did not consider here the possibility of unit model errors having a complex correlation structure. Although a design-consistent estimator obtains using our methods when such a structure exists but is ignored, the efficiency of the estimator likely suffers. It will be interesting to investigate whether allowing certain correlation structures in the data can make the estimation procedure more efficient.

In this article, we only transformed y-variable by the Box-Cox technique. In future, this method can be extended to different functional forms of the Box-Cox transformation.

The variance estimator of the $\hat{T}_{G\_BC}$ estimator proposed in this article is based on large sample approximation. Some studies showed poor performance for this type of variance estimator for some choices of the assisted models for the $\hat{T}_{G\_L}$. Theoretical and empirical studies on the coverage property of the proposed variance estimator need further investigation.

Acknowledgements

I wish to thank the Editor and three referees for a number of constructive suggestions which led to substantial improvement of the original manuscript. I thank Dr. Phil Kott for his generous line editing of the text and Dr. Alan Dorfman for supplying the survey data used in the paper. The author is grateful to her Ph.D. advisor Professor Partha Lahiri, supervisory committee members Professors Katherine Abraham, Wolfgang Jank, Fritz Scheuren, Paul Smith and Mr. Paul D. Williams for their generous guidance throughout her dissertation research.

Appendix

Write

$$\hat{T}_{G\_BC} = \sum_{i\in s} y_i/\pi_i + \left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} g_i(\hat{\boldsymbol{\theta}}_w)/\pi_i\right),$$

by Taylor Series expansion, with assumption (2) we have

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + (\partial g_i(\mathbf{t})/\partial\mathbf{t})\,|_{t=\boldsymbol{\theta}^*}\,(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

where $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ or $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$, and $(\partial g_i(\mathbf{t})/\partial\mathbf{t})$ is a row vector.

By assumptions (1) and (2),

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n})$$

$$N^{-1}\sum_{i\in s}\pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s}\pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Also note that by assumption (3),

$$N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) = N^{-1}\sum_{i\in s}\pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Therefore,

$$N^{-1}\left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s}\pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w)\right) = O_p(1/\sqrt{n}).$$

Also, by assumption (3),

$$N^{-1}\left(\sum_{i\in U} y_i - \sum_{i\in s}\pi_i^{-1} y_i\right) = O_p(1/\sqrt{n}).$$

Therefore, $N^{-1}(\hat{T}_{G\_BC} - T) = O_p(1/\sqrt{n})$, *i.e.*, $\hat{T}_{G\_BC}$ converges in probability to T with the order of $O_p(1/\sqrt{n})$.

In addition, with assumption (4), a second-order Taylor series approximation to $g_i(\hat{\boldsymbol{\theta}}_w)$ can be expanded as:

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + (\partial g_i(\mathbf{t})/\partial\mathbf{t})'\,|_{t=\boldsymbol{\theta}_N}$$
$$(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N)'\left(\partial^2 g_i(\mathbf{t})/(\partial\mathbf{t}\partial\mathbf{t}')\right)|_{t=\boldsymbol{\theta}^*}\,(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

where $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ or $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$. It follows from assumptions (1) and (4) that

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N)$$
$$+ N^{-1}\sum_{i\in U}(\partial g_i(\mathbf{t})/\partial\mathbf{t}')\,|_{t=\boldsymbol{\theta}_N}\,(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}),$$

$$N^{-1}\sum_{i\in s}\pi_i^{-1}g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s}\pi_i^{-1}g_i(\boldsymbol{\theta}_N)$$

$$+ N^{-1}\sum_{i\in s}\pi_i^{-1}\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}'\right)|_{t=\boldsymbol{\theta}_N}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}).$$

By assumptions (1) and (3), $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$ and

$$N^{-1}\sum_{i\in s}\pi_i^{-1}\partial g_i(\mathbf{t})/\partial \mathbf{t}\ |_{t=\boldsymbol{\theta}_N} = N^{-1}\sum_{i\in U}\partial g_i(\mathbf{t})/\partial \mathbf{t}\ |_{t=\boldsymbol{\theta}_N}$$

$$+\ O_p(1/\sqrt{n}).$$

Hence,

$$N^{-1}\left(\sum_{i\in U}g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s}\pi_i^{-1}g_i(\hat{\boldsymbol{\theta}}_w)\right) =$$

$$N^{-1}\left(\sum_{i\in U}g_i(\boldsymbol{\theta}_N) - \sum_{i\in s}\pi_i^{-1}g_i(\boldsymbol{\theta}_N)\right) + O_p(n^{-1}).$$

Therefore,

$$\hat{T}_{G\_BC} = \sum_{i\in s}y_i/\pi_i$$

$$+\left(\sum_{i\in U}g_i(\boldsymbol{\theta}_N) - \sum_{i\in s}\pi_i^{-1}g_i(\boldsymbol{\theta}_N)\right) + O_p(N/n).$$

The asymptotic design-variance of $\hat{T}_{AG}$ is:

$$AV_d(\hat{T}_{G\_BC})$$

$$\approx \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)(y_i - g_i(\boldsymbol{\theta}_N))(y_j - g_j(\boldsymbol{\theta}_N))/(\pi_i\pi_j),$$

which can be estimated by

$$\hat{V}_d(\hat{T}_{G\_BC})$$

$$\approx \sum_{i\in s}\sum_{j\in s}(\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\,(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))/(\pi_i\pi_j).$$

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *International Statistical Review*, 51, 279-92.

Bickel, P.J., and Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296-311.

Box, G.E., and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

Boylan, T.A., Cuddy, M.P. and O'Muircheartaigh, I.G. (1980). The functional form of the aggregate import demand equation: A comparison of three european economies. *Journal of international economics*, 10, 561-566.

Boylan, T.A., Cuddy, M.P. and O'Muircheartaigh, I.G. (1982). Import demand equations: An application of a generalized Box-Cox methodology. *International Statistical Review*, 50, 103-112.

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

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

Carrol, R.J., and Ruppert, D. (1988). *Transformations and weighting in Regression*. London: Chapman and Hall.

Cassel, C.M., Särndal, C.-E. and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chambers, R.L., and Dorfman, A.H. (2003). Transformed variables in survey sampling. Technical Report.

Chen, G., and Chen, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.

Davison, C.W., Arnade, C.A. and Hallahan, C.B. (1989). Box-Cox estimation of U.S. soyabean exports. *Journal of Agricultural Economics Research*, 41, 8-15.

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

Estevao, V., Hidiroglou, M. and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, 11, 181-204.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5–25.

Fuller, W.A., Loughin, M. and Baker, H. (1994). Regression weighting in the presence of nonreponse with application to the 1987-1988 nationwide food consumption survey. *Survey Methodology*, 20, 75-85.

Gemmil, G. (1980). Using the Box-Cox Form for Estimating Demand: A Comment. *The review of economics and statistics*, 62, 147-148.

Gurka, M. (2004). The Box-Cox transformation in the general linear mixed model for longitudinal data. Ph.D. Dissertation.

Gurka, M. (2006). Expanding the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society*, Series A, 169, 273-288.

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

Jayasuriya, B., and Valliant, R. (1996). An application of restricted regression estimation to post-stratification in a household survey. *Survey Methodology*, 22, 127-137.

Jiang, J., and Lahiri, P. (2006). Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

John, N.R., and Draper, J.A. (1980). An alternative family of transformations. *Applied Statistics*, 29, 190–197.

Karlberg, F. (2000). Survey estimation for highly skewed population in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.

Khan, M.S., and Ross, K.Z. (1977). The functional form of aggregate import demand. *Journal of International Economics*, 7, 149-160.

Korn, E., and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24, 193-201.

Li, Y., and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664-673.

Miner, A.G. (1982). The contribution of weather and technology to U.S. soybean yield. *Unpublished Dissertation*. University of Minnesota.

Montanari, G.E., and Ranalli, M.G. (2003). Nonparametric model calibration estimation in survey sampling. Manuscript.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.

Newman, P. (1977). Malaria and Mortality. *Journal of the American Statistical Association*, 72, 257-263.

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

Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.

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

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

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

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

Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *Journal of the Royal Statistical Society*, Series B, 33, 307-311.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys.* Chichester: Wiley.

Spitzer, J.J. (1976). The demand for money, the liquidity trap and functional forms. *The International Economics Review*, 17, 220-227.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79-92.

Taylor, M.J.G. (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114-118.

Tukey, J.W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics*, 28, 601-632.

Wu, C. (1999). The effective use of complete auxiliary information from survey data. Ph.D. dissertation, Simon Fraser University, Canada.

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

Zarembka, P. (1968). Functional form in the demand for money. *Journal of the American Statistical Association*, 63, 502-511.

Zheng, H., and Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30, 209-218.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 91-103
Statistics Canada, Catalogue No. 12-001-X

The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data

Cédric Béguin and Beat Hulliger [1]

Abstract

With complete multivariate data the BACON algorithm (Billor, Hadi and Vellemann 2000) yields a robust estimate of the covariance matrix. The corresponding Mahalanobis distance may be used for multivariate outlier detection. When items are missing the EM algorithm is a convenient way to estimate the covariance matrix at each iteration step of the BACON algorithm. In finite population sampling the EM algorithm must be enhanced to estimate the covariance matrix of the population rather than of the sample. A version of the EM algorithm for survey data following a multivariate normal model, the EEM algorithm (Estimated Expectation Maximization), is proposed. The combination of the two algorithms, the BACON-EEM algorithm, is applied to two datasets and compared with alternative methods.

Key Words: Forward search method; Outlier detection; Multivariate data; Missing value; Sampling; Robustness; EM-algorithm.

1. Introduction

The problem underlying the methods presented in this article is a sample survey on quantitative data like sales of different products where missing values and outliers occur. Often in the editing phase of the survey outliers are detected by inspection of individual questionnaires or by univariate outlier detection methods. However, there are few systematic methods which allow multivariate outlier detection in incomplete survey data.

Outlier detection is an important aspect of statistical data editing. Undetected outliers may have a large and undesirable impact on survey results. Most existing outlier-detection methods are designed for complete univariate or bivariate data. However, real outliers in survey data are often multivariate in nature. The problem of outliers becomes much more difficult in three or more dimensions than in one dimension or two. While an outlier can only be very small or very large in one dimension (at least for unimodal distributions) in higher dimensions the issue of the "direction" of the outlier becomes more and more important. Outliers may be quite close to the bulk of the data or to a model if the distance is measured in a Euclidean metric because this metric only checks the axis directions. However, if a metric appropriate to the correlation structure of the bulk of the data is used the outlier may be far away. Thus in higher dimensions the form of the point cloud of the bulk of the data must be well reflected in the metric used to detect outliers.

Outlier detection needs a model for the bulk of the data to be able to distinguish observations which are not fitted well by the model. Thus outlier detection is inherently tied to models and their robust estimation. In a sampling context the model should be appropriate for the bulk of the population and not only for the bulk of the sample. Therefore, the sample design should be taken into account when detecting outliers in sample survey data. The discussion on the role of sampling weights is taken up again in Sections 1 and 5.

Survey data often contains missing values. Outlier detection with missing items must estimate the model for the bulk of the data taking into account the missingness. This estimation under missing values will be based on the relationship among the observed and missing variables. The relationship must be modeled robustly to protect it from outliers. If an observation would be classified as an outlier based on complete information but the values causing the outlyingness are missing then the outlier will not show up compared with a robust model. Therefore it will be difficult to detect an outlier which is outlying only in its missing values. This is analogue to the conception of missingness at random (MAR) (Little and Rubin 1987): We need information in the observed values to infer that an observation is an outlier. We may call this situation "outlying at random". We can formalize it by stating that the outlier mechanism does not depend on unobserved data, which includes unobserved true values of the outlier in case the outlier is an error. However, for outlier detection this condition is too strict because we may be able to detect outliers in observed values even if the mechanism depends on unobserved values. This is possible because the model must hold for the bulk of the data only and not for the outliers. If the observed values of the outlier deviate enough from the model the outlier will be detected. However, when it comes to imputation of true values for nominated outliers we are in the same situation as for missing values. If the

1. Cédric Béguin, University of Neuchâtel, 2010 Neuchâtel, Switzerland; Beat Hulliger, University of Applied Sciences Northwestern Switzerland, 4600 Olten, Switzerland. E-mail: beat.hulliger@fhnw.ch.

outlier mechanism, conditionally on the observed values, still depends on the true unobserved values of the outlier we cannot estimate a model for the unobserved values. In this article we use imputation only as an ad hoc device for better outlier detection. Nevertheless we assume that, conditionally on the observed data, the non-response mechanism and the outlier-mechanism are independent and that both mechanisms do not depend on unobserved data.

In a workpackage of the EUREDIT project on "The development and evaluation of new methods for editing and imputation" (EUREDIT 2003) the authors developed outlier detection methods which cope with this difficult set-up: multivariate incomplete sample survey data. Two of these methods, Transformed Rank Correlations and the Epidemic Algorithm are presented in (Béguin and Hulliger 2004). The third method, BACON-EEM, is presented here.

In this article we concentrate on outlier detection. The scenario we have in mind is that once an outlier is detected, either it may be checked and treated manually or it may be treated by imputation. Robust estimation would replace both detection and imputation but is less adapted to the practice of official statistics. We do not distinguish between representative and non-representative outliers (Chambers 1986) since both types of outliers have to be detected, though they may have to be treated differently.

For complete data the existing multivariate methods can be classified into two major families. Many methods suppose that the data follow some elliptical distribution and try to estimate robustly the center and the covariance matrix. Then they use a corresponding Mahalanobis distance to detect outliers. The second class of methods does not rely on a distributional assumption but uses some measure of data-depth (see Liu, Parelius and Singh 1999, for a review) to be used as an outlyingness measure. The second family is at first sight more appealing, but, unfortunately, it often fails to yield methods computationally feasible with large datasets.

Many robust estimators of the covariance matrix have been reported in the literature. M-estimators (Huber 1981; Maronna 1976) have the advantage of being relatively simple to compute with a straightforward iteration from a good starting point (Rocke and Woodruff 1993). But their breakdown point - *i.e.*, the smallest fraction of the data whose arbitrary modification can carry an estimator beyond all bounds - is at most $1/(p+1)$ where p is the dimension of the data (Donoho 1982; Maronna 1976; Stahel 1981). This handicap is important when dealing with data from official statistics, which is often high dimensional. Many other affine equivariant robust estimators, *i.e.*, estimators which transform coherently when the data is transformed linearly, were studied by (Donoho 1982) but all have breakdown points of at most $1/(p+1)$. Other approaches ended up with affine equivariant high breakdown point estimators, *e.g.* the Stahel-Donoho (SD) estimator (Stahel 1981; Donoho 1982) or the Minimum Covariance Determinant (MCD) estimators (Rousseeuw 1985; Rousseeuw and Leroy 1987), but had the disadvantage of being computationally expensive. An approach of Gnanadesikan and Kettenring, using a componentwise construction of the covariance matrix, sacrificed affine equivariance but gained simplicity and speed. This approach has been re-actualized in (Maronna and Zamar 2002) and in one of the methods presented in (Béguin and Hulliger 2004), called Transformed Rank Correlations (TRC). TRC calculates an initial matrix of bivariate Spearman Rank correlations. To ensure positive definiteness of the covariance matrix the data is transformed into the space of eigenvectors of the initial matrix. The coordinatewise medians and median absolute deviations in this new space are then backtransformed into the original space to obtain an estimate of the center and a positive definite covariance matrix.

Another idea from (Gnanadesikan and Kettenring 1972) is related to the so-called forward search methods, which are closely related to the method proposed in this paper. These so called forward search methods are based on the concept of "growing a good subset of observations". By "good subset" one means a subset free or almost free of outliers. The idea is to start with a small subset of the data and then to add non-outlying observations until no more non-outliers are available.

The idea of a forward search algorithm was first suggested in (Wilks and Gnanadesikan 1964) and described in detail in (Gnanadesikan and Kettenring 1972). The articles of (Hadi 1992) and (Atkinson 1993) demonstrated the efficiency of such methods. In both articles the "good subset" grows one point at a time using Mahalanobis distances to rank the observations. Then research concentrated on developing faster and more sophisticated methods based on the same idea. The last two and most efficient were developed in (Billor, Hadi and Vellemann 2000) and (Kosinski 1999). These algorithms were compared in (Béguin 2002) and the BACON algorithm (Billor *et al.* 2000) turned out to be the most robust and fastest forward search method with complete multivariate normal data. In particular the breakdown point turned out to be high in practical applications. Also when comparing with other Mahalanobis type methods the performance of BACON on complete data is very good (Béguin and Hulliger 2003).

None of the above methods is designed to deal with incomplete data stemming from surveys, *i.e.*, with missing values and sampling weights. The first article to address the problem of multivariate outlier detection in incomplete data is (Little and Smith 1987). The authors propose Mahalanobis distances to detect outliers, with robust

estimations of center and scatter obtained by the ER algorithm. The ER algorithm replaces the maximum-likelihood estimator in the maximization step of the EM algorithm (Dempster, Laird and Rubin 1977) by a robust one-step *M*-estimator. However, the starting point of the ER algorithm is the classical non-robust mean and covariance and therefore the breakdown point of the ER algorithm is 0. In other words even one single outlier can carry the estimator beyond any limit. To correct the low breakdown point of that algorithm (Cheng and Victoria-Feser 2000) used an MCD algorithm for the maximization step of the EM-algorithm. However, the combination of the iterative procedures of MCD and EM makes the computation for large datasets too slow for practical applications. Moreover the introduction of sampling weights is not straightforward.

The TRC algorithm in (Béguin and Hulliger 2004) uses robust linear regression imputations by the best univariate predictor to cope with missing values. The Spearman rank correlations are expressed as functionals of the empirical distribution function of the sample to obtain estimates for the Spearman rank correlations in the population.

The BACON algorithm is based on the multivariate normal distribution and thus the EM algorithm for multivariate normal data was chosen to impute missing values within the BACON iterations. To take into account the sampling aspect, the estimates of the BACON algorithm have to be replaced by Horvitz-Thompson type estimators and a special version of the EM algorithm is developed where the expectations on the population level are estimated from the sample. Section 2 sets up the notation, recalls quickly the BACON algorithm and presents its adaptation to sampling weights. Section 3 introduces the Estimated-EM (EEM) algorithm and Section 1 discusses the adaptation of the Mahalanobis distance to missing values. Section 4 explains how BACON and EEM are merged in an efficient way to become the BACON-EEM algorithm. Section 5 shows the application of BACON-EEM to two datasets. The results are compared to the competitor methods, Transformed Rank Correlations, developed in (Béguin and Hulliger 2004), the ER-algorithm and a baseline algorithm which uses MCD after non-robust imputation based on the EM-algorithm.

