Catalogue no. 12-001-XIE

Survey Methodology

June 2007

Statistics Canada
Statistique Canada

Canada

How to obtain more information

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

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

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

Accessing and ordering information

This product, catalogue no. 12-001-XIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

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

The following additional shipping charges apply for delivery outside Canada:

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

All prices exclude sales taxes.

The printed version of this publication can be ordered by

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

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

Standards of service to the public

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

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2007

Published by authority of the Minister responsible for Statistics Canada

June 2007

Catalogue no. 12-001-XIE, Vol. 33 no. 1
ISSN 1492-0921

Catalogue no. 12-001-XPB, Vol. 33 no. 1
ISSN: 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF 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

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

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 33, Number 1, June 2007

Contents

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.
∞

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 1-2
Statistics Canada, Catalogue No. 12-001

In This Issue

This issue of *Survey Methodology* includes papers covering a variety of methodological subjects such as modeling and estimation, weighting and variance estimation, non-response and sampling.

In the first paper of the issue, Skinner and Vieira investigate the effect of clustered sampling on variance estimation in longitudinal surveys. They present theoretical arguments and empirical evidence of the effects of ignoring clustering in longitudinal analyses, and find that these effects tend to be larger than for corresponding cross-sectional analyses. They also compare traditional survey sampling based methods to account for clustering in variance estimation to a multi-level modeling approach.

Kovačević and Roberts compare three models for analyzing multiple spells arising from data collected through longitudinal surveys with complex survey designs, which can involve stratification and clustering. These models are variations of the Cox proportional hazards model along the same lines as those proposed in the literature by Lin and Wei (1989), Binder (1992) and Lin (2000). These three models are compared using data from Statistics Canada's Survey on Labor and Income Dynamics (SLID). This paper gives new insight into fitting Cox models to survey data containing multiple spells per individual, a situation that arises quite frequently. The paper also illustrates some of the challenges in fitting Cox models to survey data.

Elliott, in his paper, presents a method for balancing elevated variance due to extreme weights with potential bias using a Bayesian weight trimming method in generalized linear models. This is accomplished by using a stratified hierarchical Bayesian model in which strata are determined by the probabilities of inclusion or survey weights. He illustrates and evaluates the approach using simulations based on linear and logistic regression models, and an application using data from the Partners for Child Passenger Safety dataset.

The paper by Breidt, Opsomer, Johnson and Ranalli explores the use of semiparametric methods for the estimation of population means. In semiparametric estimation, some variables are assumed to be linearly related to the variable of interest while the other variables may have a complicated, unspecified relation to the variable of interest. The authors study theoretically the properties under the sampling design of the resulting estimators. In particular, they show the design-consistency and the asymptotic normality of their estimator. Their method is then applied to data from a survey of lakes in the northeastern United States.

Tanguay and Lavallée address the problem of estimating the depreciation of assets based on a database of price ratios. In their paper, the issue is that the ratios do not come from a random sample from the population of ratios. The authors argue that the distribution of ratios should converge to a Uniform distribution and propose a weighting scheme that will make the weighted empirical distribution function approximately uniform. The proposed method is illustrated by an example using data on the depreciation of automobiles.

Steel and Clark present an empirical and theoretical comparison of person-level generalized regression survey weights and integrated household-level weights in the case of a simple random sample of households in which all household members selected. They conclude that there is little or no loss in efficiency associated with integrated weighting.

Saigo, in his paper, proposes a bootstrap variance estimation procedure for two-phase designs with high sampling fractions. The method uses common bootstrap techniques, but adjusts the values of the auxiliary variables for units that are selected in the first phase sample only. The proposed technique is illustrated using several commonly used estimators such as the ratio estimator, and estimators of the distribution function and quantiles. Results from a simulation study comparing the proposed method to several others are presented.

In the paper by Longford the problem of estimating the MSE of small area estimates is investigated. A composite estimator of the MSE of small area means is obtained by combining a model-based variance estimator and a naïve estimator of the MSE. The coefficient that combines the two estimators minimizes the expected MSE of the resulting composite estimator of the MSE. The proposed estimator is compared with existing estimators through several simulation studies.

Shao considers the problem of imputing for missing values when the nonresponse is nonignorable. In the situation where the nonresponse depends on a cluster level random effect, he shows that the commonly used mean imputed estimator is biased unless the mean of the cluster is used. For variance estimation, a jackknife variance estimation procedure for the proposed estimator is provided. The proposed estimator is compared with the mean imputed estimator by means of a simulation study.

In the final paper of this issue, Tiwari, Nigam and Pant make use of the idea of nearest proportional to size sampling designs to obtain optimal controlled sample designs where non-preferred samples have zero selection probabilities. The optimal controlled sampled designs are obtained by combining an initial inclusion probability proportional to size design and quadratic programming techniques to ensure that non-preferred samples have a zero selection probability. Their method is illustrated using several examples.

Harold Mantel, Deputy Editor

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 3-12
Statistics Canada, Catalogue No. 12-001

Variance estimation in the analysis of clustered longitudinal survey data

Chris Skinner and Marcel de Toledo Vieira [1]

Abstract

We investigate the impact of cluster sampling on standard errors in the analysis of longitudinal survey data. We consider a widely used class of regression models for longitudinal data and a standard class of point estimators of a generalized least squares type. We argue theoretically that the impact of ignoring clustering in standard error estimation will tend to increase with the number of waves in the analysis, under some patterns of clustering which are realistic for many social surveys. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses. We illustrate this theoretical argument with empirical evidence from a regression analysis of longitudinal data on gender role attitudes from the British Household Panel Survey. We also compare two approaches to variance estimation in the analysis of longitudinal survey data: a survey sampling approach based upon linearization and a multilevel modelling approach. We conclude that the impact of clustering can be seriously underestimated if it is simply handled by including an additive random effect to represent the clustering in a multilevel model.