2. The BACON algorithm

The BACON algorithm is presented in (Billor *et al.* 2000). Two versions are described: one for multivariate data in general and one for regression data. Only the first case will be considered here.

The data are stocked in a matrix X of n rows (observations $x_1, ..., x_n$) and p columns (variables $x^1, ..., x^p$). We assume that the bulk of the data is unimodal and roughly elliptical symmetric. The coordinatewise mean (resp. covariance matrix) computed on X is denoted by m_X (resp. C_X). The squared Mahalanobis distance of a point y based on m_X and C_X is $\mathrm{MD}_X^2(y) = (y - m_X)^\top C_X^{-1}(y - m_X)$. If the mean and covariance are calculated only on a subset G of the data then we denote them m_G and C_G with corresponding Mahalanobis distance MD_G.

The first step of the algorithm is the choice of an initial subset G of "good data". Two versions are proposed in the literature. The first version simply selects the cp points with smallest Mahalanobis distances $\mathrm{MD}_X(x_i)$, $i \in \{1, ..., n\}$, with c being an integer chosen by the data analyst. It may be set to $c = 3$ by default. The second version selects the cp points with smallest Euclidean distances from the coordinatewise median, with c as before. The second version is more robust but it looses affine equivariance. Other starting points than the coordinatewise median might be considered like a spatial median. In this article we concentrate on the second version of the basic good subset. In both versions if C_G is singular then the basic subset is increased by adding observations with smallest distances until C_G has full rank. Then an iterative process starts.

Denote by $\chi^2_{p,\beta}$ the $1-\beta$ percentile of the χ^2 distribution with p degrees of freedom and by $|G|$ the number of elements in the set G. The steps of the BACON algorithm are:

1. Compute the squared Mahalanobis distances $\mathrm{MD}_G^2(x_i)$ for $i \in \{1, ..., n\}$;
2. Define a subset G' including all points with $\mathrm{MD}_G^2(x_i) < c_{npr}\, \chi^2_{p,\alpha/n}$, where $c_{npr} = c_{np} + c_{hr}$ is a correction factor with $c_{np} = 1 + (p+1)/(n-p) + 1/(n-h-p)$, $c_{hr} = \max\{0, (h-r)/(h+r)\}$, $h = \lceil (n+p+1)/2 \rceil$ and $r = |G|$.
3. If $G' = G$ then stop, else set G to G' and go to Step 1.

Note that the correction factor c_{npr} is close to 1 for large n. The observations that are not contained in the final G are declared outliers. Alternatively a threshold for the Mahalanobis distance, above which observations are nominated outliers, may be chosen by inspecting the distribution of the Mahalanobis distance.

The computing effort required by the BACON algorithm depends on the configuration of the data. Compared with other algorithms it is small and in particular this effort grows slowly with increasing sample size (see also Section 5). This makes the BACON method particularly well suited for large datasets.

Note that the original selection criterion of Step 2 is designed for a multivariate normal distribution, which

implies that the squared Mahalanobis distances follow asymptotically a χ^2 distribution with p degrees of freedom. Suppose all points follow a multivariate normal distribution and that the Mahalanobis distance is computed using the sample mean and covariance matrix. The test $\mathrm{MD}_X^2(x_i) > \chi^2_{p,\alpha}$ declares about 100α percent of the points as outliers. Instead of α we often use α/n. Using Bonferroni inequalities one can show that under normality the test with level α/n will declare no outlier with probability larger than $1-\alpha$ (*i.e.*, $P(\mathrm{MD}_X^2(x_i) < \chi^2_{p,\alpha/n}, \forall i \in \{1,...,n\}) \geq 1-\alpha$). The test with α/n very rarely detects points that are not outliers but it also reduces its sensitivity to close outliers when n becomes large. One may also want to run the method with both types of the test level and compare the results.

2.1 Adaptation to sampling weights

For the sampling context we use the following notation. The data stem from a random sample s of the finite population U with N elements. The sample of size n is drawn with the sample design $p(s)$ and the first order inclusion probabilities are denoted $\pi_i = \sum_{s|i\in s} p(s)$. The weights will be the inverse of the inclusion probabilities of the observations $w_i = 1/\pi_i$ such that the Horvitz-Thompson estimator of the population total, $\sum_{i\in U} x_i$, is $\sum_s w_i x_i = \sum_{i=1}^n w_i x_i$. Furthermore it is assumed that $\sum_s w_i \approx N$. The mean m_X and the covariance matrix C_X may be estimated by the Hájek estimators

$$\hat{m}_X = \frac{\sum_s w_i x_i}{\sum_s w_i}$$

and (1)

$$\hat{C}_X = \frac{\sum_s w_i (x_i - \hat{m}_X)(x_i - \hat{m}_X)^\top}{\sum_s w_i}.$$

The sample estimate of the median is defined as in (Béguin and Hulliger 2004): let x_u^k be the smallest value such that $\sum_s w_i 1_{x\leq x_u^k}(x_i^k) \geq 0.5\sum_s w_i$ and x_v^k the smallest value such that $\sum_s w_i 1_{x\leq x_v^k}(x_i^k) > 0.5\sum_s w_i$, then the estimate is given by

$$\widehat{\mathrm{med}}_X = (w_u x_u^k + w_v x_v^k)/(w_u + w_v). \quad (2)$$

To adapt the BACON algorithm to sampling the initial subset is selected using Hájek estimators $\hat{m}_X$ and $\hat{C}_X$ or the median $\widehat{\mathrm{med}}_X$. For the iterative process, denote by s_G the selected "good observations" of the sample. These observations are representatives of a "virtual good subset" G of the whole population with estimated size $\hat{r} = \sum_{s_G} w_i$. The mean and covariance matrix of this subset are estimated by the Hájek estimators

$$\hat{m}_G = \frac{\sum_{s_G} w_i x_i}{\sum_{s_G} w_i}$$

and (3)

$$\hat{C}_G = \frac{\sum_{s_G} w_i (x_i - \hat{m}_G)(x_i - \hat{m}_G)^\top}{\sum_{s_G} w_i}.$$

These estimates are used to compute the estimates of the Mahalanobis distances $\widehat{\mathrm{MD}}_G(x_i), x_i \in s$. Finally the correction factor $c_{Npr} = c_{Np} + c_{hr}$ of the selection criteria is computed using the estimates $\hat{N} = \sum_s w_i$ and $\hat{r} = \sum_{s_G} w_i$. If N is known, its actual value is used.

If there are no missing values in the data the BACON algorithm can be used to estimate the population mean and covariance. The basic assumption for the BACON algorithm is still that the bulk of observations of the population has an elliptical distribution. We may use the BACON algorithm without weighting and compare the result with the weighted version. Different results indicate that the design-variables or a model used for non-response weighting are not well reflected in the model. We advocate the use of weights, in particular in routine applications, to give some protection against miss-specification of the model. In any case the estimand should be the mean and covariance of the bulk of the population.

Note that the Mahalanobis distance does not involve the sampling weights directly. The weight of a possible outlier influences the Mahalanobis distance only through the model, *i.e.*, the mean and the covariance.

3. The EEM algorithm

Nonresponse issues are important in official statistics and many surveys cannot deliver a complete dataset. The problem of unit-nonresponse, *i.e.*, completely missing observations, is usually dealt with by using appropriate weights and is not treated here. Item-nonresponse, *i.e.*, observations with only partially available information, cannot be treated by discarding all incomplete observations because too much information is lost. The approach followed here will retain high efficiency under multivariate normal data. At each BACON iterative step the mean and covariance matrix of the good subset of observations will be computed using a modified version of the EM algorithm for multivariate normal data. The expectations computed in the E-step are replaced by sample estimates. The modified algorithm is therefore named the EEM (Estimated-Expectation/Maximization) algorithm. Note that this adaptation is presented here for multivariate normal data but the results can be generalized to other distributions of the regular exponential family.

This paragraph re-uses the description and notation of the EM algorithm given in (Schafer 2000). All details about EM not given here can be found within the first three chapters and in Section 5.3 of this book. The following abuse of notation is also used here: X will denote simultaneously a p-dimensional random variable and the $N \times p$ matrix containing the realized values of the variable X of the population U. If a census were taken of the whole population to measure the variable X it would result in some observed and missing values $X = X_o \cup X_m$. The EM-algorithm assumes that the missingness mechanism is ignorable (Schafer 2000, section 2.2). Here we assume in addition that the missingness is independent from the sampling. The observations of the data can be modeled as independent, identically distributed (iid) draws from a multivariate normal probability distribution with density $f(x, \theta)$. Using the assumptions and the factorization $P(X \mid \theta) = P(X_o \mid \theta)P(X_m \mid X_o, \theta)$ the complete-data log-likelihood can be written as $l(\theta \mid X) = l(\theta \mid X_o) + \log(P(X_m \mid X_o, \theta)) + c$, where $l(\theta \mid X_o)$ is the observed-data log-likelihood and c is an arbitrary constant. The term $P(X_m \mid X_o, \theta)$ captures the interdependence between X_m and θ on which the EM-algorithm capitalizes. Because $P(X_m \mid X_o, \theta)$ is unknown the average of $l(\theta \mid X)$ over $P(X_m \mid X_o, \theta^{(t)})$ is taken at each E-step, where $\theta^{(t)}$ is a preliminary estimate of the unknown parameter. The next estimate $\theta^{(t+1)}$ is found by maximizing the result of the expectation step (M-step). The sequence of E and M-steps is iterated until convergence. Conditions under which this sequence $\theta^{(t)}$ converges to a stationary point of the observed-data likelihood are provided in (Dempster *et al.* 1977). In well-behaved problems this stationary point is a global maximum.

For a probability distribution of the regular exponential family the complete data log-likelihood may be written as

$$l(\theta \mid X) = \eta(\theta)^{\top} \cdot T(X) + Ng(\theta) + c, \qquad (4)$$

where $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), ..., \eta_k(\theta))^{\top}$ is the canonical form of the parameter θ and $T(X) = (T_1(X), T_2(X), ..., T_k(X))^{\top}$ is the vector of complete-data sufficient statistics. Moreover, each of the sufficient statistics has an additive form $T_j(X) = \sum_{i=1}^{N} h_j(x_i)$, for some function h_j. Because $l(\theta \mid X)$ is a linear function of the sufficient statistics, the E-step replaces $T_j(X)$ by $E(T_j(X) \mid X_o, \theta^{(t)})$. In other words the E-step fills in the missing portions of the complete-data sufficient statistics. For a multivariate normal distribution $X = (X^1, \ldots, X^p)$ the sufficient statistics are composed of two types of elements: the sums $\sum_{i=1}^{N} x_i^k$ and the sums of products $\sum_{i=1}^{N} x_i^k x_i^l$, $1 \le k, l \le p$. The E-step reduces to computing the conditional expectations of these sums given the observed data X_o and the preliminary parameter $\theta^{(t)}$.

For a single summand i one can show (Schafer 2000, section 5.3) that these expectations depend only on the observed components of the same observation, *i.e.*, on x_i^{obs}. This leads to

$$E\left(\sum_{i=1}^{N} x_i^k \,\middle|\, X_o, \theta^{(t)}\right) = \sum_{i=1}^{N} E(x_i^k \mid X_o, \theta^{(t)}) = \sum_{i=1}^{N} E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}), 1 \le k \le p \qquad (5)$$

and the analogue form of the sum of products. Of course $E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}) = x_i^k$ if $x_i^k \in x_i^{\text{obs}}$. If x_i^k is missing, then this expectation is the fitted value of a regression of x^k given the parameter $\theta^{(t)}$ on the variables which are observed for observation i. Thus the sufficient statistics are composed of population sums of observed values (T_o) and sums of fitted values (T_m).

In the situation where our data stem from a sample of a finite population we consider the finite population as a realization of a multivariate normal distribution and the sums (5) and sums of products have to be estimated from the sample. The form of (5) allows the use of simple Horvitz-Thompson estimators. The estimate of (5) is

$$T^{k0} = \sum_{s} w_i E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}), \; 1 \le k \le p, \qquad (6)$$

and $E(\sum_{i=1}^{N} x_i^k x_{i_s}^l \mid X_o, \theta^{(t)})$ is estimated by

$$T^{kl} = \sum_{s} w_i E(x_i^k x_i^l \mid x_i^{\text{obs}}, \theta^{(t)}), \; 1 \le k, l \le p. \qquad (7)$$

In short: We replace the population sums of T_o and T_m by their Horvitz-Thompson estimators $\hat{T}_o$ and $\hat{T}_m$. We call the calculation of the T^{k0} and T^{kl} the estimated expectation step (EE-step). Plugging these estimators into (4) we obtain an estimator of the average population likelihood function.

For the M-step, the maximization of the estimate of the average population likelihood, the weighted normal equations have to be solved. The solution is found by a simple matrix operation using the sweep operator (Schafer 2000, section 5.3) applied to the symmetric $(p+1) \times (p+1)$ matrix $(T^{kl})_{0 \le k, l \le p}$ of the estimated expectations of the sufficient statistics (with T^{00} set to 1) divided by N, which is estimated by the sum of weights if unknown:

$$\theta^{(t+1)} = \text{SWP}[0]\left(\frac{(T^{kl})_{0 \le k, l \le p}}{\sum_s w_i}\right), \qquad (8)$$

where SWP[0] is the sweep operator on the first line/column of the matrix.

The EEM algorithm iterates the EE and the M-step. Computationally the difference between the EE-step and the E-step of the original EM-algorithm comes down to using

weighted sums instead of un-weighted sums with weights that do not change over the iterations. We therefore expect that the convergence of the EEM-algorithm will remain similar to the EM-algorithm. For the BACON-EEM algorithm we only need a rough approximation to the solution in each BACON-step. Thus we use only a small number of iterations of the EEM algorithm.

3.1 Mahalanobis distance with missing values

The Mahalanobis distance is developed for complete observations and needs to be adapted to missing values. One option is to use the EEM estimate to impute the conditional mean for the missing values given the observed values and then calculate the Mahalanobis distance with imputed values. Under a MAR (Missing At Random) assumption there is a valid model based on the observed part of the data to impute the missing values. In the case of outlier detection we suppose that the imputation model may hold for the bulk of the data only and is estimated in a robust way. But then we may not expect that an outlier value is predicted by the model, except if already the observed part of an observations is outlying. Therefore there is no advantage to use imputation before outlier detection and we prefer to directly adapt the Mahalanobis distance to missingness. Two different versions of the Mahalanobis distance are possible in this situation.

We call the first version *marginal* Mahalanobis distance. It uses the Mahalanobis distance in the space of observed variables and scales it up with a factor p/q, where $q = \sum_k r_{ik}$ is the number of non-missing variables and p is the total number of variables. More precisely, we assume an observation x is partitioned into $x = (x_o^\top, x_m^\top)^\top$ (after possible rearrangement), where x_o denotes the observed part and x_m the unobserved part of the observation. Then the marginal Mahalanobis distance is

$$\mathrm{MD}^2_{\mathrm{marg}} = \frac{p}{q}(x_o - m_o)^\top (S_{oo})^{-1}(x_o - m_o), \qquad (9)$$

where S_{oo} is the part of the covariance matrix corresponding to x_o. This version is also used in (Little and Smith 1987).

The second version of Mahalanobis distance with missing values is obtained by reducing the contribution of the missing values to the Mahalanobis distance to zero. This amounts to replacing all missing values by their mean, *i.e.*, $x_m = m_m$. In other words we would impute a mean without consideration of the covariance matrix and the above arguments against outlier detection with imputed values apply here as well. Nevertheless we tested this second version of Mahalanobis distance. It yields erratic Mahalanobis distances (Béguin 2002) and (Béguin and Hulliger 2003) and we did not use it any further.

4. The BACON-EEM algorithm

Both algorithms, BACON and EEM, are computationally demanding. By merging them in a convenient way we gain performance. The "growing" structure of the BACON algorithm implies redundancies which may be used to avoid extra-computations in the EEM-algorithm at each step. The crucial point at each BACON step is that the estimations of the mean and the covariance matrix from the EEM-algorithm allow the exclusion of outlying points from the good subset and this does not need extremely precise estimates. Thus it is not necessary to iterate EEM to convergence each time the mean and covariance are needed. We use only 5 iterations by default. Furthermore we use the result of the last EEM-iteration of the last BACON-step as a starting value for EEM.

As much information from past iterations as possible should be reused. In fact the sufficient statistic T^G computed on some good subset G have an observed part of the sum T_o^G and a missing part of the sum T_m^G. The expectation computed by the E-step can therefore be written as

$$E(T^G \mid X_o^G, \theta) = T_o^G + E(T_m^G \mid X_o^G, \theta). \qquad (10)$$

As the subsets G are usually growing, $\hat{T}_o^G$ is not recomputed at each step of the BACON loop, but a global variable for $\hat{T}_o^G$ is updated each time G changes (usually only adding points, sometimes removing a few).

At each iteration of the BACON-EEM algorithm, once the EEM algorithm has obtained the estimations of the center and the scatter of the good subset, marginal Mahalanobis distances for all observations are used in step 2 of the BACON algorithm.

Note the crucial point for the robustness of the algorithm: EEM is not robust, but at each BACON-step EEM is run only on points that have the smallest and therefore non-outlying marginal Mahalanobis distance in the preceding step. In other words the observation x will be used by EEM if and only if x_o is sufficiently small for the metric given by $(S_{oo})^{-1}$ at the preceding step. Therefore if the first subset of good points is free or almost free of outliers, the imputation process in EEM will never create outlying values throughout the whole BACON-EEM algorithm. In other words, the non-robust EEM-algorithm is protected by the general forward search approach of the BACON algorithm in the same way as the non-robust mean and covariance of the original BACON algorithm is protected.

Summing up, the steps of the BACON-EEM algorithm are the following:

1. Calculate the weighted coordinate-wise median $\widehat{\mathrm{med}}(x)$ ignoring missing values in each variable separately. Determine the Euclidean distance from the median of

each observation omitting missing values but standardizing for the number of present values: $a_i = \|x_i - \widehat{\text{med}}(x)\|\sqrt{p/q}$. Select the $m = cp$ observations with least a_i to constitute the initial subset G.

2. Compute a center $\hat{m}_G$ and scatter $\hat{C}_G$ using the EEM-algorithm and update the estimate of the sufficient statistic of the observed part $\hat{T}_o^G$.
3. Compute the squared marginal Mahalanobis distances $\text{MD}_G^2(x_i)$ for $i = 1, ..., n$. The new set G' contains the observations with $\text{MD}_G^2(x_i) < c_{\hat{N}p\hat{r}}\ \chi^2_{p,\alpha}$.
4. If $G' = G$ then stop, else set G to G' and go to step 2.

If instead of outlier detection the mean and covariance estimates of BACON-EEM are the main objectives the EEM-algorithm may be iterated further without changing G. In step 3 one may alternatively use α/n instead of α (see Section 2).

5. Applications

In this section we compare the BACON-EEM algorithm (BEM) with Transformed Rank Correlations (TRC) from (Béguin and Hulliger 2004) and the ER-algorithm from (Little and Smith 1987). As a further benchmark we use an imputation under the multivariate normal model with estimates of the mean and covariance by the EM algorithm. In other words we create a non-robust imputation. Then robust estimates of the multivariate location and the covariance matrix are obtained by the Minimum Covariance Determinant estimator computed on the imputed data and finally outliers are detected using the corresponding Mahalanobis distances. The benchmark method is called GIMCD for "Gauss Imputation followed by MCD detection". The algorithms are implemented in `R` (R Development Core Team 2006) with the help of the `R`-packages `norm` (Novo and Schafer 2002) and `MASS` (Venables and Ripley 2002).

5.1 Bushfire data

The reaction of the BACON-EEM to the introduction of missing values is illustrated with a real dataset of 38 observations and 5 variables. It was used by (Maronna and Zamar 2002) to locate bushfire scars. This well known example is also studied in (Maronna and Yohai 1995) and (Maronna and Zamar 2002). It allows a two dimensional plot (in variable 2 and 3) that reveals most of the outliers (see Figure 1). The data contains an outlying cluster of observations 33 to 38 a second outlier cluster of observations 7 to 11 and a few more isolated outliers, namely observations 12, 13, 31 and 32. We have added observation 31 to the list of potential outliers because it is indicated as a borderline case by MCD, BACON and also other methods studied in (Maronna and Zamar 2002). Missing values are created with a MCAR (Missing Completely At Random) mechanism. Two datasets are created with respectively 20 and 40% of missing items. The dataset with 40% of missing values have observations with up to 4 out of 5 missing values and therefore are a challenge for any method. As the size n of the dataset is small, BACON-EEM is run with the $\chi^2_{p,\alpha/n}$ test. The results are given in Table 1. Observations 7 to 13 and 31 to 38 are individually shown as detected or not, while for the other 23 good points the number of observations declared as outliers is indicated. The limit above which a Mahalanobis distance indicates an outlier, was determined for each run by inspection of the quantile plot of the Mahalanobis distance.

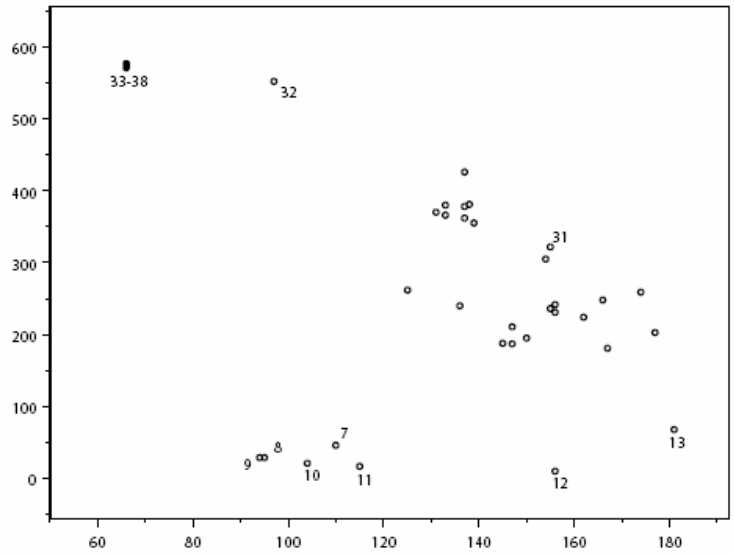

Figure 1 Bushfire Data

With complete observations none of the methods declared any good observations as outlier. The Mahalanobis distance with non-robust mean and covariance detects 3 outliers but misses the others. MCD and BACON-EEM end up with the same subset of data as good points and therefore give the same result, *i.e.*, exactly the same Mahalanobis distance for all observations. Both do not declare observations 12 and 13 as outliers but all the others. ER does detect the group of outliers 7,...,12 but none of the other outliers. TRC detects the group 32,...,38 and two more outliers.

With 20% missing values BEM and TRC declare one good data point as outlier. TRC detects 14 of the 15 potential outliers. ER misses all outliers except observation 7. GIMCD detects the same 13 outliers as without missing values. Note however, that there is some variability in the results for GIMCD due to the random imputation. BEM detects the same outliers as with complete data except observation 11.

Table 1 Outliers detected for 3 missingness rates

(1 - *q*)%	Method	7-11	12,13,21	32-38	n. good
0	MD	11100	000	0000000	0
0	MCD	11111	001	1111111	0
0	ER	11111	100	0000000	0
0	BEM	11111	001	1111111	0
0	TRC	01100	000	1111111	0
20	GIMCD	11111	001	1111111	0
20	ER	10000	000	0000000	0
20	BEM	11111	000	1111111	1
20	TRC	01111	111	1111111	1
40	GIMCD(1)	11100	000	1111111	0
40	GIMCD(2)	11100	000	0100000	5
40	ER	10000	010	0000000	2
40	BEM	11111	000	1111111	1
40	TRC	11111	010	1111111	1

MD: Classical Mahalanobis distance, MCD: Minimum Covariance Determinant, GIMCD: Non-robust imputation under Gaussian model followed by MCD (GIMCD(1) and GIMCD(2) are two realisations of the GIMCD-algorithm), ER: Expectation-Maximization with one M-step at maximization, BEM: BACON-EEM, TRC: Transformed Rank Correlations. The first column indicates the proportion of missing values, the last column gives the number of other points (non-outliers) declared outliers, the intermediate columns are detection indicators for the observations in the first row.

With 40% missing values ER nominates observation 7, 13 and two good observations as outliers. Since the imputation is random the result of GIMCD has some variability. Two realizations, GIMCD(1) and GIMCD(2) are reported in Table 1. In a good case GIMCD detects 10 of the 15 outliers and does not declare any good observation as outlier. In a bad case GIMCD detects only 4 outliers but declares 5 good observations as outliers. BEM detects 12 of the outliers and declares one good observation as outlier. TRC detects 13 outliers and declares one good observation as outlier.

5.2 MU281 data

The MU284 data set from (Särndal, Swensson and Wretman 1992) contains data about Swedish municipalities. We use the variables *population in 1975* and *population in 1985* (pop75 and pop85), *revenue from municipal taxes 1985* (RMT85), *number of municipal employees 1984* (ME84) and *real estate value 1984* (REV84). The largest three cities according to pop75 are discarded because they are huge outliers and would be treated separately in practice. The remaining municipalities are supposed to be a stratified sample of a larger population. Strata are defined according to $0 < \text{pop75} < 20, 20 \le \text{pop75} < 100, 100 \le \text{pop75}$. Table 2 shows the assumed population sizes and the corresponding weights. This sample design reflects a typical stratification for establishment surveys with a take-all stratum of the largest establishments, where in the end 8 of 10 establishments answer the survey.

Table 2 MU281 population and sample sizes

	stratum		
	1	2	3
pop75	0-19	20-99	100+
N	1,600	250	10
n	171	102	8
w	9.36	2.45	1.25

The three variables RMT85, ME84 and REV84 are divided by pop85 to obtain figures per capita. The per capita variables are denoted by lower case names (rmt85, me84 and rev84). Figure 2 shows the distribution of these 3 variables plus the auxiliary variable pop75. The per capita figures are roughly elliptically distributed. There is a linear relationship between rmt85 and me84 and a slightly non-linear relationship of these variables with pop75. There is no apparent relationship between rev84 and rmt85 and between rev84 and me84 but there is clearly more variability in rev84 for low pop75. The distributions of variable pop75 and rev84 are skew. There is a large outlier in rmt85 and me84 and at least two in rev84.

We include pop75 in all our calculations. In practice one would include the auxiliary variable which defines the sample design in a model. Note that pop75 has no missing values.

The qq-plot of Mahalanobis distances based on MCD shows only the two clear outliers in rev84. The large outlier jointly in rmt85 and me84 has 25th largest Mahalanobis distance. We call these largest 25 observations the unweighted basic outliers. In the original MU284 dataset these unweighted basic outliers have LABEL 3, 4, 29, 31, 46, 47, 56, 79, 83, 117, 126, 131, 140, 158, 199, 211, 222, 246, 248, 252, 254, 260, 262, 272, and 273. With classical non-robust Mahalanobis distances only 12 of the basic outliers are nominated outliers, *i.e.*, are among the 25 observations with

largest (classical) Mahalanobis distance. Robust methods are necessary to detect the outliers in the MU281 data. To allow a comparison between the methods we fix the number of observations which are to be considered outliers to 25 for the moment. Thus we consider for each method the 25 observations with largest Mahalanobis distance as the outliers. Note that this may not be the threshold one would choose after inspection of the qq-plot of the Mahalanobis distances.

Table 3 shows the number of basic outliers detected by the methods run on the complete dataset. MCD detects its own 25 outliers, of course. The ER, BEM and TRC algorithm detect 25 or 24 of these outliers if no weights are applied but only 11or 15 if weighted. A suffix "w" behind the acronym of the method indicates that the sampling weights were used. Since the small municipalities have more weight the estimates are attracted towards them and other outliers will come up among the 25 largest Mahalanobis distances for ERw, BEMw and TRCw (see also Table 4). Closer inspection shows that many of the unweighted outliers are located in the tail of rev84 while most of the weighted outliers are located in the tail of pop75. The weighted methods coincide on 20 observations as outliers. We will call them weighted basic outliers. The weighted basic outliers have original MU284-LABEL 16, 28, 36, 45, 46, 55, 97, 113, 115, 121, 155, 185, 196, 208, 233, 241, 245, 265, 267, and 270. Only 10 of these observations are also among the unweighted basic outliers.

Table 4 shows the number of weighted and unweighted basic outliers in the strata. There are 12 unweighted but only 2 weighted basic outliers in stratum 1. Thus the weights have a clear influence on outlier detection. The influence is on the model primarily which is attracted towards the small observations with larger weights. Of course this can be seen as a sort of masking of outliers but in the context of modeling a better explanation is that the model is not completely adequate over all the strata and the weighted model fits the population better than the unweighted.

The second row of Table 3 gives the computation time for the algorithms. The ER algorithm is much slower than its competitors. This may be due to an inefficient implementation, however. The fastest algorithm is BEM, followed by TRC and, at some distance MCD. TRC may become slow, however, when the missingness rate is high.

Table 3 Complete MU281, number of detected unweighted basic outliers

Method	MCD	ER	ER_W	BEM	BEM_W	TRC	TRC_W
Number detected	25	25	11	24	15	24	15
Computation time	0.81	3.17	2.52	0.07	0.04	0.14	0.14

Suffix w indicates that the algorithm is run with sampling weights.

Table 4 Number of basic outliers per stratum

stratum	1	2	3	Total
unweighted	12	5	8	25
weighted	2	10	8	20

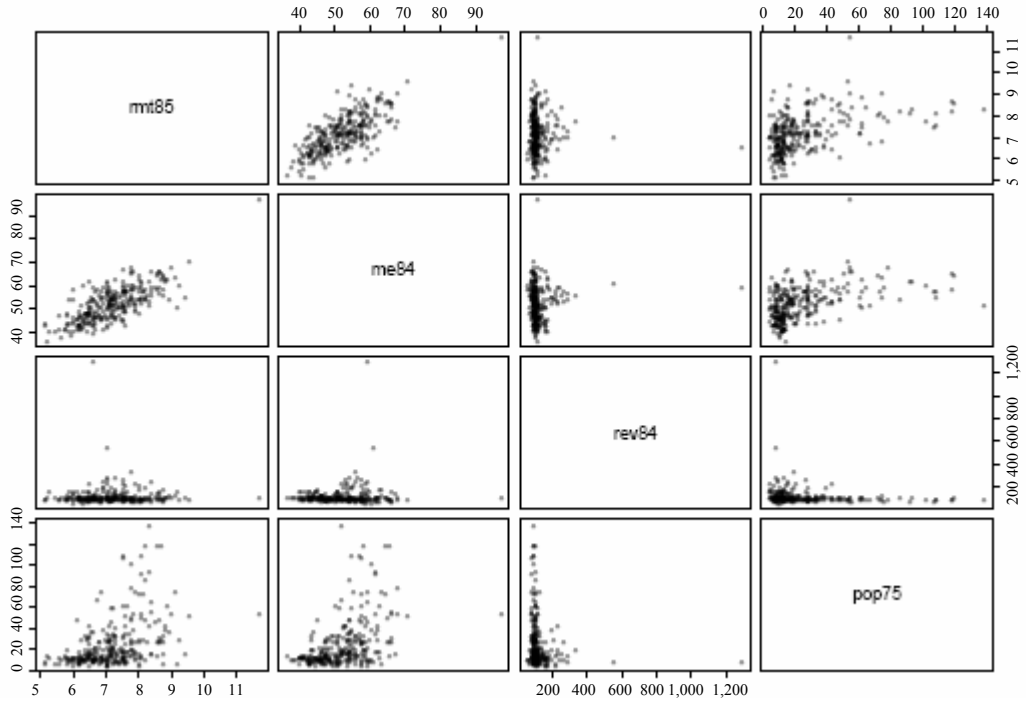

Figure 2 MU281 per capita figures and pop75

5.2.1 Missing values

We now introduce missingness into the variables rmt85, me84 and rev84 according to a mild MAR mechanism. Missingness patterns are assigned to the observations according to the modulo of their label in the original MU284 data. The MAR mechanism is reflected in higher missingness rates for stratum 1 and 2 (see Appendix). The response patterns and missingness rates per stratum are shown in Table 5. For example rev84 is the only missing value in 15 observations of stratum 1 and 2 of stratum 2. Overall 187 observations remain complete and the proportion of observations with missing values (missingness rate) is 33%.

Table 5 Frequency of response patterns per stratum for rmt85, me84, rev84

Response indicator			stratum		
rmt85	me84	rev84	1	2	3
0	1	1	11	4	0
1	0	1	13	2	0
1	1	0	15	2	0
1	0	0	13	2	1
0	1	0	14	2	0
0	0	1	11	4	0
1	1	1	94	86	7
missingness rate			0.450	0.157	0.125

Among the 35 weighted or unweighted basic outliers there are 17 observations with missing values. Table 6 shows how many of the basic outliers have been detected after the introduction of missingness. The 20 weighted basic outliers are detected well by the weighted algorithms ERw, BEMw and TRCw. GIMCD detects 4 of the weighted basic outliers and 14 of the unweighted basic outliers. Thus the missingness affects the capability of the MCD algorithm which was actually used to define the unweighted basic outliers. One word of caution: Several runs of random Gaussian imputation have been made and there is some variability in the results of GIMCD. However, also with a favorable imputation outcome GIMCD did not beat ER, BEM or TRC in detecting the unweighted basic outliers. The weighted versions ERw, BEMw and TRCw detect the weighted basic outliers well. The number of complete observations among the outliers nominated by the different methods is indicated in the last row of Table 6. All methods nominate as outliers also observations with missing values. Since the missingness rate is larger in the stratum of small observations and the weighted versions of the methods nominate less outliers in this stratum, the number of complete outliers is usually larger for the weighted algorithms. Overall the introduction of missingness has not altered the capabilities of ER, BEM and TRC by much, while GIMCD is moderately affected.

Table 6 MU281 data set with missing values, number of detected basic outliers

Method	GIMD	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Weighted	14	4	10	20	9	19	12	17
Unweighted	12	14	23	11	22	15	22	17
Complete	16	8	17	20	16	18	19	18

GIMD: Non-robust imputation under a Gaussian model followed by classical Mahalanobis distance.

5.2.2 Additional outliers

In addition to the outliers in the original data we now introduce new outliers. The observations which should become additional outliers are determined by the modulo of the original LABEL. If (LABEL mod $8 = 1$ and pop75 ≥ 10) or (LABEL mod $16 = 1$ and pop75 < 10) then the observation is an additional outlier. Thus the rate of outlyingness is larger for large municipalities. But the outlyingness is not influenced by the values of the other variables. We may say that the outlyingness is at random. Note that we could have taken a random sample instead of the above systematic sample. We preferred the systematic sample to simplify the replication of the results and to avoid additional randomness.

Two of the weighted and one of the unweighted basic outliers happen to be also additional outliers. We continue to treat them as basic outliers. Taking this into account there are 32 additional outliers in the sample. Together with the 25 unweighted or the 20 weighted basic outliers defined above there are 57 or 52 outliers to detect (20.3% or 18.5% outliers). From now on the threshold for the Mahalanobis distances is set at the 57^{th} largest distance to simplify the comparison of the methods.

The values of the additional outliers are created as follows: $\text{rmt} = 0.2 * \text{rmt85} + 8$, $\text{me} = 0.1 * \text{me84} + 50$, $\text{rev} = 0.4 * \text{rev84} + 300$. Note that we omit the suffix indicating the year for the contaminated variables. The dependence on the old values is negligible. It is only used to avoid an explicit model for the error around the point $(\text{rmt}, \text{me}, \text{rev}) = (8, 50, 300)$. This is the type of contamination that is difficult to detect for robust covariance estimators (Rocke and Woodruff 1996): concentrated and close to the point cloud of good observations.

Figure 3 shows the three variables with contamination and the location of the additional outliers.

Table 7 shows the number of detected outliers. GIMCD detects 31 of the 32 additional outliers, while BEM, BEMw, TRC and TRCw detect many of them but not all. ER and ERw detect less of the additional outliers. The weighted basic outliers are all detected by ER and ERw, BEMw and TRCw. The unweighted versions of BEM and TRC detect less of the weighted basic outliers and GIMCD detects only 4 of the weighted basic outliers. BEM and TRC whether weighted or not detect the unweighted basic outliers best.