Key Words: Clustering; Design effect; Misspecification effect; Multilevel model.

1. Introduction

It is well known that it is important to take account of sample clustering when estimating standard errors in the analysis of survey data. Otherwise, standard error estimators can be severely biased. In this paper we investigate the impact of clustering in the regression analysis of longitudinal survey data and compare it with the impact on corresponding cross-sectional analyses. Kish and Frankel (1974) presented empirical work which suggested that the impact of complex designs on variances decrease for more complex analytical statistics and so one might conjecture that the impact on longitudinal analyses might also be reduced. We shall argue that, in fact, the impact of clustering on longitudinal analyses can tend to be greater, at least for a number of common types of analysis and for some common practical settings. An intuitive explanation is that some common forms of longitudinal analysis of individual survey data 'pool' data over time and enable much temporal 'random' variation in individual responses to be 'extracted' in the estimation of regression coefficients. In contrast, it may only be possible to extract much less variation in the effects of clustering since such clustering, representing geography for example, often tends to generate more stable effects than repeated measurements of individual behaviour. As a consequence the relative importance of clustering in standard errors can increase the more waves of data are included in the analysis.

In addition to considering the impact of clustering on variance estimation, we shall also consider the question of how to undertake the variance estimation itself. It is natural for many analysts to represent clustering via multilevel models (Goldstein 2003, Chapter 9; Renard and Molenberghs 2002) and we shall consider how variance estimation methods based upon such models compare with survey sampling variance estimation procedures in the case of cluster sampling.

There is a well established literature on methods for taking account of complex sampling schemes in the regression analysis of survey data. See *e.g.*, Kish and Frankel (1974), Fuller (1975), Binder (1983), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003). We restrict attention here to 'aggregate' regression analyses (Skinner *et al.* 1989), where regression coefficients at the 'population level' are the parameters of interest, where suitable estimates of these coefficients may be obtained by adapting standard model-based procedures using survey weights and where the variances of these estimated regression coefficients may be estimated by linearization methods (Kish and Frankel 1974; Fuller 1975). In this paper, we extend this work to the case when longitudinal survey observations are obtained, based upon an initial sample drawn according to a complex sampling scheme, focussing again on the case of a clustered design. We consider a standard class of linear regression models for such longitudinal data, as considered in the biostatistical literature (*e.g.*, Diggle, Heagerty, Liang and Zeger 2002), the multilevel modelling literature (*e.g.*, Goldstein 2003) and the econometric literature (*e.g.*, Baltagi 2001). We consider an established class of point estimators of a generalized least squares type, modified by survey weighting. For some applications of such methods to survey data, see Lavange, Koch and Schwartz (2001); Lavange, Stearns, Lafata, Koch and Shah (1996).

1. Chris Skinner, University of Southampton, United Kingdom; Marcel de Toledo Vieira, Universidade Federal de Juiz de Fora, Brazil.

The impact of a complex sampling scheme on variance estimation will be measured by the 'misspecification effect', denoted meff (Skinner 1989a), which is the variance of the point estimator of interest under the actual sampling scheme divided by the expectation of a specified variance estimator. This is a measure of the relative bias of the specified variance estimator. If it is unbiased then the meff will be one. If the actual sampling scheme involves clustering but the specified variance estimator is 'misspecified' by ignoring the clustering, then the expectation of the variance estimator will usually be less than the actual variance and the meff will be greater than one. This concept is closely related to that of the 'design effect' or deff of Kish (1965), defined as the variance of the point estimator under the given design divided by its variance under simple random sampling with the same sample size, a concept more relevant to the choice of design than to the choice of standard error estimator.

We shall illustrate our theoretical arguments with analyses of data from the British Household Panel Survey (BHPS) on attitudes to gender roles, where the units of primary analytic interest are individual women and the clusters consist of postcode sectors, used as primary sampling units in the selection of the first wave sample from an address register.

The framework, including the models and estimation methods, is described in Section 2. The theoretical properties of the variance estimation methods are considered in Section 3. Section 4 illustrates these properties numerically, using an analysis of BHPS data. Some concluding remarks are provided in Section 5.

2. Regression model, data and inference procedures

Consider a finite population $U = \{1, ..., N\}$ of N units, assumed fixed across a series of occasions $t = 1, ..., T$. We shall refer to the units as individuals, although our discussion is applicable more generally. Let y_{it} denote the value of an outcome variable for individual $i \in U$ at occasion t and let $y_i = (y_{i1}, ..., y_{iT})'$ be the vector of repeated measurements. Let x_{it} denote a corresponding $1 \times q$ vector of values of covariates for individual i at occasion t and let $x_i = (x_{i1}', ..., x_{iT}')'$. We assume that the following linear model holds for the expectation of y_i conditional on $(x_1, ..., x_N)$:

$$E(y_i) = x_i\beta, \qquad (1)$$

where β is a $q \times 1$ vector of regression coefficients and the expectation is with respect to the model. We suppose that β is the target for inference, that is the regression coefficients are the parameters of primary interest to the analyst. Although we shall consider further features of this model, such as the covariance matrix of y_i, these will be assumed to be of secondary interest to the analyst.