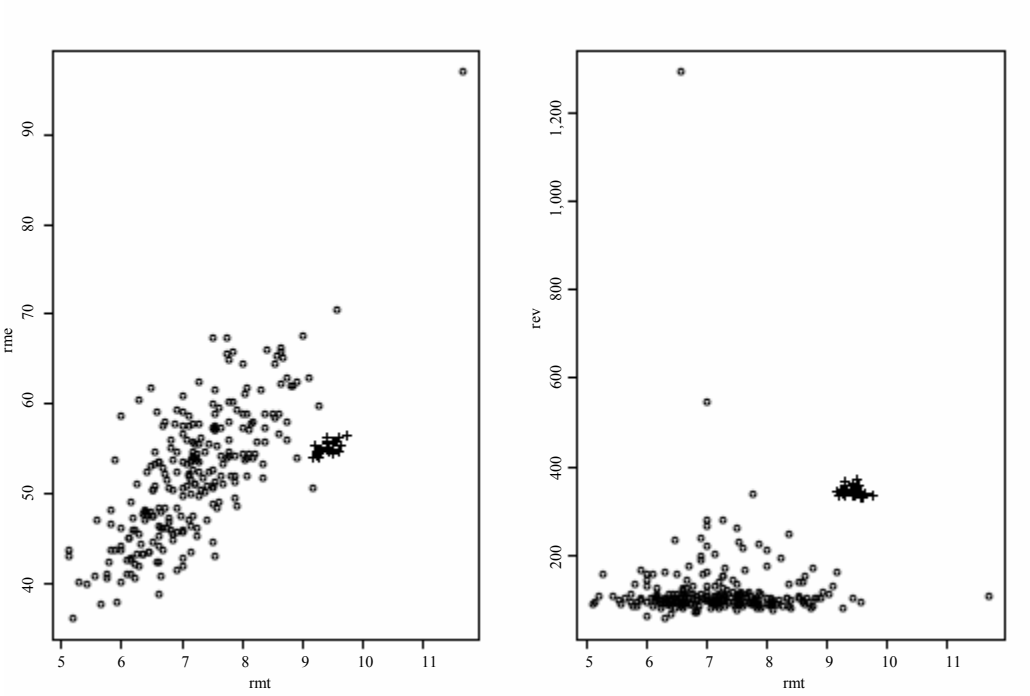

Figure 3 Variables me and rev *vs.* rmt. Additional outliers are marked with a cross

The last row in Table 7 shows that its non-robust imputation leads GIMCD to nominate more outliers with missing values than the other methods which robustify their imputations already before the detection phase.

Table 7 MU281 with missingness and moderate additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	32	31	19	6	27	27	28	27
Weighted basic	20	4	20	20	11	20	12	20
Unweighted basic	25	13	15	12	23	19	23	18
Complete		24	34	43	38	40	40	40

n. out: number of outliers, n. complete: number of complete observations among detected outliers.

In order to check the breakdown of the methods when a high number of additional outliers contaminates the data we set outliers if (LABEL mod $2 = 0$ and pop75 ≥ 20) or (LABEL mod $3 = 1$ and pop75 < 20). Excluding observations which are already basic outliers there are 98 additional outliers. Thus together with the 25 unweighted basic outliers we obtain 43.8% outliers. The threshold for the methods is therefore set at the 123rd largest Mahalanobis distance. Table 8 shows that, due to this large threshold, all weighted basic outliers are detected by all methods. The methods GIMCD, ER and ERw cannot cope with the high rate of outliers. BEM detects most of the outliers with BEMw and TRC only slightly behind. TRCw detects somewhat less of the unweighted basic outliers and of the additional outliers.

Table 8 MU281 with missingness and high additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	98	20	19	37	85	85	85	80
Basic weighted	20	20	20	20	20	20	20	20
Basic unweighted	25	21	19	17	23	18	18	13

6. Conclusions

The EM-algorithm for multivariate normal data can be adapted to a sampling context. The BACON algorithm protects the non-robust EEM-algorithm from outliers when the latter is applied within an iteration of the BACON algorithm. The ER-algorithm uses robustification within the EM-algorithm. The applications showed that this may not yield enough robustification. A possible reason, however, may also be the non-robust starting point of the M-step in the ER-algorithm.

GIMCD, a non-robust EM-algorithm followed by an imputation and detection with MCD covariance worked remarkably well for moderate missingness and contamination. Its variability with high missingness rate is a disadvantage. More stable solutions which also can take into account the sampling design should be explored. The

BACON-EEM algorithm showed very good detection capabilities in particular when the missingness rate and the contamination rate are high.

In spite of its simplicity the TRC algorithm is a good method in many circumstances. Its main problem seems to be the *ad-hoc* imputation with only one covariable, which can be a problem with high missingness rates.

In order to find a good model for the population it is important to use the sampling weights. Nevertheless, it is advisable to use also a non-weighted version and to check the differences. It is possible that outliers are masked by large sampling weights because they may then dominate the model estimate.

Acknowledgement

The EUREDIT research project was part of the Information Society Technology Program (IST) of Framework Program 5 of the European Union. The Swiss participation in EUREDIT was supported by the Swiss Federal Office for Education and Science. A large part of this research was carried out while both authors worked at the Swiss Federal Statistical Office. The authors wish to thank the referees and editors for their valuable remarks.

Appendix

Missingness in MU281

The default response pattern is 111, indicating that the three variables rmt85, me84 and rev84 are all present. A 0 in the string indicates a missing value for the corresponding variable. First, for all strata the response pattern is changed accordingto the following scheme with parameters $(a, b, c) = (1, 2, 3)$:

$$\text{response pattern} = \begin{cases} 011, & \text{if LABEL mod } 20 = a; \\ 101, & \text{if LABEL mod } 20 = b; \\ 110, & \text{if LABEL mod } 20 = c; \\ 100, & \text{if LABEL mod } 30 = a; \\ 010, & \text{if LABEL mod } 30 = b; \\ 001, & \text{if LABEL mod } 30 = c\,. \end{cases}$$

Additionally, the above scheme with parameters $(a, b, c) = (5, 6, 7)$ is applied for stratum 1 again.

References

Atkinson, A. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. In *Data Analysis and Robustness*. (Eds., S. Morgenthaler, E. Ronchetti and W. Stahel), Birkäuser.

Béguin, C. (2002). Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel.

Béguin, C., and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.

Béguin, C., and Hulliger, B. (2004). Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society*, A 167(Part 2.), 275-294.

Billor, N., Hadi, A.S. and Vellemann, P.F. (2000). BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. *Computational Statistics and Data Analysis*, 34(3), 279-298.

Campbell, N. (1989). Bushfire mapping using noaa avhrr data. Technical report, CSIRO.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396), 1063-1069.

Cheng, T.-C., and Victoria-Feser, M.-P. (2000). Robust correlation estimation with missing data. Technical Report 2000.05, Université de Genève.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society*, Series B, *Methodological*, 39, 1-22.

Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators*. Ph.d. qualifying paper, Department of Statistics, Harvard University.

EUREDIT (2003). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1 and 2. EUREDIT consortium. http://www.cs.york.ac.uk/euredit/results/results.html.

Gnanadesikan, R., and Kettenring, J.R. (1972, March). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, 81-124.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, B, 54(3), 761-771.

Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.

Kosinski, A.S. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis*, 29, 145-161.

Little, R., and Smith, P. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

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

Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3), 783-858.

Maronna, R., and Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.

Maronna, R.A. (1976). Robust *M*-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51-67.

Maronna, R.A., and Yohai, V.J. (1995). The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429), 330-341.

Novo, A.A., and Schafer, J.L. (2002). *norm*: *Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.

R Development Core Team (2006). *R*: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rocke, D., and Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47, 27-42.

Rocke, D., and Woodruff, D. (1996). Identification of outlier in multivariate data. *Journal of the American Statistical Association*, 91(435), 1047-1061.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Volume B, 283-297. Elsevier.

Rousseeuw, P.J., and Leroy, A.M. (1987). Robust Regression and Outlier Detection. New York: John Wiley & Sons, Inc.

Schafer, J. (2000). *Analysis of Incomplete Multivariate Data,* Volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

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

Stahel, W. (1981). *Robuste Schätzungen: infinitesimale optimalitätund Schätzungen von Kovarianzmatrizen*. Ph.D. Thesis, Swiss Federal Institute of Technology.

Venables,W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S* (Fourth Ed.). New York: Springer. ISBN 0-387-95457-0.

Wilks, S.S., and Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics*, 35, 623-631.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 105-117
Statistics Canada, Catalogue No. 12-001-X

Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias

Annette Jäckle and Peter Lynn [1]

Abstract

Respondent incentives are increasingly used as a measure of combating falling response rates and resulting risks of nonresponse bias. Nonresponse in panel surveys is particularly problematic, since even low wave-on-wave nonresponse rates can lead to substantial cumulative losses; if nonresponse is differential, this may lead to increasing bias across waves. Although the effects of incentives have been studied extensively in cross-sectional contexts, little is known about cumulative effects across waves of a panel. We provide new evidence about the effects of continued incentive payments on attrition, bias and item nonresponse, using data from a large scale, multi-wave, mixed mode incentive experiment on a UK government panel survey of young people. In this study, incentives significantly reduced attrition, far outweighing negative effects on item response rates in terms of the amount of information collected by the survey per issued case. Incentives had proportionate effects on retention rates across a range of respondent characteristics and as a result did not reduce attrition bias in terms of those characteristics. The effects of incentives on retention rates were larger for unconditional than conditional incentives and larger in postal than telephone mode. Across waves, the effects on attrition decreased somewhat, although the effects on item nonresponse and the lack of effect on bias remained constant. The effects of incentives at later waves appeared to be independent of incentive treatments and mode of data collection at earlier waves.

Key Words: Attrition; Item nonresponse; Mail survey; CATI; Unconditional incentive; Conditional incentive.

1. Introduction

Surveys around the world are facing declining response rates and, with this, increasing risks of nonresponse bias if nonrespondents' characteristics systematically differ from respondents' characteristics. For panel surveys this is particularly problematic, since even low nonresponse rates at each wave can lead to large cumulative losses. If nonresponse is differential, bias could increase with the duration of the panel. In order to boost participation rates, survey organisations increasingly offer respondent incentives. This paper provides new evidence on the cumulative effects of incentives on attrition, attrition bias and item nonresponse, using data from a large scale, multi-wave, mixed mode incentive experiment on a UK government panel survey of young people.

The effects of incentives have been studied in many settings: monetary incentives increase response more than gifts or lotteries (Church 1993; Singer, Hoewyk, Gebler, Raghunathan and McGonagle 1999); unconditional incentives (*i.e.*, those incentives that are given at the time of the survey request) increase response more than conditional incentives (those that are promised in return for participation) (Church 1993; Goyder 1994; Hopkins and Gullickson 1992; Singer *et al.* 1999); response rates increase with the value of the incentive (Armstrong 1975; Church 1993; Fox, Crask and Kim 1988; Hopkins and Gullickson 1992; Rodgers 2002; Yu and Cooper 1983); incentives have larger effects in studies with low response rates and larger effects in postal than interviewer administered surveys (Singer *et al.* 1999). Most evidence of differences between modes in the effect of incentives, however, stems from comparisons of separate studies and fails to control for differences in other measures affecting response. As a result, differences in the effects of incentives are not necessarily genuine mode effects. The study by Ryu, Couper and Marans (2006) is an exception. The authors compared the effects of monetary incentives and gifts in a mixed mode postal and face-to-face survey. Their study did not, however, include a no-incentive condition and so did not allow an evaluation of the magnitude of incentive effects across modes. We compared the effects of incentives in a mixed postal and computer-assisted telephone interviewing (CATI) survey and, in postal mode, also examined the effects of conditional and unconditional incentives.

Research on the effects of incentives has focused on response rates and little is known about the effects on bias, the ultimate reason for concern about low response. Incentive studies are mostly limited to studying effects on bias in sample composition and some studies have found that incentives disproportionately increase participation of respondents typically under-represented, for example those with low education (Singer, Van Hoewyk and Maher 2000), poor (James 1997), black or poor (Mack, Huggins, Keathley and Sundukchi 1998), of black or Indian minority ethnic groups, living in larger households or households with dependent children, aged 0-20, or single (Stratford, Simmonds and Nicolaas 2003). Biases in sample composition are

1. Annette Jäckle, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. E-mail: aejack@essex.ac.uk; Peter Lynn, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. E-mail: plynn@essex.ac.uk.

however not necessarily correlated with biases in important survey estimates, especially since typically only a handful of compositional characteristics are studied. Ultimately, studies of nonresponse bias are limited by the lack of information about nonrespondents, a limitation which can be overcome to some extent by panel studies, where information about nonrespondents is available from waves prior to the dropout. We estimate the extent of bias due to attrition in terms of sample composition and survey variables. We then assess the effectiveness of incentives at reducing bias, exploiting the information on attriters available from the panel.

Additionally, little is known about the effects of incentives over waves of a panel survey, whether the same treatment is administered repeatedly or whether the treatment changes between waves. In a review of the use of incentives in longitudinal studies, Laurie and Lynn (in press) concluded that, given the cost implications of changing incentive conditions, there was surprisingly little evidence about the longer term effects of such changes to guide survey practitioners. Quoting an internal memorandum of the US Census Bureau, Ward, Boggess, Selvavel and McMahon (2001) wrote (see page 2) that a "review of the well-known longitudinal studies (Downs 1999) found that all non-Census Bureau studies used a monetary incentive during each wave, but there had been no scientific tests to determine the effectiveness of the incentives". If attrition leads to dropout of the least co-operative, the sample might increasingly be composed of committed respondents who are less responsive to incentives, because they are sufficiently motivated to participate even without the incentive (Laurie and Lynn in press). In this case, incentives may have decreasing marginal effects on response rates over the life of the panel. By the same token, incentives may have increasing effects on attrition bias, if they have disproportionate effect on sample members who would otherwise be more likely to drop out. Although some incentive studies have been carried out in the context of panels, they mostly only covered one wave, or examined the effect of changes in incentive treatment from one wave to the next. Martin, Abreu and Winters (2001) and Ward *et al.* (2001), for example, studied the effects of incentives on conversion rates of previous wave nonrespondents; Rodgers (2002) and Laurie (2007) examined the effects of changes in incentive values in a panel. The only studies which examined the effects of incentives over more than two waves appear to be those by James (1997), Mack, Huggins, Keathley and Sundukchi (1998) and Laurie and Lynn (in press), who reported that the positive effect of an incentive paid early in a panel persisted for several waves even without repeated incentive payment. These studies, however, only examined the effect of an incentive paid in a single wave and did not examine the cumulative effects of incentives offered over successive waves. We examine the cumulative effects of continued incentive payments across three waves spanning a time frame of three years, as well as the effects of changes from telephone to postal mode and from conditional to unconditional incentive treatment.

Finally, there is conflicting evidence in the literature about the effects of incentives on data quality. Although concern is frequently voiced that incentives may lead to lower data quality, by marginally increasing the motivation of respondents who would otherwise have dropped out of the study and are not sufficiently able or motivated to respond diligently, existing studies have either found that incentives lead to improved respondent effort and less item nonresponse (James and Bolstein 1990; Mack *et al.* 1998; Singer *et al.* 2000), or have found no relationship (Berk, Mathiowetz, Ward and White 1987; Davern, Rockwood, Sherrod and Campbell 2003; Goyder 1994; Shettle and Mooney 1999; Singer *et al.* 1999; Teisl, Roe and Vayda 2005; Tzamourani and Lynn 1999; Willimack, Schuman, Pennell and Lepkowski 1995). Item nonresponse is potentially critical, because analysts typically only use cases with complete data. This leads to losses in efficiency due to reductions in sample sizes and, similar to unit nonresponse, can lead to biased estimates and invalid inference if item nonrespondents are not a random subset of the sample (Mason, Lesser and Traugott 2002). Problems of item nonresponse increase for multivariate analysis, if the patterns of missingness vary across items, and for analysis of change, which in addition depends on complete information at different points in time. Since incentives may affect both unit and item nonresponse, it is then not clear what their net effect may be on repeated measures derived from a panel study. We examined the effect of incentives on item nonresponse rates and calculated their net effect on attrition and item response.

2. Hypotheses tested

The outcomes measured for this analysis were the attrition rate, item nonresponse rate and attrition bias. Attrition was an absorbing state, since the survey did not reissue nonrespondents at later waves. Item nonresponse was measured as the number of non-filtered items missing, either due to refusals or 'don't know' answers. (Non-filtered items are those which apply to all sample members: items for which eligibility is determined by the response to an earlier question are excluded from our measure of item nonresponse.) Attrition bias was measured in terms of sociodemographic characteristics and wave 1 survey measures. These three outcome measures were used to test the following:

*H*1: *Effects of incentives on attrition, item nonresponse and attrition bias.*
In previous studies incentives have generally increased response rates, be it because norms of social exchange oblige the respondent to return a "favour" (norm of reciprocity, Gouldner 1960) or because the incentive substitutes for a lack of motivation to participate for other reasons, such as civic duty or topic interest (leverage-salience theory, Groves, Singer and Corning 2000). Incentives may in addition motivate respondents to provide better quality responses, reducing item nonresponse. At the same time incentives may change the sample composition to include more respondents who are not diligent about answering the survey questions, and as a result increase item nonresponse. Finally, incentives may have differential effects on attrition across sample members. Those with a high propensity to participate in the survey without the incentive may be less likely to be affected by incentives, while those more likely to drop out of the survey may be more susceptible. As a result, incentives may reduce attrition bias.
*Null hypothesis H*1: *Incentives have no effect on attrition, item nonresponse or attrition bias.*

*H*2: *Effects of incentives across waves*.
The effect of incentives in increasing unit and item response rates may weaken across waves, if attrition leads to dropout of the least motivated sample members and the remaining members are sufficiently motivated to participate for other reasons and hence less susceptible to incentives (Laurie and Lynn in press). However, the extent to which incentives reduce nonresponse bias could increase over waves, if incentives disproportionately retain those in the sample who are most likely to otherwise drop out.
*Null hypothesis H*2: *The effects of incentives do not change across waves.*

*H*3: *Effects of unconditional and conditional incentives in a panel context*.
Previous studies, carried out on cross-sectional surveys, suggest that unconditional incentives have larger effects on unit nonresponse, possibly because the prepayment signals that the survey organisation trusts the sample member will participate, reinforcing the norm of reciprocity. Whether the different incentive conditions have different effects on item nonresponse is not clear.
*Null hypothesis H*3: *Unconditional and conditional incentives have similar effects in a panel context.*

*H*4: *Effects of incentives in postal and telephone mode*.
Comparisons of previous studies suggest that incentives have a larger effect in postal mode, possibly because in telephone mode the interviewer already functions as an external motivator to increase both unit and item response (Singer *et al.* 1999) and the scope for additional improvements is smaller. The same may not necessarily be true in a panel context where the effect of mode on response may be mediated by the respondent's experience of previous waves.
*Null hypothesis H*4: *Incentives have similar effects in postal and telephone mode.*

*H*5: *Effects of changes over waves in mode or incentive treatment*.
Compared to sample members allocated to the same mode and treatment across waves, those who were allocated to different treatments or different modes may differ in their experiences of previous survey waves and their expectations about future waves. As a result, the effect of incentives may not only be conditional on mode at the current wave, but may be influenced by the incentive treatment and mode in previous waves.
*Null hypothesis H*5: *Changes in mode or incentive treatment over waves do not have lasting effects.*

*H*6: *Effects of incentives across ability levels*.
Sample members with low education levels are typically more likely to drop out of surveys. If incentives reduce attrition bias, they should therefore disproportionately reduce attrition among lower achievers. Low ability respondents may at the same time be more likely to provide incomplete responses, if they find the task of completing the postal questionnaire more difficult. Therefore, incentives *may* increase mean levels of item nonresponse.
*Null hypothesis H*6: *Incentives have similar effects across ability levels.*

3. Study design

The Youth Cohort Study of England and Wales (YCS) investigates transitions from compulsory education to further or higher education or the labour market and typically samples cohorts of 16 to 17 year-olds every two years, who are surveyed on several occasions at annual intervals. The incentives experiment was embedded in waves 2, 3 and 4 of cohort 10. The survey is managed and funded by the Department for Children, Schools and Families, who jointly designed the incentive experiment

with the National Centre for Social Research, the survey contractors for waves 2 and 3 of YCS cohort 10.

3.1 The survey

The population studied in the YCS cohort 10 consisted of pupils in England and Wales who had reached minimum school leaving age of 16 in the 1998/1999 school year (Russell and Phelps 2001), that is, a one year age cohort of pupils born between 1-9-1982 and 31-8-1983. A 10% random sample was drawn from the registers of schools (excluding special schools and schools with fewer than 20 pupils of that age) in 1999, by asking schools to provide the names and addresses of pupils born on the 5th, 15th and 25th of every month. From the resulting file of 31,424 names and addresses a systematic random sample of 25,000 pupils was drawn. The first wave of the survey took place a year later in spring 2000, the second at the end of 2000, the third in spring 2002 and the fourth in spring 2003. Nonrespondents were not issued in subsequent waves and, as a result, attrition was monotonic.

Wave 1 was a postal survey with telephone follow-up of nonrespondents after 4 mailings (initial questionnaire mailing and three reminders). Based on reported examination results, wave 1 respondents were classified as either 'higher achievers' (if they had obtained 5 General Certificate of Secondary Education examination passes at grades A* to C) or 'lower achievers' otherwise. This led to around one-third of wave 1 respondents being classified as lower achievers. At wave 2 roughly one third of issued sample members were randomly selected for additional questions on particular topics and assigned to computer assisted telephone interviewing (CATI). In addition to the core questionnaire, telephone respondents were administered a module on decisions about entering higher education (for higher achievers) or on educational and employment aspirations (for lower achievers). The remaining sample members were administered the core questionnaires by post. At wave 3 all lower achievers received the core mail questionnaire, although the telephone module continued to be carried for a third of higher achievers. At wave 4 all respondents were assigned to the core postal survey. Figure 1 illustrates the allocation to modes and incentives.

The core questionnaire remained mainly unchanged for the three experimental waves. Telephone respondents were asked the core questions before the additional modules. The core questionnaire was the same as the postal questionnaire, although some items were adapted for administration over the telephone. The average telephone interview took around 20 minutes. (The questionnaires and technical reports are available via the UK Data Archive in the appendices of the YCS User Guide at http://www.data-archive.ac.uk/findingdata/snDescription.asp?sn=4571&key=YCS.)

3.2 The incentives experiment

Facing growing concerns over declining response rates, an experiment was introduced in the second wave of cohort 10, to study the effect of incentive payments on response rates and nonresponse bias. A proportion of wave 1 respondents on both the postal and telephone surveys were sent a GBP5 voucher (approx. USD10 or EUR7), while the control groups received no such incentive. Additionally, in the postal survey the incentives were either unconditional (the incentive was sent with the initial mailing) or conditional (the voucher was promised in the original mailing, but only sent on receipt of a completed questionnaire). At waves 3 and 4, all incentives were paid unconditionally.

Wave 1	Postal core questionnaire. Reported exam results used to classify respondents for wave 2 allocation.									
	Higher Achievers					Lower Achievers				
Wave 2	T_x	T_u	P_x	P_u	P_c	T_x	T_u	P_x	P_u	P_c
Wave 3	T_x	T_u	P_x	P_u	--	--	--	P_x	P_u	--
Wave 4	--	--	P_x	P_u	--	--	--	P_x	P_u	--
Questionnaire	Core + Higher education		Core			Core + Education and employment		Core		

Notes: T = telephone, P = postal, x = control, u = unconditional incentive, c = conditional incentive. Arrows indicate changes in incentive treatment or mode allocation between waves.

Figure 1 Experimental design

3.3 Allocation of respondents to modes and incentive treatments

At wave 2, wave 1 respondents were randomly assigned to either telephone or postal mode. The allocation of incentive treatments was however done at the school level (randomised cluster assignment by mode). Each school represented in the sample was allocated to one telephone treatment (control or incentive) and independently allocated to one, potentially different, postal treatment (control, unconditional incentive or conditional incentive), so that all sample members from the same school approached in the same mode received the same incentive treatment.

The 4,712 wave 1 lower achiever respondents were stratified by identification number within school within Government Office Region and alternately allocated to telephone and postal treatments. We have excluded from the analysis 627 cases for which there was no valid telephone number on file, as those amongst this group who had been allocated to telephone mode were approached by post. Consequently, analysis of lower achievers is restricted to 2,097 approached by telephone and 1,988 approached by post.

A similar procedure was carried out for higher achievers, except that a larger proportion was allocated to postal treatment. There were 8,909 wave 1 higher achiever respondents of which 751 had no valid telephone number and are excluded from the analysis. After these exclusions there are 2,922 higher achievers allocated to telephone mode and 5,236 allocated to postal mode.

For the allocation of schools to incentive treatment groups, the schools containing telephone sample members (i.e. all schools apart from a few of the very smallest schools with fewer than five pupils in the sample) were stratified according to the ratio of lower to higher achievers in the sample and randomly assigned to incentive treatments within strata. (The proportion of schools assigned to incentives was 1/2 if the ratio of lower to higher achievers in the sample was ≥ 2; 1/3 for $1/2 \leq$ ratio < 2 and 1/4 for all remaining schools.) The procedure was repeated for the allocation of schools in the postal treatment groups, where those selected for incentive treatment were randomly split into a conditional and an unconditional treatment group. (The proportions allocated to incentives were 2/3 if the ratio was ≥ 2; 1/3 for $0 \leq$ ratio $< 1/2$ and 1/6 for all other schools.) All estimates of significance presented in this text account for the clustered sampling design of the incentive experiment.

Table 1 shows the issued sample sizes at each wave for the different treatment and mode combinations, excluding cases of known ineligibility who had either moved abroad or died ($n = 13$ at wave 2; $n = 3$ at wave 3). Ineligible cases at wave 4 are not identified in the data, but the number is likely to be small. The analysis also excludes wave 1 respondents for whom no telephone number was known at the time of the allocation to modes for wave 2, as described above, and 117 higher achievers assigned to telephone mode at wave 2, who responded by post and were subsequently allocated to postal mode.

Table 1 also documents the observed wave-on-wave and cumulative response rates (AAPOR RR1). The rates are shown by achievement level and sequential mode/incentive combination. Wave-on-wave response rates for the higher achiever sample allocated to telephone control at wave 2 and moved to postal control at wave 4 (Col 1) were, for example, 76.82%, 69.13% and 72.21%. The issued numbers of cases declined from 2,075 to 1,101 across the three waves, because nonrespondents were not issued in subsequent waves.

Table 1 Conditional and cumulative response rates

Wave	Response Rate %	Higher Achievers					Lower Achievers				
		$T_xT_xP_x$	$T_uT_uP_u$	$P_xP_xP_x$	$P_uP_uP_u$	$P_cP_uP_u$	$T_xP_xP_x$	$T_uP_uP_u$	$P_xP_xP_x$	$P_uP_uP_u$	$P_cP_uP_u$
2	Conditional	76.82	80.91	78.23	86.45	82.32	65.21	70.41	64.93	75.00	71.35
	(Issued *n*)	(2,075)	(728)	(3,262)	(1,004)	(967)	(1,282)	(811)	(807)	(608)	(569)
3	Conditional	69.13	73.17	73.07	81.91	81.36	59.09	70.93	63.36	71.93	70.20
	(Issued *n*)	(1,594)	(589)	(2,551)	(868)	(794)	(836)	(571)	(524)	(456)	(406)
	Cumulative	53.11	59.20	57.16	70.82	66.94	38.53	49.94	41.14	53.95	50.09
4	Conditional	72.21	85.61	76.11	85.65	86.82	63.16	74.26	65.36	75.30	81.34
	(Issued *n*)	(1,101)	(431)	(1,863)	(711)	(645)	(494)	(404)	(332)	(328)	(284)
	Cumulative	38.31	50.69	43.48	60.66	58.03	24.34	36.99	26.89	40.63	40.60

Notes: AAPOR Response Rate 1. Treatment groups are identified by T = telephone, P = postal, x = control, u = unconditional incentive, c = conditional incentive. $T_x T_x P_x$ for example, refers to the sample allocated to telephone control at waves 2 and 3 and to postal control at wave 4. Conditional response rates are conditional on response at the previous wave. The base is the number of issued cases, which excludes previous wave nonrespondents and ineligible cases. Cumulative response rates are the percentage of wave 1 respondents remaining in the respondent sample. The base is the wave 2 number of issued cases, excluding three higher achievers ineligible at wave 3 (1 $P_x P_x P_x$ and 2 $P_c P_u P_u$).

4. Outcome measures and methods

The analysis is based on the sample of wave 1 respondents, since allocation to experimental treatments used information collected in the first wave and the corresponding characteristics of wave 1 nonrespondents are unknown. Our focus is therefore on attrition, conditional upon wave 1 response. This is the aspect of non-response that is particular to panel surveys, though of course it must be recognised that the characteristics of attrition are conditional on the characteristics of wave 1 response. The response rate at wave 1 (AAPOR RR1) was 54.80%, excluding 5 cases of known ineligibility (Russell and Phelps 2001). This section describes the outcome measures and methods used to evaluate the hypotheses about the effects of incentives.

4.1 Attrition

To test the effect of incentives on attrition, we estimated the probability of attrition as a function of the experimental design variables (*tel*ephone mode, *unc*onditional incentives, *cond*itional incentives, *l*ower *a*chievers) and their interactions. For each of the three experimental waves ($t = 2, 3, 4$), we estimated a separate probit model of the probability of attrition, in each case using the wave 1 respondent sample as the base:

$$\begin{aligned}\Pr(attrition_{it}) = F(&\beta_{0t} + \beta_{1t} tel_i + \beta_{2t} unc_i + \beta_{3t} cond_i \\ &+ \beta_{4t} la_i + \beta_{5t} tel_i * unc_i + \beta_{6t} la_i * tel_i \\ &+ \beta_{7t} la_i * unc_i + \beta_{8t} la_i * cond_i \\ &+ \beta_{9t} la_i * tel_i * unc_i + \varepsilon_i) \end{aligned} \quad (1)$$

where F is the probit link function. The estimated coefficients and standard errors from this model were then used to calculate predicted probabilities of attrition under different treatment conditions and to test for differences due to incentives.

4.2 Item nonresponse

To test the effect of incentives on item nonresponse, we estimated count models of the number of items missing, using all non-filtered items from the core questionnaires in waves 2 ($n = 44$), 3 ($n = 48$) and 4 ($n = 46$), where 'don't know' was counted as a missing value. We used the same specification of the predictors as for model (1) to estimate separate negative binomial regression models for each of the three experimental waves, conditional on response to the given wave. (Overdispersion meant that Poisson models did not fit the data: the P-value of the Likelihood Ratio test of equal mean and variance was 0.0000 for all three waves.) The estimated coefficients and standard errors from these models were used to calculate predicted item nonresponse under different treatment conditions and to test for differences due to incentives.

4.3 Attrition bias

To test the effect of incentives on attrition bias, we estimated the probability of attrition using model (1) but including wave 1 respondent characteristics and their interactions with the experimental design variables as predictors. We estimated separate probit models for attrition at each of the experimental waves ($t = 2, 3, 4$) and for each characteristic, again using the wave 1 respondent sample as the base:

$$\begin{aligned}\Pr(attrition_{it}) = F(&\beta_{0t} + \beta_{1t} tel_i + \beta_{2t} unc_i + \beta_{3t} cond_i \\ &+ \beta_{4t} la_i + \beta_{5t} tel_i * unc_i + \beta_{6t} la_i * tel_i \\ &+ \beta_{7t} la_i * unc_i + \beta_{8t} la_i * cond_i \\ &+ \beta_{9t} la_i * tel_i * unc_i + \beta_{10} w1char \\ &+ \beta_{11} w1char_i * tel_i + ... \\ &+ \beta_{19} w1char_i * la_i * tel_i * unc_i + \varepsilon_i) \end{aligned} \quad (2)$$

where β_{11} to β_{19} are the coefficients for the interactions of the characteristic with the design variables. The coefficient for the respondent characteristic, β_{10}, provides information about the direction, magnitude and, in combination with its standard error, the significance of attrition bias for the postal, no incentive, higher achiever reference group. The interaction of the characteristic and the incentive indicators provide information about the change in attrition bias due to incentives. The significance of all interactions presented in this text was calculated following recommendations for nonlinear models by Norton, Wang and Ai (2004) using the command 'predictnl' in Stata version 9.

The characteristics tested were gender, school type, exam results, current activity (full-time education, employment, not in education, employment or training ("neet")), experience of unemployment, studying for vocational or academic qualifications, household composition (living with parent, partner, neither) and a set of attitudinal questions about employment and training. The wording of all questions is documented in Table 6. The characteristics chosen were those for which respondents and non-respondents could be expected to differ, based on previous studies of nonresponse in the YCS and other surveys and on nonresponse theories (Groves and Couper 1998; Lynn, Purdon, Hedges and McAleese 1994).

4.4 Reported results

Since coefficients from non-linear models cannot be interpreted substantively (Long 1997), we report predicted values based on the model estimates, rather than coefficients. Unless stated otherwise, the results are for the higher achiever group. To convey a sense of the magnitude of differences in outcomes across treatments, we report transformations of the predicted values, comparing each treatment with the comparison group, the higher achiever postal control.

5. Attrition, item nonresponse and attrition bias in the control groups

As a background to the evaluation of the effects of incentives, this section documents the extent of attrition, item nonresponse and attrition bias in the control groups, highlighting differences across waves, achievement levels and modes. Throughout the discussion the higher achiever postal no-incentive group is the reference category, with which all other treatments are compared.

5.1 Attrition

The predicted cumulative attrition rate among higher achievers allocated to the postal control group, increased from 21.77% in wave 2 to 56.53% in wave 4 (Table 2, Col 1). For *lower achievers* (Col 2), attrition rates in the postal control group were 61% higher at wave 2, but this difference decreased across waves to 29% at wave 4. The difference by achievement level was nonetheless significant in all three waves (P-value of $\beta_4 = 0.0000$ for $t = 2, 3, 4$). In *telephone mode* (Col 3), attrition rates in the control group were not significantly different at wave 2, but 9% higher at wave 3 (P-value of $\beta_5 = 0.0034$ for $t = 3$). This is contrary to findings from other studies, where nonresponse is generally lower in telephone mode due to the role of the interviewer in persuading respondents to take part in the survey. One possible reason for finding the opposite in this study is that for both the postal and CATI treatment groups, further attempts to obtain responses from initial non-respondents were made by telephone, so that only the postal group had a multi-mode treatment. Secondly, the burden of the wave 2 survey (measured by the interview length) was higher for the telephone respondents due to the additional modules, possibly leading to higher nonresponse at wave 3 than among the postal sample. The predicted cumulative response rates, which were the base for the calculation of percentage differences across treatment groups, are documented in the first three columns of Table 5.

5.2 Item nonresponse

The predicted number of missing items in the higher achiever postal control group was 2.89 at wave 2, falling to 1.75 at wave 4 (Table 3, Col 1). For *lower achievers* (Col. 2), the expected count for the control group was 21% higher at wave 2, with the gap increasing to 45% at wave 4. The differences by achievement level were significant in all three waves ($P \leq 0.0001$ for β_4, $t = 2, 3, 4$). For *telephone mode* (Col 3), the predicted count was 4% lower at wave 2 and 12% lower at wave 3 ($P = 0.0000$ for β_5, $t = 2, 3$), compared with postal mode. The predicted item non-response counts, used as the base for the calculations presented in Table 3, are documented in columns 4 to 6 of Table 5.

5.3 Attrition bias

Nonresponse in the higher achiever postal control group was differential for all of the domains tested (Table 4). The respondent samples significantly over-represented those living with their parents, in full-time education or studying for academic qualifications. Predicted attrition rates for those in full-time education in the higher achiever postal control group, for example, were 14% lower than for those not in full-time education at wave 2, with the difference increasing to 17% by wave 4 ($P = 0.0000$ for β_{10}, $t = 2, 3, 4$). At the same time, the respondent samples under-represented males, those in secondary modern schools, with low or no exam results, who thought employers did not give young people the right training and that making plans for the future was a waste of time, those in full-time employment, those who had experienced unemployment and those who were studying for vocational qualifications. Bias was particularly strong with respect to qualifications. Those without any or with very low exam qualifications were around 50% more likely to have attrited from the sample by waves 3 and 4, compared to sample members with better qualifications. Similarly, those in full-time employment were 17% more likely than those not in employment (most of whom were still in education) to drop out at wave 2, with the difference increasing to 22% by wave 4.