The data available to make inference about β are from a longitudinal survey in which values of y_{it} and x_{it} are observed at each occasion (wave) $t = 1, ..., T$ for individuals i in a sample, s, drawn from U at wave 1 using a specified sampling scheme. For simplicity, we assume no non-response here, but return to this possibility in Section 4.

In order to formulate a point estimator of β, we extend the specification of (1) to the following 'working' model:

$$y_{it} = x_{it}\beta + u_i + v_{it}, \qquad (2)$$

where u_i and v_{it} are independent random effects with zero means and variances $\sigma_u^2 = \rho\sigma^2$ and $\sigma_v^2 = (1 - \rho)\sigma^2$ respectively, conditional on $(x_1, ..., x_N)$. This model may be called a uniform correlation model (Diggle *et al.* 2002, page 55) or a two-level model (Goldstein 2003). The parameter ρ is the intra-individual correlation.

The basic point estimator of β we consider is

$$\hat{\beta} = \left(\sum_{i \in s} w_i \, x_i' V^{-1} x_i \right)^{-1} \sum_{i \in s} w_i \, x_i' V^{-1} y_i, \qquad (3)$$

where w_i is a survey weight and V is a $T \times T$ estimated covariance matrix of y_i under the working model (2), *i.e.*, it has diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\,\hat{\sigma}^2$, where $(\hat{\rho}, \hat{\sigma}^2)$ is an estimator of (ρ, σ^2). (Note that in fact $\hat{\sigma}^2$ cancels out in (3) and hence σ^2 does not need to be estimated for $\hat{\beta}$). In the absence of the weight terms and survey considerations, the form of $\hat{\beta}$ is motivated by the generalized estimating equations (GEE) approach of Liang and Zeger (1986). The idea here is that $\hat{\beta}$, as a generalized least squares estimator of β, would be fully efficient if the working model (2) held. However, $\hat{\beta}$ remains consistent under (1) and may still be expected to combine within- and between-individual information in a reasonably efficient way even if the working model for the error structure does not hold exactly.

The survey weights are included in (3) following the pseudo-likelihood approach (Skinner 1989b) to ensure that $\hat{\beta}$ is approximately unbiased for β with respect to the model and the design, provided (1) holds.

There are a number of alternative ways of estimating ρ. In a non-survey setting, Liang and Zeger (1986) provide an iterative approach which alternates between estimates of β and ρ. Shah, Barnwell and Bieler (1997) describe how survey weights may be incorporated into this approach and implement this method in the REGRESS procedure of the software SUDAAN. By default, SUDAAN implements only one step of this iterative method and, in the non-survey setting, Lipsitz, Fitzmaurice, Orav and Laird (1994) conclude there is little to be lost by using only a single step.

For the working model in (2), the approach of Liang and Zeger (1986) to the estimation of β and ρ is virtually identical to the iterative generalized least squares (IGLS) estimation approach of Goldstein (1986). Both methods iterate between estimates of β and ρ and both use GLS to estimate β given the current estimate of ρ. The only slight difference is in the method used to estimate ρ. Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) show how to incorporate survey weights into the IGLS approach and their method may be expected to lead to very similar estimates of ρ to those in the SUDAAN REGRESS procedure. For the purposes of this paper, the precise form of $\hat{\rho}$ will not be critical and we may view $\hat{\beta}$ as either a weighted GEE or a weighted IGLS estimator.

We now turn to the estimation of the covariance matrix of $\hat{\beta}$ under the complex sampling scheme. We shall generally assume that a stratified multistage sampling scheme has been employed. We consider two main approaches to variance estimation.

Our first approach is the classical method of linearization (Skinner 1989b, page 78). The estimator of covariance matrix of $\hat{\beta}$ is

$$v(\hat{\beta}) = \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \times \left[\sum_h n_h/(n_h-1)\sum_a (z_{ha}-\bar{z}_h)(z_{ha}-\bar{z}_h)'\right] \times \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \quad (4)$$

where h denotes stratum, a denotes primary sampling unit (PSU), n_h is the number of PSUs in stratum h, $z_{ha} = \sum_i w_i\, x_i' V^{-1} e_i$, $\bar{z}_h = \sum_a z_{ha}/n_h$ and $e_i = y_i - x_i\hat{\beta}$. Similar estimators are considered by Shah *et al.* (1997, pages 8-9) and Lavange *et al.* (2001). If the weights, the sampling scheme and the difference between $n/(n-1)$ and 1 are ignored, this estimator reduces to the 'robust' variance estimator presented by Liang and Zeger (1986).

Our second approach is more directly model-based. The model is first extended to represent the complex population underlying the sampling scheme and inference then takes place with respect to the extended model. We consider only the case of two-stage sampling from a clustered population, where the two-level model in (2) is extended to the three-level model (Goldstein 2003):

$$y_{ait} = x_{ait}\beta + \eta_a + u_{ai} + v_{ait}. \quad (5)$$

The additional subscript a denotes cluster and the additional random term η_a with variance σ_η^2 represents the cluster effect (assumed independent of u_{ai} and v_{ait}). We let σ_u^2 and σ_v^2 denote the variances of u_{ai} and v_{ait} respectively. Inference then takes place using IGLS, which may be weighted to avoid selection bias. This approach generates an estimated covariance matrix of the estimator of β directly. It should be noted, however that the estimator of β derived using weighted IGLS under model (5) may differ slightly from the estimator in (3) (although, for given estimates of the three variance components in (5), it will be the same as a weighted GEE estimator with a working covariance matrix based on this three-level model). Nevertheless, from our experience of social survey applications, such as in Section 4, and from theory (Scott and Holt 1982) the difference between these alternative point estimators will often be negligible.