Including background information used by the YCS for weighting (gender, school type, exam results and region) in the models did not affect the bias for any of the characteristics (in each wave and for each item, the P-value > 0.05 from Wald tests of the equality of β_{10} estimated with and without background characteristics; not reported), except for bias with respect to qualifications, which was somewhat reduced when the background information was included.

The extent of attrition bias was mostly stable across waves, except for a few characteristics. In the higher achiever postal control sample, the under-representation of males significantly increased from waves 2 to 4 (P-value

from a Wald test of the equality of β_{10} across the two waves = 0.0295; not reported). For some of the other characteristics, the bias significantly decreased across waves. Nonresponse bias associated with attending a modern school fell between waves 3 and 4 and bias associated with not having any qualifications fell between waves 2 and 3 and again between waves 3 and 4.

For lower achievers there were few differences in the extent of attrition bias (not reported). Bias by gender, that is the difference in predicted nonresponse rates between males and females, was 12% less than for higher achievers at wave 4 (*P*-value of the interaction between achievement level and gender was 0.0425 for $t = 4$), and bias by full-time employment was 4% less at wave 2 (*P*-value = 0.0269 for $t = 2$); bias according to attitudes on training provided by employers was 9% higher at wave 2 (*P*-value = 0.0056); bias according to whether studying for academic or vocational qualifications was higher at wave 2 (22% and 13%), 6% lower and 1% higher at wave 3, and lower at wave 4 (81% and 92%).

Table 2 Effect of incentives on attrition rates

	Control groups			Incentives		Incentives by ability		Incentives by mode and ability	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Wave	P_x^{ha}	$\frac{P_x^{la} - P_x^{ha}}{P_x^{ha}}$	$\frac{T_x^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{P_u^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{(P_c^{ha} - P_x^{ha})/P_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(P_c^{la} - P_x^{la})/P_x^{la}}{(P_c^{ha} - P_x^{ha})/P_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	21.77	0.6112	0.0650	-0.3777	0.4966	0.7602	0.9763	0.4669	0.8471
(*P*-Value)		(0.0000)	(0.2268)	(0.0000)	(0.0142)	(0.5085)	(0.4332)	(0.0556)	(0.6810)
3	42.86	0.3734	0.0941	-0.3191	0.7066	0.6820	0.6743	0.4074	1.4275
(*P*-Value)		(0.0000)	(0.0034)	(0.0000)	(0.0592)	(0.7834)	(0.8287)	(0.0057)	(0.0861)
4	56.53	0.2933	-	-0.3040	0.8402	0.6179	0.7340	0.6597	0.8338
(*P*-Value)		(0.0000)	-	(0.0000)	(0.2244)	(0.2535)	(0.8177)	(0.0911)	(0.9265)

Notes: *P* = postal, *T* = telephone, *x* = control, *u* = unconditional incentive, *c* = conditional incentive, *ha* = higher achievers, *la* = lower achievers. Column (1) shows the predicted attrition rate for the postal control higher achiever sample. The remaining columns show proportionate change in predicted rates. *P*-values of columns 2-4 represent standard errors of the main effects in the probit model; column 5 represents *P*-values from a Wald test of the equality of the coefficients for conditional and unconditional incentives; columns 7-9 represent *P*-values for the relevant interactions calculated using 'predictnl' in Stata version 9, according to Norton *et al.* (2004).

Table 3 Effect of incentives on item nonresponse (counts)

	Control groups			Incentives		Incentives by ability		Incentives by mode and ability	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Wave	P_x^{ha}	$\frac{P_x^{la} - P_x^{ha}}{P_x^{ha}}$	$\frac{T_x^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{P_u^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{(P_c^{ha} - P_x^{ha})/P_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(P_c^{la} - P_x^{la})/P_x^{la}}{(P_c^{ha} - P_x^{ha})/P_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	2.89	0.2068	-0.9579	0.1008	1.3849	2.4825	0.6927	0.1820	-0.4094
(*P*-Value)		(0.0005)	(0.0000)	(0.0173)	(0.4790)	(0.1308)	(0.6472)	(0.6251)	(0.9202)
3	2.54	0.3879	-0.8828	0.1660	1.5599	1.6788	1.2445	-0.9526	-0.1378
(*P*-Value)		(0.0001)	(0.0000)	(0.0049)	(0.2372)	(0.4339)	(0.6796)	(0.0442)	(0.3890)
4	1.75	0.4533	-	0.0085	17.5491	16.8405	0.4621	13.8706	2.3073
(*P*-Value)		(0.0013)	-	(0.9262)	(0.2133)	(0.4724)	(0.6481)	(0.5049)	(0.4530)

Notes: *P* = postal, *T* = telephone, *x* = control, *u* = unconditional incentive, *c* = conditional incentive, *ha* = higher achievers, *la* = lower achievers. Column (1) shows the predicted number of missing items of 44 non-branched items at wave 2, 48 at wave 3 and 46 at wave 4. The remaining columns show proportionate change in predicted item nonresponse counts. *P*-values of columns 2-4 represent standard errors of the exponentiated coefficients from the count model; column 5 represents *P*-values from a Wald test of the equality of the exponentiated coefficients for conditional and unconditional incentives; columns 7-9 represent *P*-values for the relevant interactions calculated using'predictnl' in Stata version 9, according to Norton *et al*.(2004).

Table 4 Attrition bias (higher achiever postal control group)

	Wave 2	*P*-Value	Wave 3	*P*-Value	Wave 4	*P*-Value
Male	0.0807	(0.0000)	0.1330	(0.0000)	0.1474	(0.0000)
School type						
Comprehensive 16	0.0196	(0.2645)	0.0102	(0.6178)	0.0259	(0.2060)
Comprehensive 18	-0.0197	(0.1966)	-0.0138	(0.4444)	-0.0200	(0.2650)
Selective	-0.0188	(0.3661)	-0.0547	(0.0407)	-0.0213	(0.4577)
Modern	0.2310	(0.0001)	0.2423	(0.0004)	0.1597	(0.0261)
Independent	-0.0142	(0.4639)	0.0147	(0.5245)	-0.0068	(0.7756)
Exam results						
5+ grades A-C	-0.0977	(0.1778)	-0.0866	(0.3060)	-0.1795	(0.0320)
1-4 grades A-C	0.0831	(0.2857)	0.0721	(0.4298)	0.1696	(0.0606)
5+ grades D-G	0.0324	(0.8769)	0.0715	(0.7739)	0.1849	(0.4536)
1-4 grades D-G	-0.2177	(0.0000)	0.5714	(0.0000)	0.4347	(0.0000)
None	0.7826	(0.0000)	0.5716	(0.0000)	0.4348	(0.0000)
Attitudes						
Employers don't give training	0.0842	(0.0000)	0.0882	(0.0000)	0.0798	(0.0001)
Training more important than pay	0.0108	(0.4808)	-0.0070	(0.6979)	-0.0062	(0.7370)
Plans for future are a waste of time	0.0656	(0.0959)	0.1457	(0.0015)	0.1371	(0.0030)
Information about opportunities	0.0034	(0.8431)	-0.0204	(0.3266)	-0.0236	(0.2549)
Enough support planning future	0.0063	(0.6771)	0.0043	(0.8233)	-0.0105	(0.5848)
Current activity						
In full-time education	-0.1371	(0.0000)	-0.1462	(0.0000)	-0.1728	(0.0000)
In full-time employment	0.1661	(0.0003)	0.1983	(0.0001)	0.2201	(0.0000)
Neither in employment, education or training	0.0898	(0.1387)	0.1036	(0.1495)	0.1098	(0.1184)
ILO unemployed	0.0112	(0.6272)	0.0573	(0.0421)	0.0475	(0.0879)
Unemployed during past 12 months	0.0246	(0.4216)	0.0731	(0.0523)	0.0891	(0.0146)
Studying for academic qualifications	-0.1173	(0.0000)	-0.1351	(0.0000)	-0.1341	(0.0000)
Studying for vocational qualifications	0.0677	(0.0001)	0.0882	(0.0000)	0.0721	(0.0003)
Living arrangements						
Living with parent	-0.1348	(0.0111)	-0.1916	(0.0027)	-0.1033	(0.0986)
Living with partner	0.0904	(0.4457)	-0.0441	(0.7475)	-0.1042	(0.4525)

Notes: Predicted differences in attrition rates based on $\hat{\beta}_{10,t}$, *i.e.*, prediction for each category compared to all residual categories. Each table entry is from a different model as explained in the text. *P*-values based upon estimated standard errors of the coefficient for the characteristic in the probit model.

Attrition bias in telephone mode was no different from postal mode, except for differential nonresponse by gender: the bias was 7% less at wave 2, 2% less at wave 3 and 1% more at wave 4 (*P*-value of the interaction between telephone mode and gender was ≤ 0.002 for $t = 2, 3, 4$).

6. Evaluation of hypotheses

The evidence discussed here is summarised in Table 2 (effects of incentives on attrition), Table 3 (effects on item nonresponse), Table 4 (effects on attrition bias) and Table 5 (net effect on unit and item nonresponse).

*H*1: *Effects of incentives on attrition rate, attrition bias and item nonresponse.*

Incentives reduced attrition and increased item nonresponse but did not impact on attrition bias. Unconditional incentives reduced cumulative *attrition* in the postal higher achiever sample (Table 2, Col 4) by 38% (corresponding to an 8 percentage point difference) at wave 2, 32% at wave 3 and 30% at wave 4 (*P*-value of $\beta_2 = 0.0000$ for $t = 2, 3, 4$). At the same time, the incentive increased *item nonresponse* by 10% at wave 2 and 17% at wave 3 (*P*-value of $\beta_2 \leq 0.05$ for $t = 2, 3$), but had no effect at wave 4 (Table 3, Col 4). The difference across waves was however not significant (see *H*2).

Incentives had a proportionate effect on attrition across all respondent characteristics tested and therefore did not reduce *attrition bias*: the *P*-value of the interaction of unconditional incentives and respondent characteristics was > 0.05 for all characteristics and waves (not reported). The exception was the proportion of pupils in 'modern' schools who were under-represented in all three waves. (Modern schools were the smallest category, representing only 2.8% of the wave 1 respondent sample.) Unconditional incentives reduced this bias by 60%, 47% and 78% at waves 2, 3 and 4 respectively (*P*-values of the interaction of incentives and modern school ≤ 0.01 for $t = 2, 3, 4$).

Since incentives had a positive effect on unit response and a negative effect on item response, Table 5 documents the net effect on the amount of information collected in the survey. The benefits of incentives in terms of unit nonresponse clearly outweighed the cost in terms of item nonresponse. For each sample person issued at wave 2, the predicted unit and item response

rates for the postal higher achiever sample implied that by wave 4, 40% more valid items were collected with unconditional incentives compared to the control group. For lower achievers, 50% more information was collected with incentives. This is, however, a crude measure of the net effect of incentives, since in a multivariate analysis or for analyses of change, different patterns of missingness across items or across waves may lead to large numbers of cases being dropped by pairwise deletion.

*H*2: *Effects of incentives across waves.*
The effect on attrition decreased somewhat across waves, while the effects on item nonresponse and attrition bias were constant. Incentives reduced attrition by 38% at wave 2, 32% at wave 3 and 30% at wave 4 (Table 2, Col 4). The effects were similar at waves 2 and 3, but significantly different between waves 2 and 4 and between waves 3 and 4 (*P*-value from a Wald test of the equality of β_2 across waves was ≤ 0.05). Although the relative effect of incentives decreased, the absolute effect increased across waves (-17 percentage points at wave 4, compared to -8 and -14 at waves 2 and 3, see Table 5). The effect of incentives on *item nonresponse* was not significantly different across waves (*P*-value of equality of β_2 across waves was > 0.05), although the predicted numbers of missing items fell across waves. Similarly, the effects of incentives on *attrition bias* did not differ across waves.

*H*3: *Conditional compared to unconditional incentives.*
Unconditional incentives had a greater effect in reducing attrition than conditional incentives, but similar effects on item nonresponse and attrition bias. For higher achievers, the conditional incentives used at wave 2 were only half as effective at reducing attrition as unconditional incentives (Table 2, Col 5) and the difference between the two conditions was significant (*P*-value from a Wald test of the equality of β_2 and β_3 was 0.0142). At the same time, conditional incentives increased *item nonresponse* by 38% more than unconditional incentives (Table 3, Col 5), but the difference was not significant. Conditional incentives somewhat reduced *attrition bias* for a single characteristic: sample members in the control group studying for vocational qualifications at wave 1 were 6.8% more likely to drop out than those not studying for vocational qualifications. With conditional incentives the difference was 6.4% (*P*-value of the interaction of conditional incentives with this characteristic was ≤ 0.05 for $t = 2$).

*H*4: *Differential effects by mode.*
Incentives had more effect on attrition and item nonresponse in postal than telephone mode, but no effect on attrition bias in either mode. In telephone mode, unconditional incentives had less than half the effect on *attrition* they had in postal mode for the higher achiever group (Table 2, Col 8). The difference was significant at wave 3 (*P*-value of the interaction between telephone mode and unconditional incentives was 0.0057) but not at wave 2. At wave 3, incentives increased *item nonresponse* 5% less in telephone mode than in postal mode (*P*-value of the interaction was 0.0442), but the difference at wave 2 was not significant. The lack of effect of unconditional incentives on *attrition* was no different across the two modes.

*H*5: *Effects of changes in mode or incentive treatment.*
Changing the incentive condition or mode did not have lasting effects. Changing the treatment from conditional to unconditional incentives had no lasting effect on either *attrition* or *item nonresponse* ($P > 0.05$ from Wald tests of the equality of β_2 and β_3 for $t = 3, 4$) and the effects after the change in treatment were similar to those for the sample allocated to unconditional incentives from the start (Tables 2 and 3, Col 5). Changing the survey mode from telephone to postal did not have a lasting effect on *attrition* or *item nonresponse* either ($P =$ value of the interaction for telephone mode and unconditional incentives > 0.05 at $t = 4$) and the effects after the change in mode were no different from the effects for the sample allocated to postal unconditional incentives from the start (Tables 2 and 3, Col 8).

*H*6: *Differential effects by ability level.*
The effects of incentives were similar across achievement levels. Differences between achievement levels in the proportional effects of unconditional and conditional incentives on *attrition* and *item nonresponse*, were not significant (Cols 6 and 7 in Tables 2 and 3 report the *P*-values of the interactions of achievement level with each of the incentive treatments), since the absolute effects were comparable. Unconditional incentives, for example, reduced attrition at wave 2 by 8 percentage points among higher achievers and 10 percentage points among lower achievers. However, since the level of nonresponse in the control group was 61% higher for the lower achiever group, the similar absolute effect implied a smaller proportional effect of only 76% of the effect for higher achievers.
Similarly, the difference between modes was not differential by achievement (Tables 2 and 3, Col 9 report the *P*-values of the interaction between achievement level, unconditional incentives and telephone mode) and the lack of effect on attrition bias was no different for lower achievers (not reported).

Table 5 Net effect of incentives on unit and item response

		Predicted cumulative RR (%)			Predicted mean # INR			# valid items per unit issued at w2: incentive/control		
		w2	w3	w4	w2	w3	w4	w2	w3	w4
Higher	P_x	78.23	57.14	43.47	2.89	2.54	1.75	-	-	-
Achievers	P_u	86.45	70.82	60.66	3.19	2.96	1.77	1.097	1.228	1.395
	P_c	82.32	66.80	57.91	3.30	3.20	2.01	1.042	1.152	1.324
	T_x	76.82	53.11	38.31	0.12	0.30	1.61	-	-	-
	T_u	80.91	59.20	50.69	0.12	0.25	1.80	1.053	1.116	1.317
Lower	P_x	64.93	41.14	26.89	3.49	3.52	2.54	-	-	-
Achievers	P_u	75.00	53.95	40.63	4.37	4.51	2.91	1.130	1.282	1.498
	P_c	71.35	50.09	40.60	3.83	4.66	2.72	1.090	1.186	1.504
	T_x	65.21	38.53	24.34	0.50	3.48	2.35	-	-	-
	T_u	70.41	49.94	36.99	0.49	3.56	2.99	1.080	1.294	1.498

Notes: RR = response rate, INR = item nonresponse, # = number. T = telephone, P = postal, x = control, u = unconditional incentive, c = conditional incentive. Calculation based on 44 non-branched items at wave 2, 48 at wave 3 and 46 at wave 4. The number of valid items is calculated as RR_4*(44-INR_2 + 48-INR_3 + 46-INR_4).

Table 6 Question wording of items included in analysis of nonresponse bias

Variable	Question wording
Year 11 exam results	"Please tell us: a) Which GCSE subjects you studied in Years 10 and 11, b) Which GCSE subjects you have taken an exam in, c) Your GCSE results (do not record any re-sit results obtained in Year 11)."
Attitudes:	"Here are some things which people have said. We would like to know what you think. Please put a cross in one box for each statement: Agree, Disagree, Don't know."
ATT: employers	Agree: "Most employers don't give young people the right king of training at work."
ATT: training/pay	Agree: "In looking for a job, I am more concerned to find one with training than one that pays the best."
ATT: plans	Agree: "I think that making plans for the future is a waste of time."
ATT: information	Agree: "I know how to find out about future work, training or education opportunities."
ATT: support	Agree: "I get enough support in planning my future."
Current activity:	"Please put a cross against one box to tell us your main activity at the moment: a) Out of work/unemployed, b) Modern Apprenticeship, National Traineeship, Youth Training or other government supported training, c) In a full-time job (over 30 hours a week), d) In a part-time job (if this is your *main* activity), e) In full-time education at school or college, f) Looking after home or family, g) Doing something else (please specify)."
In ft education	In full-time education.
In ft employment	In full-time employment.
NEET	Not in employment, education or training.
ILO unemployed	Unemployed and searching for job among economically active (YCS derived variable).
Unemployed	Unemployed in one or more months from April 1999 to March 2000: "We would also like to know what you have been doing over the past months. Please put a cross in one box for each month to show us what you were doing for all, or most of each month".
	Response options as for current activity, including 'On holiday'.
Studying (ac)	Yes: "At present, are you studying for GCSE, A/S or A-level qualifications?"
Studying (voc)	Yes: "At present, are you studying for any GNVQs (General National Vocational Qualifications)?" or "At present, are you studying for NVQ (National Vocational Qualification) or any other vocational or professional qualification including BTEC, City & Guilds or RSA qualifications?"
Household:	"Who lives in the same household as you? a) Father, b) Stepfather, c) Mother, d) Stepmother, e) Your own children, f) Brothers and sisters g) Other persons (please write in their relationship to you)."
Living with parent	Living with one or more of father, stepfather, mother or stepmother.
Living with partner	Living with boyfriend, girlfriend, husband, wife or partner.

7. Summary and discussion

This study has provided new evidence on the effects of continued incentive payments in a multi-mode panel study. We tested the effects of incentives on attrition, item nonresponse and attrition bias and whether these effects changed across waves. We also tested whether conditional and unconditional incentives had similar effects, whether incentive effects were differential across modes and ability levels, and whether changes in the incentive treatment or mode had lasting impact on the effect of incentives in subsequent waves.

The findings showed that unconditional incentives significantly reduced attrition and, although they also

increased item nonresponse, the net effect on the amount of information collected by the survey was positive. Incentives had proportionate effects across a range of respondent characteristics and as a result did not impact on attrition bias in terms of those characteristics. Item nonresponse increased more with unconditional than conditional incentives, and more in postal than in telephone mode. Attrition bias was not affected by either incentive treatment in either mode. Across waves, incentives had a somewhat decreasing effect on attrition, but similar effects on item nonresponse. The lack of effect on attrition bias was also a constant across waves. Changes in incentive treatment from conditional to unconditional, and in mode from telephone to postal, did not affect outcomes at later waves.

The findings imply that respondent incentives are an effective means of maintaining sample sizes of a panel and ensuring its value in terms of efficiency of estimation and feasibility of subgroup analyses. Among lower achievers, fully 50% more information was collected during the three experimental waves, in terms of the number of valid items per case issued at the start. Incentives were safe, in the sense that increased response rates did not inadvertently increase nonresponse bias in terms of observed characteristics.

Changes in incentive treatment did not have lasting effect; however, in this study the only change implemented was an improvement for the respondent, from conditional to unconditional incentives. Expectations formed on the basis of previous incentive treatments may well mean that changes have lasting effect, if the change reduces the value of the incentive in the eyes of the respondent (see, Singer, Van Hoewyk and Maher 1998).

Incentives had no effect on attrition bias. We could however not evaluate the effect on bias of nonresponse at wave 1. Ideally, we would assess both the magnitude of bias due to nonresponse at wave 1 and due to subsequent attrition, and the effects of incentives on both. It is possible that nonresponse at wave 1 is more detrimental in terms of bias than later attrition, especially in studies such as the present one with low initial response rates. In this case, the effect of incentives on bias at wave 1 may be more important than any effect on bias caused by attrition. In addition, the discussion of the effects of incentives on attrition bias has focused entirely on observed characteristics and although incentives did not have differential effects in terms of these, they may nonetheless have differential effects in terms of unobserved factors. If this were the case, the use of respondent incentives could introduce sample selection bias in multivariate estimates, if the unobservables determining the responsiveness to incentives are correlated with outcomes measured by the survey (Kennedy 2003). For example, if responsiveness to incentives depends on time preferences for money and this factor also determines the decision to leave further education and work instead, then models of the determinants of educational outcomes will lead to biased estimates.

Finally, there was little evidence that the respondent sample became less sensitive to incentives across waves as potentially less committed sample members dropped out. This finding is consistent with Laurie (2007), who reported that an increase in the value of an incentive in the British Household Panel Survey significantly increased response, even after 14 waves of the panel, with already high annual response rates of around 95% each year. Since previous studies have found that the effects of one-off incentives can carry over across waves (James 1997; Laurie and Lynn in press; Mack *et al.* 1998), a formal test of marginal effects of incentives would however require comparisons with a treatment group only offered an incentive at the first wave.

Acknowledgements

We would like to thank Iain Noble at the Department for Children Schools and Families for facilitating the data and commenting on an earlier version of the paper, Tim Thair and Rory Fitzgerald for patient and helpful response to data queries, Noah Uhrig, Heather Laurie for comments and Mark Bryan for advice on the calculation of interactions in nonlinear models. The views expressed are those of the authors and not necessarily those of the Department for Education and Skills.

References

AAPOR (2006). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 4th Edition: American Association for Public Opinion Research.

Armstrong, J.S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly*, 39, 111-116.

Berk, M.L., Mathiowetz, N.A., Ward, E.P. and White, A.A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics*, 3, 449-457.

Church, A.H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.

Davern, M., Rockwood, T.H., Sherrod, R. and Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment. *Public Opinion Quarterly*, 67, 139-147.

Downs, B. (1999). *Incentive Use in Panel Surveys*. Internal Census Bureau Memorandum. Washington, DC: US Census Bureau.

Fox, R.J., Crask, M.R. and Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161–178.

Goyder, J. (1994). An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society*, 36, 360-366.

Groves, R., Singer, E. and Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.

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

Hopkins, K.D., and Gullickson, A.R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education*, 61, 52-62.

James, J.M., and Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly*, 54, September 1, 1990, 346-361.

James, T.L. (1997). Results of wave 1 incentive experiment in the 1996 survey of income and program participation. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.

Kennedy, P. (2003). *A Guide to Econometrics*. Oxford: Blackwell.

Laurie, H. (2007). The effect of increasing financial incentives in a panel survey: An experiment on the british household panel survey, Wave 14. *ISER Working Paper* No. 2007-05. Colchester: University of Essex.

Laurie, H., and Lynn, P. (in press). The use of respondent incentives on longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed. P. Lynn), Chichester: Wiley.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Lynn, P., Purdon, S., Hedges, B. and McAleese, I. (1994). England and Wales Youth Cohort Study: An Assessment of Alternative Weighting Strategies. Employment Department Research Series YCS No. 30. Sheffield: Employment Department.

Mack, S., Huggins, V., Keathley, D. and Sundukchi, M. (1998). Do monetary incentives improve response rates in the survey of income and programme participation? *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.

Martin, E., Abreu, D. and Winters, F. (2001). Money and motive: effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17, 267-284.

Mason, R., Lesser, V. and Traugott, M.W. (2002). Effect of item nonresponse on nonresponse error and inference. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little), New York: John Wiley & Sons, Inc., 149-161.

Norton, E.C., Wang, H. and Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4, 154-167.

Rodgers, W. (2002). Size of incentive effects in a longitudinal study. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.

Russell, N., and Phelps, A. (2001). Youth Cohort Study Cohort 10 Sweep 1: Technical Report. London: Taylor Nelson Sofres.

Ryu, E., Couper, M.P. and Marans, R.W. (2006). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; Response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18, 89-106.

Shettle, C., and Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15, 231-250.

Singer, E., Hoewyk, J.V., Gebler, N., Raghunathan, T. and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15, 217-230.

Singer, E., Van Hoewyk, J. and Maher, M.P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly*, 62, June 1, 1998, 152-164.

Singer, E., Van Hoewyk, J. and Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, June 1, 2000, 171-188.

Stratford, N., Simmonds, N. and Nicolaas, G. (2003). *National Travel Survey 2002*: *Report on Incentives Experiment*. London: National Centre for Social Research.

Teisl, M.F., Roe, B. and Vayda, M. (2005). Incentive effects on response rates, data quality, and survey administration costs. *International Journal of Public Opinion Research*, 18, 364-373.

Tzamourani, P., and Lynn, P. (1999). The Effect of Monetary Incentives on Data Quality - Results from the British Social Attitudes Survey 1998 Experiment. CREST Working Paper No. 73. Oxford: University of Oxford.

Ward, R.K., Boggess, S., Selvavel, K. and McMahon, M.F. (2001). The use of targeted incentives to reluctant respondents on response rates and data quality. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.

Willimack, D.K., Schuman, H., Pennell, B.-E. and Lepkowski, J.M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly*, 59, March 1, 1995, 78-92.

Yu, J., and Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 119-126
Statistics Canada, Catalogue No. 12-001-X

Balancing sample design goals for the National Health and Nutrition Examination Survey

Leyla Mohadjer and Lester R. Curtin [1]

Abstract

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States National Center for Health Statistics. A unique feature of NHANES is the administration of a complete medical examination for each respondent in the sample. To standardize administration, these examinations are carried out in mobile examination centers. The examination includes physical measurements, tests such as eye and dental examinations, and the collection of blood and urine specimens for laboratory testing. NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The major analytic goals of NHANES include estimating the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors. The sample design for NHANES must create a balance between the requirements for efficient annual and multiyear samples and the flexibility that allows changes in key design parameters to make the survey more responsive to the needs of the research and health policy communities. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the goals of NHANES.

Key Words: Multistage Sampling, Domain Sampling, Weighted Measure of Size, Mobile Examination Centers.

1. Introduction

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States Centers for Disease Control and Prevention through its National Center for Health Statistics (NCHS). The NHANES surveys have been used to assess the health and nutritional status of the non-institutionalized civilian population of the United States for over 45 years. The data collected in NHANES are used to estimate the prevalence of major diseases and risk factors for diseases. The nutritional data from NHANES provide temporal monitoring for the nation with respect to such factors as diet, cholesterol, hypertension, iron deficiency, anemia, and obesity. NHANES has also been designed to assess the relationship between diet, health, and the environment so that nutritional assessments can be linked to such diseases as cardiovascular disease, diabetes, hypertension, and osteoporosis.

Data collection for NHANES includes at least three stages: a household screener, an interview, and a medical examination. The primary objective of the screener is to determine whether any household members are eligible for the interview and examination. The screener collects basic information on household composition and demographic characteristics. The interview collects household-, family-, and person-level data on demographic and socioeconomic background, health, and nutritional characteristics. Upon completion of the interview, respondents are asked to participate in a medical examination. To standardize administration and protocols, these examinations are carried out in a specially designed and equipped mobile examination center (MEC). The examination includes physical measurements, tests such as eye and dental examinations, physiological measurements, and the collection of blood and urine specimens for laboratory testing. The NHANES website (http://www.cdc.gov/nchs/ nhanes.htm) provides detailed information about the NHANES medical components.

The development of an efficient design has involved consideration of several design issues unique to NHANES in addition to the ones normally involved in survey samples. This paper is focused on the unique and challenging aspects of the NHANES design. However, it is helpful to provide an overall summary of the NHANES design, as given below, before discussing the unique features.

The NHANES sample represents the total non-institutionalized civilian population of the United States. Active military and institutionalized persons are not part of the population of inference. NHANES is not an equal probability design; sampling fractions are set to "over-sample" Mexican Americans (and Hispanics in the 2007 and beyond samples), black Americans, low-income white/other Americans, persons below age 20, and persons above age 60. A four-stage sample design is being used. The primary sampling units (PSUs), often referred to as stands, are selected from a frame of all U.S. counties. The PSUs are mostly single counties; in a few cases, adjacent counties are combined to keep PSUs above a certain minimum size. There are close to 3,000 PSUs in the NHANES sampling frame. NHANES PSUs are selected with probabilities proportionate to a measure of size (PPS). There are 15 stands in each annual sample.

The second sampling stage is area segments comprising Census blocks or combinations of blocks. Because PSUs

1. Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850. E-mail: LeylaMohadjer@Westat.com; Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, U.S.A. 20782. E-mail: lrc2@cdc.gov.

vary in size, there is some variability in segment size and in the number of segments per PSU. Segments are formed with an average of about 150 households (or dwelling units) per segment. An average of about 5,000 segments are created within each PSU, and an average of 24 segments are sampled. The sample is designed to produce approximately equal sample sizes per PSU, and most PSUs have exactly 24 segments. The segments are also selected with PPS. The measures of size (MOS) of the segments, when combined with the subsampling rates used within the segments, provide approximately equal numbers of sample persons (SPs) per segment, although the relative variation in workload is greater among segments than among PSUs.

The third stage of sample selection consists of households and noninstitutional group quarters, such as dormitories. In a given PSU, following the selection of segments, all dwelling units (DUs) in the sampled segments are listed, and a subsample of households and group quarters within the DUs are designated for screening in order to identify potential SPs for interview and examination. SPs within the households or group quarters are the fourth stage of sample selection. All eligible members within a household are listed, and a subsample of individuals is selected. The subsampling rates for households within segments and for individuals within households are determined in advance. The combination of screening and differential sampling rates provides the increased sample size for those demographic subdomains of special interest (age, sex, race/ethnicity, and income). For example, in the 30 PSUs in which data were collected during the 2-year data cycle 2005-2006, 716 segments were selected and 26,529 households were selected for screening. After being screened for age, sex, and race/ethnicity composition and low-income status, 6,372 households had one or more individuals selected into the sample. A total of 12,862 individuals were selected, of whom 9,950 completed interviews and examinations.

The NHANES examination requires both highly specialized personnel and laboratory processing of collected specimens. As a result, examination components can be very costly to implement. To limit costs and reduce respondent burden, certain examination components are administered to only a subsample of MEC respondents. A single subsampling algorithm controls the amount of overlap among the various subsamples to allow analyses of correlations between various examinations and laboratory components. The SP's assignments to subsamples are fully determined before the SP arrives at the MEC.

The data collected in NHANES surveys have been extremely important in providing needed information about the health and nutritional status of the U.S. population. As a result, beginning with NHANES 1999, the survey has been implemented as a continuous, ongoing, annual survey (Montaquila, Mohadjer and Khare 1998). It is critical to devote a lot of attention to the development and maintenance of an efficient sample design for such an important and complex survey. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the multiple goals of NHANES. The paper focuses on the sample design used through 2006 (in response to emerging analytical requirements, some aspects of the sample design changed starting in 2007).

Section 2 outlines the major purposes and goals of the survey, followed by an overview of the major factors affecting the design given in Section 3. The unique features of the NHANES sample design are described in Section 4. Finally, Section 5 provides a brief summary of the paper.

2. Major purposes and goals of NHANES

NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The main objectives of NHANES are to (1) estimate the national prevalence of selected diseases and risk factors; (2) estimate national population reference distributions of selected health parameters and environmental contaminants; (3) document and investigate reasons for secular trends in selected diseases and risk factors; (4) contribute to the understanding of disease etiology; (5) investigate the natural history of selected diseases; (6) study the relationship between diet, nutrition, environment, genetics, and health; and (7) explore emerging public health issues.

3. Major factors affecting sample design

As mentioned above, a unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the design needs to produce efficient sample sizes for a large number of subdomains of the general population. Many health and nutritional characteristics differ considerably by age, sex, and race/ethnicity and are also affected by income status. As a result, most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Therefore, the survey is designed to produce efficient sample sizes for a very large number of subdomains of the U.S. population.

In general, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. More specifically, the NHANES design attempts to (1) obtain prespecified self-weighting sample sizes for a set of about 75 predesignated subdomains; (2)

produce sample sizes per PSU that will result in an efficient workload for the interview and examination staff at the MEC; (3) design samples that are likely to achieve high response rates; (4) be as cost effective as possible; (5) produce efficient annual samples; (6) allow for accumulation of samples, especially for rare subdomains or rare diseases over time; and (7) be flexible to allow changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

In the remainder of this section, we provide brief summaries of how each of these seven goals affects the design and implementation of NHANES.

NHANES subdomains - The sample design for NHANES meets a prespecified level of precision for cross-sectional data and comparisons over time for a set of predesignated subdomains. Specifically, 77 sampling domains (in the 2006 sample) are defined by race/ethnicity, sex, age, income, and pregnancy status. The sample includes oversamples of blacks, Mexicans, the very young, adolescents, the elderly, pregnant women, and the low-income population.

When estimates of universe totals for the entire population are considered to be of the greatest importance, then the best available estimate of the total population is used as an MOS in the sample selection process. For NHANES, where the interest is in subdomains of the total population, an alternative MOS is needed to improve the accuracy of the estimates and provide better control of the sample size. Section 4 describes the MOS used for sampling PSUs and segments in NHANES.

The objective of oversampling (using differential probabilities of selection) is to achieve a sample containing proportionately more members of certain population subdomains than there are in the population. The goal is to obtain adequate sample sizes to make inferences for subdomains representing relatively small proportions of the total universe of interest and to do it in such a way as to minimize variances for the budget available for the survey. Different oversampling strategies are used depending on the domains of interest. For example, oversampling of the minority subpopulations is accomplished through stratifying geographic areas by concentration of these minority groups and selecting segments in high-density areas at a higher rate. On the other hand, a large screening sample may be required to oversample persons within specific age categories. The subsection on Cost Ratios below describes why oversampling procedures used in NHANES are different from those commonly used in many area frame sample surveys.

Workload for mobile examination centers (*MECs*) - The MEC consists of four specially designed and equipped trailers and contains all of the medical equipment. Each trailer is approximately 45 feet long and 10 feet wide. Detachable truck tractors drive the trailers from one location to another. MECs travel to survey locations throughout the country. The trailers are set up side by side and connected by enclosed passageways. The area in the MEC is divided into rooms to allow privacy during the examinations and interviews. The examination includes a variety of physical and dental assessments and measurements, laboratory tests, and health interviews.

Because of the logistical issues related to the traveling MECs, the sample size in each sampled location must be derived ahead of time and considered fixed so that field operations can be scheduled in an efficient and manageable way. Also, it is necessary to establish a firm time schedule for each stand so that appointments can be made for examinations. It is not possible to change the time schedule since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

Response rates - Achieving high response rates is a concern for practically every sample survey. With NHANES, this is a particular challenge because of the extensive nature of the interviews and examinations. Remunerations have been used in NHANES as a means of improving response rates. In addition, NHANES has an extensive outreach program that includes contacts with local organizations and individuals to gain cooperation, as well as local media coverage to reach as many SPs as possible. As a sample design issue, one approach that has been proven to favorably affect response rates is selecting larger sample sizes within sampled households. One of the factors thought to be responsible for the increased response rates in multiple-SP households is that each person is given remuneration for his or her time and participation, and it is generally more convenient for household members to come to the MEC at the same time. Table 1 shows the examination response rates for SPs coming from households where only one person was selected compared to the response rates for SPs coming from multiple-SP households. As the table indicates, response rates increase by about 4 to 7 percent depending on the type of household.