Two broad approaches to deriving variance estimators from (5) are available. First, ignoring survey weights, the standard IGLS method (Goldstein 1986) may be employed, assuming that each random effect follows a normal distribution. Second, to avoid the assumption of normal homoscedastic random effects, a 'robust' variance estimation method (Goldstein 2003, page 80) may be employed. This approach is extended to handle survey weights in Pfeffermann *et al.* (1998). Leaving aside stratification, their variance estimator is identical to the linearization estimator in (4) for a given value of $\hat{\rho}$.

3. Properties of variance estimators

In this section we consider the properties of the estimators of the covariance matrix of $\hat{\beta}$ described in the previous section. We focus first on the linearization estimator $v(\hat{\beta})$ in (4).

The consistency of $v(\hat{\beta})$ for the covariance matrix of $\hat{\beta}$ follows established arguments in a suitable asymptotic framework (*e.g.*, Fuller 1975; Binder 1983). The one non-standard feature is the presence of V^{-1} in $\hat{\beta}$ and $v(\hat{\beta})$ and the dependence of V on $\hat{\rho}$. In fact, in large samples the covariance matrix of $\hat{\beta}$ depends on $\hat{\rho}$ only via its limiting value ρ^* (in a given asymptotic framework). To see this, write $\hat{\beta} - \beta = (\sum_s u_i)^{-1}\sum_s \tilde{z}_i$, where $u_i = w_i\, x_i' V^{-1} x_i$, $\tilde{z}_i = w_i\, x_i' V^{-1}\tilde{e}_i$ and $\tilde{e}_i = y_i - x_i\beta$. Note that, under weak regularity conditions (Fuller and Battese 1973, Corollary 3), the asymptotic distribution of $\hat{\beta} - \beta$ is the same as that of $\beta^* - \beta = (\sum_s u_i^*)^{-1}\sum_s z_i^*$, where $u_i^* = w_i\, x_i' V^{*-1} x_i$, $z_i^* = w_i\, x_i' V^{*-1}\tilde{e}_i$ and V^* takes the same form as V with $\hat{\rho}$ replaced by $\rho^* = p\lim(\hat{\rho})$, the probability limit of $\hat{\rho}$ in the asymptotic framework. Writing $\bar{z}^* = \sum_s z_i^*/n$ and $\bar{U} = p\lim(\sum_s u_i^*/n)$, we may thus approximate the covariance matrix of $\hat{\beta}$ asymptotically by $\mathrm{var}(\hat{\beta}) \approx \bar{U}^{-1}\mathrm{var}(\bar{z}^*)\bar{U}^{-1}$. If the working model (2) holds then $\rho^* = \rho$ and this covariance matrix will be the same for any consistent method of estimating ρ. Even if the working model does not hold, $v(\hat{\beta})$ will be consistent for $\bar{U}^{-1}\mathrm{var}(\bar{z}^*)\bar{U}^{-1}$ within the kinds of asymptotic frameworks considered by

Fuller (1975) and Binder (1983) and under the kinds of regularity conditions they and Fuller and Battese (1973) set out.

We next explore the impact on the linearization method of ignoring a complex sampling design. We denote by $v_0(\hat{\beta})$ the linearization estimator obtained from expression (4) by ignoring the design, *i.e.*, by assuming only a single stratum with PSUs identical to individuals so that $n_h = n$ is the overall sample size and z_{ha} is replaced by $z_i = w_i x_i' V^{-1} e_i$. We shall be concerned with the bias of $v_0(\hat{\beta})$ when in fact the design is complex. Let $\hat{\beta}_k$ denote the k^{th} element of $\hat{\beta}$ and let $v_0(\hat{\beta}_k)$ denote the k^{th} element of $v_0(\hat{\beta})$. Then, following Skinner (1989a, page 24), we shall measure the relative bias of the 'incorrectly specified' variance estimator $v_0(\hat{\beta}_k)$ as an estimator of $\text{var}(\hat{\beta}_k)$ by the *misspecification effect*, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \text{var}(\hat{\beta}_k)/E[v_0(\hat{\beta}_k)]$. Since $v(\hat{\beta}_k)$ is a consistent estimator of $\text{var}(\hat{\beta}_k)$, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be estimated by $v(\hat{\beta}_k)/v_0(\hat{\beta}_k)$ and is closely related to the idea of design effect.

To investigate the nature of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$, we first write:

$$v_0(\hat{\beta}) = \left(\sum_s u_i\right)^{-1} [n/(n-1)] \times \left[\sum_s (z_i - \bar{z})(z_i - \bar{z})'\right]\left(\sum_s u_i\right)^{-1}, \qquad (6)$$

where $\bar{z} = \sum_s z_i / n$. Then, as an asymptotic approximation, we have $E[v_0(\hat{\beta})] \approx \bar{U}^{-1}[n^{-1}S_z^*]\bar{U}^{-1}$, where S_z^* is the probability limit of the finite population covariance matrix of z_i^*. Using the fact that the numerator of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be approximated by $\bar{U}^{-1}\,\text{var}(\bar{z}^*)\bar{U}^{-1}$, we can thus write:

$$\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \frac{(\bar{U}^{-1})_k \,\text{var}(\bar{z}^*)\,(\bar{U}^{-1})_k'}{(\bar{U}^{-1})_k [n^{-1}S_z^*](\bar{U}^{-1})_k'}, \qquad (7)$$

where $(\bar{U}^{-1})_k$ is the k^{th} row of $\bar{U}^{-1}$. This simplifies in the case $q = 1$ to:

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = \text{var}(\bar{z}^*)/[n^{-1}S_z^*]. \qquad (8)$$

We may explore more specific forms of these expressions under different models and assumptions about the weights and the sampling scheme. We focus here on the impact of clustering, assuming equal weights and no stratification. Consider the three-level model in (5) and, to simplify matters, suppose that $q = 1$ and $x_{ait} \equiv 1$ and β is the mean of y_{ait}. Then, straightforward algebra shows that the value of z_i^* for individual i within cluster a is $[1 + \rho^*(T-1)]^{-1}\sum_t(\eta_a + u_{ai} + v_{ait})$. Now suppose that two-stage sampling is employed with a common sample size m per cluster. Then, evaluating the variance $\text{var}(\bar{z}^*)$ and probability limit S_z^* in (8) with respect to the model in (5), we find, in a similar manner to Skinner (1989a, page 38):

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = 1 + (m-1)\tau, \qquad (9)$$

where $\tau = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2/T)$ is the intracluster correlation of z_i^*. We see that, under this model, the meff increases as T increases (provided $\sigma_v^2 > 0$) and thus the impact of clustering on variance estimation is greater in the longitudinal case than for the cross-sectional problem (where $T = 1$).

This finding depends on the rather strong assumption that the cluster effects η_a are constant over time. In fact, (9) still holds if we replace η_a by a time-varying effect η_{at} provided we replace τ by $\tau = \text{var}(\bar{\eta}_a)/[\text{var}(\bar{\eta}_a) + \sigma_u^2 + \sigma_v^2/T]$, where $\bar{\eta}_a = \sum_t \eta_{at}/T$. Now, the meff will increase as T increases if (and only if) $\sigma_u^2 + \sigma_v^2/T$ decreases faster with T than $\text{var}(\bar{\eta}_a)$. Whether this is the case will depend on the particular application. However, we suggest that for many longitudinal surveys of individuals with area-based clusters (the kind of setting we have in mind), this condition is plausible. In such applications we may often expect σ_v^2 to be large relative to σ_u^2 (*i.e.*, for the cross-sectional intracluster correlation to be small) in particular as a result of wave-specific measurement error and thus for $\sigma_u^2 + \sigma_v^2/T$ to decrease fairly rapidly as T increases. The socio-economic characteristics of areas may often be expected to be more stable and only in unusual situations might we expect measurement error to lead to much occasion-specific variance in η_{at}. Thus, we suggest that the ratio of $\text{var}(\bar{\eta}_a)$ for $T = 5$, say, compared to $T = 1$ may in such applications usually be expected to be greater than $(\sigma_u^2 + \sigma_v^2/5)/(\sigma_u^2 + \sigma_v^2)$ which will approach 1/5 as σ_u^2/σ_v^2 approaches 0. We thus suggest that in many practical circumstances it will be more important to allow for clustering in longitudinal analyses than in corresponding cross-sectional analyses. An empirical illustration is provided in Section 4.

We now consider the properties of variance estimators based upon the three-level model in (5). We consider only the approach based upon the assumption of normally distributed homoscedastic random effects, ignoring survey weights, given the (virtual) equivalence of the 'robust' multilevel approach and linearization.

If model (5) is correct and we can indeed ignore survey weights then the model-based variance estimator will be consistent (Goldstein 1986). However, as discussed in Skinner (1989b, page 68) and supported by theory in Skinner (1986), the main feature of clustering likely to impact on the standard errors of estimated regression coefficients is the variation in regression coefficients between clusters. This is not allowed for in model (5).

To see how model (5) may fail to capture the effects of clustering adequately, consider the cross-sectional case ($T = 1$) where x is scalar. Then, if the three-level model (5) holds, an approximate expression for the meff of the variance estimator of $\hat{\beta}$ based upon the two-level model (2) is:

$$\text{meff} = 1 + (m-1)\tau_1\tau_x, \qquad (10)$$

where $\tau_1 = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$ and τ_x is the intracluster correlations for x (Scott and Holt 1982; Skinner 1989b, page 68). This result extends in the longitudinal case, to:

$$1 \leq \text{meff} \leq 1 + (m-1)\tilde{\tau}\tau_z, \qquad (11)$$

where $\tilde{\tau}$ is the long-run ($T = \infty$) version of τ (see Appendix) and τ_z is an intracluster correlation coefficient for $z_{ai} = \sum_t x_{ait}/T$. The proof of this result and the simplifying assumptions required are sketched in the Appendix. The main point is that both $\tilde{\tau}$ and τ_z will often be small in which case $\tilde{\tau}\tau_z$ will be very small and thus meff may be implausibly close to one with the model-based variance estimator being subject to downward bias. We explore this empirically in Section 4. Of course, random coefficients could be introduced into model (5) and we consider this also in Section 4. However, given the difficulty of specifying a correct random coefficient model, this approach does not seem likely to be very robust.