NHANES is, therefore, designed to maximize the number of SPs per household. Such an approach is feasible for studies like NHANES, where the sample is composed of a large number of subdomains. That is, the effect of within-household clustering is not a large concern for NHANES because most analyses are done within age-sex-specific subdomains (or some limited groups of subdomains) and there is generally little within-household clustering at the subdomain level. The average number of SPs selected per household (in households where at least one SP was selected) within the defined sampling domains ranges from 1 to 1.24 in the 1999-2006 sample. Combining the domains down to 12 to 15 domains by collapsing over age and/or

race/ethnicity will result in average numbers ranging from 1.01 to 1.37 SPs per household. Therefore, some level of clustering is present to the extent that collapsed domains are used for analysis. Note that the SP sample is basically used for SP-level analysis (*e.g.*, health and nutrition statistics). The clustering of SPs is, of course, higher at the family and household levels. However, household- or family-level variables are used for such analysis (*e.g.*, household dust levels, family income, or insurance). Refer to Curtin and Mohadjer (2008) for a discussion of the impact of clustering, and unequal probabilities of selection of subdomains, on the precision levels of various estimates.

Table 1
Examination response rates by number of SPs in household, by household type, in 1999-2006 NHANES sample

Household type	Number of SPs selected per household		Response rate (%)	
	One SP	Two or more SPs	One SP	Two or more SPs
Black/Mexican	4,892	20,222	76.5	82.3
Other low-income[1]	1,362	3,349	77.6	84.5
Other non-low-income	5,597	15,508	68.8	72.6

[1] The Other group includes all SPs who are not Black or Mexican. The low-income threshold is set at 130 percent of poverty.

Cost ratios - The field data collection cost in area survey samples includes the cost of listing DUs, screening households to locate eligible respondents, and conducting the interview to collect data. In NHANES, the data collection phase includes both the household interview and the MEC examination. NHANES requires highly specialized medical equipment, personnel, and laboratory processing. As a result, the cost of an examination is very high compared to other costs in the survey. In fact, the cost of listing and screening is only about 3 to 4 percent of the cost of interviewing and examination. This cost ratio (the cost of interviewing and examination relative to the cost of listing and screening) greatly affects the design of NHANES.

As mentioned above, many of the predesignated subdomains of NHANES require some method of oversampling to achieve the required sample sizes. For the minority populations, substantial reductions in screening are possible with oversampling of highly concentrated minority areas. In general, an optimum design is developed by ascertaining the effect on cost and variance of alternative sampling procedures and choosing the one that minimizes the variance for a fixed cost. In the evaluation of trade-offs between cost and variance, suppose that a particular oversampling strategy reduces the number of households to be listed and screened while increasing the variance for most statistics. The savings brought about by the reduction in cost of listing and screening could be used to increase the size of the sample and thereby lower the variance. However, in NHANES, listing and screening a household is only a very small fraction of the cost, and thus, it takes very large savings in listing and screening costs to justify a moderate increase in variance. As a result, the oversampling procedures established for the survey reflect the NHANES cost ratio and are different from those of typical area surveys.

Annual and multiyear samples - To facilitate potential linkage with other large-scale surveys, to retain flexibility in the sample design, and to allow for the production of annual estimates for broad subdomains, NHANES became a continuous, annual survey starting in 1999. The travel requirements for nationally representative annual samples in the United States are challenging. Three MECs – two of which are stationed at PSUs and one of which is traveling at any given time – work on a very carefully designed schedule to meet the design requirements of the study.

The ability to make meaningful inferences from any survey is affected by both the precision of the estimates themselves and the precision of the variances of the estimates used in the analysis. One of the main limitations of an NHANES annual sample is the small number of PSUs (15 per year), which results in a small number of degrees of freedom for both estimation and analysis and thus design-based variance estimates that are relatively imprecise. Additionally, the effective sample sizes for most subdomains are too small in annual samples. Most subdomain analyses will need to accumulate a number of annual samples to provide both precision and statistical power for comparisons. The procedures for combining years of the survey must be relatively simple, and appropriate for commercial software packages, to maximize the usefulness to the wide variety of users of the NHANES data. Thus, it is critical to employ a sample design that allows efficient accumulation of the annual samples across years.

Flexible design - A critical objective of NHANES is to explore emerging public health issues. The survey needs to be flexible and able to adapt to changing requirements and new challenges. Thus, the sample design must balance the need for efficient subdomain samples with the flexibility needed to make changes in key parameters. To date, the current NHANES design has been able to incorporate some changes in subdomain definitions and sampling rates when these changes have been made after the selection of PSUs. However, in extreme circumstances, substantial changes in subdomain definitions or sample size requirements would necessitate the selection of a new PSU sample.

4. Unique features of the NHANES design

The factors described in Section 3 have played major roles in the development of the sample design and have resulted in some design features that are unique to NHANES. The unique features of the sample design include (1) weighted PSU and segment MOS; (2) efficient annual and multiyear samples; (3) maximized number of SPs per household; (4) controlled sample sizes for PSUs; (5) sequential release of the PSU sample; (6) special methods to deal with deterioration of the efficiency of the optimum design over time; and (7) special methods to reduce the risk of data disclosure through geographic identification.

The following paragraphs briefly describe these unique features of the NHANES design.

Clustering and measures of size (*MOS*) - In NHANES, the sample size must be large enough to produce an efficient workload for each PSU, considering the time and the cost involved in moving a MEC between survey locations and the time required to set up and break down the MECs for travel. Experience gained in earlier NHANES surveys has indicated that an average of 340 examined SPs is an approximately optimum number that provides the maximum number of PSUs while keeping the sample size in each area large enough to justify the costs associated with moving the MECs. In addition, the PSUs for NHANES are typically defined as individual counties to reduce the amount of travel necessary for respondents to visit a MEC, and thereby increase the likelihood of achieving high response rates.

The NHANES sample is designed to yield a self-weighting sample for each sampling subdomain while producing an efficient workload in each PSU. PSUs and segments are selected with probabilities proportionate to a weighted MOS, reflecting the PSU population in subdomains of interest. The selection probability of a PSU determines the maximum rate at which persons residing in that particular PSU can be selected. Refer to *Vital and Health Statistics, Series 2*, *No. 113*, *September 1992*, *CDC/NCHS*, available at http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/120-101/120-101.htm, for a description of the MOS used in NHANES.

Annual and multiyear samples and stratification - One way to achieve nationally representative annual samples is to select an independent sample of PSUs each year. Because of the limited number of NHANES PSUs and the fact that PSUs are selected proportionate to size, this approach would be likely to lead to substantial overlap in PSUs from year to year. Sample overlap, even at the PSU level, could lead to loss of precision in survey estimates when survey years are combined (due to increased clustering of the sample). Thus, rather than sampling PSUs independently each year, the approach in NHANES has been to select a 6-year sample, from a nested structure of major and minor strata (as described below), and then allocate one PSU from each major stratum to each year. This nested structure for the 6-year sample avoids overlap of non-self-representing PSUs during the 6 years.

The design for the NHANES 6-year sample is a stratified two-PSU-per-stratum design and has been developed with the primary goal of efficiency for the 6-year sample, as well as efficient multiyear samples. The stratification scheme is designed to ensure that the PSUs comprising the annual and multiyear samples are distributed evenly in terms of geography and certain population characteristics.

The NHANES design (through 2006) included 18 self-representing PSUs. These PSUs ranged from those that were self-representing for the annual samples to those that were self-representing for 3-year or 6-year samples. These PSUs were assigned such that each year had an equal number of self-representing PSUs, with 3-year self-representing PSUs being 3 years apart. The non-self-representing PSUs were stratified into 12 major strata, defined based on geography and the metropolitan statistical area status of the PSUs. Seventy-two minor strata were defined based on the demographics of the PSUs. The minor strata were constructed to be of equal size to the extent possible (in terms of total MOS). The variables used to form the boundaries of the minor strata were minority status and the percentages of the population below poverty level. Each major stratum included six minor strata, and one PSU was selected from each of these final strata. Within each major stratum, minor strata were paired to create pseudo-strata. Each pair was randomly assigned to the study 3 years apart. The assignment of the pairs to the particular sets of study years and the assignment of the study years within the pair were random within the first major stratum, and all other major strata followed the same pattern.

This stratification scheme resulted in a sample of 72 non-self-representing PSUs that produces efficient annual and multiyear estimates without compromising the efficiency of the 6-year estimates. The 6-year sample has a one-PSU-per-minor-stratum design (or a two-PSU-per-pseudo-stratum design), and each annual sample has a one-PSU-per-major-stratum design. In addition, this design allows for the flexibility needed to address changes in the sample requirements (if a new sample needs to be selected), since the first 3 years of the sample follow a one-PSU-per pseudo-stratum design.

Maximized number of SPs per household - After the sample of screened households is identified, a sample of persons to be interviewed and examined from individual households is selected. All eligible members within a household are listed, and a subsample of individuals is selected based on sex, age, race/ethnicity, and income (all

pregnant women are selected with certainty). SPs are selected at rates established to ensure that the target sample sizes by subdomains will be achieved.

The sample of SPs is selected in a way that maximizes the average number of SPs per household in order to increase the overall response rate in the survey. If independent random selections are made for the subdomains, in most cases only one person in a household would be selected and the average sample size per household would be quite low, not much above 1. Therefore, instead of unrestricted randomization, a pseudo-random procedure is used that maximizes the number of SPs per households. Refer to Waksberg and Mohadjer (1991) for a description of the approach.

Controlled sample sizes per PSU - The sample size in each PSU (stand) that is actually generated from a self-weighting sample in each domain is based on a number of assumptions such as the age and race/ethnicity distribution in the PSU. These assumptions hold only approximately. Once the sample sizes have been calculated, they are treated as quotas, and the number of SPs in each stand is forced to adhere closely to the quota. The reason for this procedure is to have a manageable and efficient field operation. It is necessary to establish a firm, and fixed, time schedule for each stand so that appointments can be made for SP examinations. The time schedule obviously takes into account the expected number of SPs in each stand. As mentioned above, it is difficult to change the time schedule for a stand since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

There is no way of knowing in advance whether the assigned quota for a particular stand is lower or higher than what would arise from self-weighting samples within the various domains. Part of the reason for the uncertainty is that the MOS used for sample selection is based on the latest decennial Census and may not be quite up to date. The issue is further complicated by variations in response rates from stand to stand, as well as sampling variation in the number of identified SPs. Consequently, it is necessary to use a sample selection procedure that can produce samples that are either somewhat larger or somewhat smaller than those arising from the application of the self-weighting sampling rates.

Sequential release of the sample in each stand - To accomplish the above objective, an initial sample is selected in each stand that uses sampling rates 50 percent larger than those required to attain the target sample sizes in each domain. Each stand's initial sample is then divided into a group of subsamples. Each subsample is a systematic subsample of the initial sample, with the households sequenced by segment number and a temporary, geographically based sequence number prior to subsampling. Thus, each subsample cuts across all segments, except when limited by sample size.

As a general rule, the 50 percent subsample (*i.e.*, subsample A) is released to the interviewers first. The yield from this subsample is monitored and used to project estimates of the total number of SPs expected when screening of this subsample has been completed. Based on these figures, additional subsamples are released as needed. The sample is monitored on a daily basis to determine whether additional subsample releases are required.

The one operational problem with the procedure for monitoring the sample yield is that it cannot completely control the subdomain sample sizes. The distribution of subdomains differs, to some extent, from the expected numbers based on the most recent Census data (used to derive the sampling rates). Experience with NHANES indicates that some population changes that will affect the sample sizes can be expected. Other factors that affect subdomain sample yield are patterns of nonresponse and undercoverage in stands. One option to correct the shortfall (or overage) in subdomain sample sizes is to change the sampling rates in future stands. However, such changes will increase heterogeneity in sample weights, thus adversely affecting the precision of the subdomain estimates, and are not advisable except under extreme circumstances.

Dealing with deterioration of the efficiency of the optimum design over time in a tightly controlled sample - The usual practice in area samples is to list all households in the sample segments and apply a prespecified sampling rate to the listed households. This approach gives all households the desired probabilities of selection. For example, if the sampling rate is 50 percent, then one-half of the housing units listed in the segments will be included in the sample. If the number of housing units has tripled due to new construction (*i.e.*, housing units built since the most recent decennial Census), the same sampling rate will produce three times as many interviews and examinations as the number originally expected. Such dramatic changes in the segment size are expected when the data collection period is several years after the most recent decennial Census for which data files are available.

For NHANES, highly variable sample sizes are not feasible because of the scheduling requirements of the MECs. Subsampling within PSUs, in an effort to obtain equal sample sizes across PSUs, is not recommended either, because it will introduce unequal weighting factors that would reduce the efficiency of the sample.

NHANES has used two procedures to update the segment MOS: (1) creation of new construction segments and (2) two-phase sampling to update the MOS. A third approach under consideration involves using purchased commercial address listings to update the MOS in a two-phase sample design.

Under the new construction approach (Bell, Mohadjer, Montaquila and Rizzo 1999), newly constructed units are excluded from area segments and new segments are created based on U.S. Census Bureau information on permits issued for new construction since the most recent decennial Census. New construction segments comprise clusters of building permits issued during one or several adjoining months by a building permit office. Census Bureau files from the Building Permits Survey are used as sources of the data on the number of residential building permits issued by the building permit offices.

Two-phase sampling is used in a number of statistical applications. One of the applications of two-phase sampling is to update a sampling frame when the sample is to be selected with respect to an MOS but a reliable estimate of the MOS is not available. With this approach, a larger sample of units (segments, in the case of NHANES) is selected. An updated value of MOS is then collected for this larger sample (also referred to as the first-phase sample). The final sample of units (segments) is selected from the first-phase sample using the updated MOS.

Starting in 2000, the NHANES segment MOS has been updated (for stands for which such updating seemed necessary) using a two-phase sampling procedure (Montaquila, Bell, Mohadjer and Rizzo 1999). In these cases, listers travel to the stand to obtain a count of the number of DUs in each segment in the first-phase sample. Using the listers' counts, an updated MOS that reflects the ratio of the actual number of DUs to the expected number of DUs is calculated for each first-phase segment. The final sample of segments is then selected by subsampling from the first-phase segments using the updated MOS.

Risk of data disclosure through geographic identification - In today's world, confidentiality concerns and the risk of data disclosure present real challenges to survey sponsors. The ability to identify survey respondents, either through unique combinations available on a single data file or by linking different databases, is of great concern. This is particularly true for NHANES, because of the extensive amount of sensitive data collected on each SP and the small number of PSUs in the sample. Therefore, NHANES evaluates the risk of disclosure on two fronts: geographic disclosure and disclosure from individual characteristics. Various methods (limited or suppressed data release) are used by NCHS to mask the individual characteristics that have a high risk of identifying individuals in the NHANES sample. Sensitive, limited, or non-released data items are available through a Research Data Center. At this time, only national estimates can be produced from publicly available data files; detailed geographic analyses must be done in the Research Data Center.

Although only national estimates can be produced, the direct estimation of sampling errors for those national estimates requires the release of design variables such as stratum and PSU identifiers. Typically, these variables indicate that a group of SPs are all in the same county but do not identify that county. Geographic disclosure is of a particular concern because (1) NHANES has a small number of PSUs, (2) PSUs are limited in geography to one county, and (3) an extensive amount of outreach activity is conducted within each PSU to improve response rates. The outreach program includes contacting various organizations and individuals at each stand to seek their support and using media (newspapers, television, and radio) to reach as many SPs as possible. It is therefore relatively easy to determine the counties in the NHANES sample. The racial/ethnic composition of a county, along with metropolitan/non-metropolitan status, is enough information to correctly match a list of known counties with groups identified as a county cluster on the public data file. To limit geographic disclosure, probabilistic record swapping methods are used at the second stage of sampling (segment swapping) to create masked variance units. The goal is to reduce the risk of identifying individuals by masking their location. Refer to Park, Dohrmann, Montaquila, Mohadjer and Curtin (2006) for a description of the swapping procedures applied to the NHANES sample.

5. Summary and conclusions

A unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the survey is designed to produce efficient sample sizes for a large number of subdomains of the U.S. population, since most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Thus, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. In addition, the design must be as cost effective as possible, produce efficient annual samples, and allow for accumulation of samples for rare subdomains or rare diseases over time. Furthermore, the design must be flexible to allow for changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

The above requirements result in a very complex design with some design features that are unique to NHANES. In particular, the current sample is designed to produce efficient annual and multiyear samples. NHANES uses weighted PSU and segment MOS to yield self-weighting

samples for each subdomain, while producing an efficient workload in each PSU. Once the sample sizes are calculated, they are treated as quotas. The sample sizes are strictly controlled in each PSU to create a manageable and efficient field operation. A very large screening sample is used to oversample most of the age and income subdomains, and oversampling of highly concentrated areas is used for some of the very rare minority subdomains. The sample of SPs is selected using a pseudo-random procedure to maximize the average number of SPs per household because it has appeared to increase the overall response rate in previous surveys.

The challenges described in this paper are focused on the main aspects of the NHANES. There remain many other features unique to NHANES that analysts must take into account when analyzing data from the survey. For example, not only are there very few PSUs in each annual sample, but data collected within these PSUs are not randomly collected across the seasons. In particular, if there is a seasonal by geographic region interaction for a variable of interest, the current NHANES design will not be able to estimate it. Because of the small number of PSUs in each data release cycle, any contextual data linkage at the geographic level must be done in the NCHS Research Data Center. Because of the many subsamples within NHANES, special care must be taken to use the appropriate subsample weight; for example, estimates for undiagnosed diabetes must use the special fasting weight.

To facilitate the efficient use of MECs for data collection, there has been no attempt to randomly allocate the sample of PSUs across time in annual samples. However, the time dimension plays a major role in some health indicators, such as nutrition. Furthermore, analysis of nutrition data may also be affected by the complex nature of the design and data collection. Special sample weights constructed for the 2 days of the 24-hour recall data account for variation in the number of examinations by day of the week. A web-based tutorial is now being developed to provide assistance in the analysis of NHANES nutrition data. A general tutorial for design-based analysis of NHANES data can be found at http://www.cdc.gov/nchs/tutorials/.

Acknowledgements

The authors are grateful to the associate editor and the referees for their helpful comments and suggestions, which have greatly improved the paper.

References

Bell, B., Mohadjer, L., Montaquila, J. and Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Curtin, L.R., and Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, to appear.

Montaquila, J., Bell, B., Mohadjer, L. and Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Montaquila, J., Mohadjer, L. and Khare, M. (1998). The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. and Curtin, L.R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Waksberg, J., and Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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 23, No. 4, 2007

Contents

Volume 24, No. 1, 2008

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 35, No. 4, December/décembre 2007