Our focus in this section has so far been on the potential bias (or inconsistency) of variance estimation methods. It is also desirable to consider their efficiency. In particular, the linearization method may be expected to be less efficient than model-based variance estimation if the model is correct. The relative importance of efficiency vs. bias may be expected to increase as the number of clusters decreases. Wolter (1985, Chapter 8) summarises a number of simulation studies investigating both the bias and variance of the linearization variance estimator and these studies suggest that the linearization method performs well even with few clusters. Possible degrees of freedom corrections to confidence intervals for regression coefficients based upon the linearization method with small numbers of clusters are discussed by Fuller (1984). A simulation study of estimators for multilevel models in Maas and Hox (2004) does not suggest that the linearization method performs noticeably worse than the model-based approach, in terms of the coverage of confidence intervals for coefficients in β, even with as few as 30 clusters.

4. Example: Regression analysis of BHPS data on attitudes to gender roles

We now present an application to BHPS data to illustrate some of the theoretical properties discussed in the previous section.

Recent decades have witnessed major changes in the roles of men and women in the family in many countries. Social scientists are interested in the relation between changing attitudes to gender roles and changes in behaviour, such as parenthood and labour force participation (*e.g.*, Morgan and Waite 1987; Fan and Marini 2000). A variety of forms of statistical analysis are used to provide evidence about these relationships. Here, we consider estimating a linear model of form (1), with a measure of attitude to gender roles as the outcome variable, y, following an analysis of Berrington (2002).

The data come from waves 1, 3, 5, 7 and 9 (collected in 1991, 1993, 1995, 1997, and 1999 respectively) of the BHPS and these waves are coded $t = 1, ..., T = 5$ respectively. Respondents were asked whether they 'strongly agreed', 'agreed', 'neither agreed nor disagreed', 'disagreed' or 'strongly disagreed' with a series of statements concerning the family, women's roles, and work out of the household. Responses were scored from 1 to 5. Factor analysis was used to assess which statements could be combined into a gender role attitude measure. The attitude score, y_{it}, considered here is the total score for six selected statements for woman i at wave t. Higher scores signify more egalitarian gender role attitudes. Berrington (2002) provides further discussion of this variable. A more sophisticated analysis might include a measurement error model for attitudes (*e.g.*, Fan and Marini 2000), with each of the five-point responses to the six statements treated as ordinal variables. Here, we adopt a simpler approach, treating the aggregate score y_{it} and the associated coefficient vector β as scientifically interesting, with the measurement error included in the error term of the model.

Covariates for the regression analysis were selected on the basis of discussion in Berrington (2002) but reduced in number to facilitate a focus on the methodological issues of interest. The covariate of primary scientific interest is economic activity, which distinguishes in particular between women who are at home looking after children (denoted 'family care') and women following other forms of activity in relation to the labour market. Variables reflecting age and education are also included since these have often been found to be strongly related to gender role attitudes (*e.g.*, Fan and Marini 2000). All these covariates may change values between waves. A year variable (scored 1, 3, …, 9) is also included. This may reflect both historical change and the general ageing of the women in the sample.

The BHPS is a household panel survey of individuals in private domiciles in Great Britain (Taylor, Brice, Buck and Prentice-Lane 2001). The initial (wave one) sample in 1991 was selected by a stratified multistage design in which households had approximately equal probabilities of inclusion. The households were clustered into 250 primary

sampling units (PSUs), consisting of postcode sectors. All resident members aged 16 or over were selected in sample households. All adults selected at wave one were followed from wave two onwards and represent the longitudinal sample. The survey is subject to attrition and other forms of wave non-response. To handle this non-response, we have simply replaced s in (3) by the 'longitudinal sample' of individuals for which observations are available for each of $t = 1, ..., T$ and have chosen not to apply any survey weighting since our aim is to study potential misspecification effects associated with clustering and we wish to avoid confounding these with weighting effects. We also ignore the impact of stratification in the numerical work in this section (but see Section 5 for some comments on the effect of weights and stratification).

Given the analytic interest in whether women's primary labour market activity is 'caring for a family', we define our study population as women aged 16-39 in 1991. Thus our data consist of the longitudinal sample of women in the eligible age range for whom full interview outcomes (complete records) were obtained in all five waves, a sample of $n = 1{,}340$ women. These women are spread fairly evenly across 248 postcode sectors. The small average sample size of around five per postcode sector combined with the relatively low intra-postcode sector correlation for the attitude variable of interest leads to relatively small impacts of the design, as measured by meffs. Since our aims are methodological ones, we have chosen to group the postcode sectors into 47 geographically contiguous clusters, to create sharper comparisons, less blurred by sampling errors which can be appreciable in variance estimation. The meffs in the tables we present therefore tend to be greater than they are for the actual design. The latter results tend to follow similar patterns, although the patterns are less clear-cut as a result of sampling error.

We first estimate meffs for the linearization estimator, as discussed at the beginning of Section 3. Using data from just the first wave and setting $x_{ait} \equiv 1$, the estimated meff for this cross-sectional mean is given in Table 1 as about 1.5. This value is plausible since, if we make the usual approximation of (9) for unequal sample cluster sizes by replacing m by $\bar{m}$, the average sample size per cluster, we find that $1 + (\bar{m} - 1)\tau = 1.5$ and $\bar{m} = 1{,}340/47 \approx 29$ imply a value of τ of about 0.02 and such a small value is in line with other estimated values of τ found for attitudinal variables in British surveys (Lynn and Lievesley 1991, Appendix D).

Table 1 Estimates for longitudinal means

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3,5	1-7	1-9
	19.83	0.12	1.51	1.50	1.68	1.81	1.84

To assess the impact of the longitudinal aspect of the data, we estimated a series of meffs using data for waves $1, \ldots, t$ for $t = 2, 3, \ldots, 5$. Although these estimated meffs are subject to sampling error, there seems clear evidence in Table 1 of a tendency for the meff to increase with the number of waves. This trend might be anticipated from the theoretical discussion in Section 3 if the average level of egalitarian attitudes in an area varies less from year to year than the attitude scores of individual women. This seems plausible since the latter will be affected both by measurement error and genuine changes in attitudes, so that $\text{var}(\bar{\eta}_a)$ may be expected to decline more slowly with T than $\text{var}(\bar{u}_a + \bar{v}_a)$. We may therefore expect τ, and consequently the meff, to increase as T increases, as we observe in Table 1.

We next elaborate the analysis by including indicator variables for economic activity as covariates. The resulting regression model has an intercept term and four covariates representing contrasts between women who are employed full-time and women in other categories of economic activity. The estimated meffs are presented in Table 2. The intercept term is a domain mean and standard theory for a meff of a mean in a domain cutting across clusters (Skinner 1989b, page 60) suggests that it will be somewhat less than the meff for the mean in the whole sample, as indeed is observed with the meff for the cross-section domain mean of 1.13 in Table 2 being less than the value 1.51 in Table 1. As before, there is some evidence in Table 2 of tendency for the meff to increase, from 1.13 with one wave to 1.50 with five waves, albeit with lower values of the meffs than in Table 1. The meffs for the contrasts in Table 2 vary in size, some greater than and some less than one. These meffs may be viewed as a combination of the traditional variance inflating effect of clustering in surveys together with the variance reducing effect of blocking in an experiment. Such variance reduction arises if the domains being contrasted share a common cluster effect (of the form η_a in model (5)) which tends to cancel out in the contrasts, implying that the actual variance of the contrast is lower than the expectation of the variance estimator which assumes independence between domains. The latter expectation will be inflated by common cluster effects. The main feature of these results of interest here is that there is again no tendency for the meffs to converge to one as the number of waves increases. If there is a trend, it is in the opposite direction. For the contrast of particular scientific interest, that between women who are full-time employed and those who are 'at home caring for a family', the meff is consistently well below one.

We next refine the model further by including, as additional covariates, age group, year and qualifications. The estimated meffs are given in Table 3. The meffs for the regression coefficients corresponding to categories of

economic activity again vary, some being above one and some below one, for the same reasons as for the contrasts (which may also be interpreted as regression coefficients) in Table 2. There is again some evidence of a tendency for these meffs to diverge away from one as the number of waves increases. A comparison of Tables 1 and 3 confirms the observation of Kish and Frankel (1974) that meffs for regression coefficients tend not to be greater than meffs for the means of the dependent variable.

Table 2 Estimates for regression with covariates defined by economic activity

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Intercept	20.58	0.11	1.13	1.01	1.09	1.38	1.50
Contrasts for							
PT employed	-1.03	0.10	0.93	0.91	0.93	1.00	0.89
Other inactive	-0.80	0.15	0.60	0.96	0.68	0.76	0.81
FT student	0.41	0.24	1.10	1.32	1.14	1.48	1.44
Family care	-2.18	0.10	0.72	0.49	0.58	0.66	0.60

Note: a) intercept is mean for women full-time employed
b) contrasts are for other categories of economic activity relative to full-time employed

Table 3 Estimates for regression coefficients with additional covariates in model

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Intercept	20.20	0.30	0.95	0.87	0.87	1.04	1.07
Year, *t*	-0.04	0.01	-	0.86	0.69	0.59	0.96
Age Group							
16-21	0.00	-					
22-27	-0.71	0.25	1.22	1.37	1.44	1.73	1.64
28-33	-0.89	0.27	1.38	1.40	1.46	1.68	1.59
34+	-1.03	0.27	0.94	1.10	1.13	1.26	1.34
Economic Activity							
FT employed	0.00	-					
PT employed	-0.93	0.10	0.97	0.95	0.96	1.06	0.91
Other inactive	-0.75	0.15	0.60	0.96	0.68	0.77	0.81
FT student	0.17	0.24	0.93	1.32	1.23	1.39	1.32
Family care	-2.09	0.10	0.77	0.59	0.70	0.78	0.67
Qualification							
Degree	0.00	-					
QF	-0.52	0.21	0.77	0.64	0.75	0.87	0.85
A-level	-0.61	0.24	0.98	0.87	0.94	0.94	1.01
O-level	-0.44	0.20	0.62	0.62	0.59	0.69	0.73
Other	-1.16	0.22	0.83	0.83	0.78	0.80	0.82

We next consider model-based standard errors obtained from the three level model in (5), as discussed in section 2. The results are given in Table 4 in the column headed '3 level model-based'. For comparison, we also estimate the standard errors under the two level model in (2) - the results are in the column headed '2 level model-based'. The estimates in the two columns are virtually identical. There is a single digit difference in the third decimal place for some coefficients and slightly greater difference for the intercept term. We suggest that this is evidence that simply adding in a random area effect term can seriously understate the impact of clustering on the standard errors of the estimated regression coefficients. This evidence is in line with the theoretical upper bound for the meff in (11). The estimated value of $\tilde{\tau}$ in (11) is 0.019 and none of the covariates may be expected to display important intra-area correlation so the expected values of the variance estimators for the two-level and three-level models would be expected to be very close.

We suggested in Section 3 that the main feature of clustering likely to impact on the covariance matrix of $\hat{\beta}$ is the variation in regression coefficients between clusters. We have explored this idea by introducing random coefficients in the model. Treating the elements of β now as the expected values of the random coefficients, we found that the estimates of β were hardly changed. We found that the estimated standard errors of these estimates were indeed inflated, much more so than from the introduction of the extra cluster random effect in model (5), and that the inflation was of an order similar to those of the meffs in Tables 2 and 3. Nevertheless, the IGLS method did lead to several negative estimates of the variances of the random coefficients, raising issues of which coefficients to allow to vary or more generally the issue of model specification. This problem is accentuated with increasing numbers of covariates, as the number of parameters in the covariance matrix of the coefficient vector increases with the square of the number of covariates. Overall, the inclusion of random coefficients seems to raise at least as many problems as it solves, if the clustering is not of intrinsic scientific interest, and thus does not seem a very satisfactory way to allow for clustering in variance estimation. It is simpler to change the method of variance estimation.

As mentioned at the end of Section 2, one alternative is a 'robust' variance estimation method based on the model in (5) (Goldstein 2003, page 80). Values of such robust standard error estimates are also included in Table 4. As anticipated in Section 2, the robust standard error estimator for the two level model performs very similarly to the linearization estimator which ignores clustering. The robust standard error estimator for the three level model performs very similarly to the linearization estimator which allows for two stage sampling. The slight differences reflect the differences between the methods of estimating V.

Table 4 Estimated standard errors of regression coefficients

	Linearization		Multilevel modelling			
	SRS	complex	2 level model-based	2 level robust	3 level model-based	3 level robust
Intercept	0.287	0.296	0.253	0.288	0.259	0.293
Year, *t*	0.014	0.014	0.013	0.014	0.013	0.014
Age Group						
16-21						
22-27	0.191	0.245	0.155	0.192	0.155	0.243
28-33	0.214	0.270	0.187	0.215	0.187	0.266
34+	0.237	0.275	0.218	0.238	0.218	0.271
Economic Activity						
FT employed						
PT employed	0.103	0.098	0.098	0.103	0.098	0.096
Other inactive	0.166	0.150	0.146	0.166	0.146	0.148
FT student	0.207	0.238	0.199	0.207	0.199	0.236
Family care	0.125	0.102	0.112	0.125	0.112	0.101
Qualification						
Degree						
QF	0.228	0.210	0.207	0.228	0.208	0.211
A-level	0.238	0.239	0.209	0.240	0.210	0.237
O-level	0.234	0.199	0.217	0.235	0.218	0.199
Other	0.247	0.224	0.229	0.249	0.230	0.223

The linearization method in the presence of two-stage sampling is thus very close to robust variance estimation methods used in the literature on multilevel modeling. The distinction between the methods becomes stronger if we allow also for stratification and weighting. Another distinction is that in the multilevel modeling approach, differences between model-based and the robust standard errors might be used as a diagnostic tool to detect departures from the model (Maas and Hox 2004). For example, the large differences in the three-level standard errors for the coefficients of age group in Table 4 might lead to consideration of the inclusion of random coefficients for age group. This contrasts with the survey sampling approach where the error structure in model (5) is only treated as a working model and it is not necessarily expected that standard errors based upon this model will be approximately valid.

5. Discussion

We have presented some theoretical arguments and empirical evidence that the impact of ignoring clustering in standard error estimation for certain longitudinal analyses can tend to be larger than for corresponding cross-sectional analyses. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses and that the findings of, for example, Kish and Frankel (1974), should not be used as grounds to ignore complex sampling in the former case.

The longitudinal analyses considered in this paper are of a certain kind and we should emphasise that the patterns observed for meffs in these kinds of analyses may well not extend to all kinds of longitudinal analyses. To speculate about the class of models and estimators for which the patterns observed in this paper might apply, we conjecture that increased meffs for longitudinal analyses will arise when the longitudinal design enables temporal 'random' variation in individual responses to be extracted from between-person differences and hence to reduce the component of standard errors due to these differences, but provides less 'explanation' of between cluster differences, so that the relative importance of this component of standard errors becomes greater.

The empirical work presented in this paper has also been restricted to the impact of clustering. We have undertaken corresponding work allowing for weighting and stratification and found broadly similar findings. Stratification tends to have a smaller effect than clustering. The sample selection probabilities in the BHPS do not vary greatly and the impact of weighting by the reciprocals of these probabilities on both point and variance estimates tends not to be large. There is rather greater variation among the longitudinal weights which are provided with BHPS data for analyses of sets of individuals who have responded at each wave up to and including a given year T. The impact

of these weights on point and variance estimates is somewhat greater. As T increases and further attrition occurs, the longitudinal weights tend to become more variable and lead to greater inflation of variances. This tends to compound the effect we have described of meffs increasing with T.

Leaving aside consideration of stratification and weighting, we have compared two approaches to allowing for cluster sampling. We have treated the survey sampling approach as a benchmark. We have also considered a multilevel modelling approach to allow for clustering. We have suggested that the use of a simple additive random effect to represent clustering can seriously understate the impact of clustering and may lead to underestimation of standard errors. If the clustering is of scientific interest, one solution would be to consider including random coefficients. Another would be to use the 'GEE2' approach (Liang, Zeger and Qaqish 1992) and specify an additional parametric model for $E(y_i\, y_i')$. If the clustering is treated as a nuisance, simply reflecting administrative convenience in data collection, we suggest the survey sampling approach has a number of practical advantages. This is discussed further by Lavange *et al.* (1996, 2001) in relation to other applications to repeated measures data.

Appendix

Justification for (11)

For simplicity, x and β are taken to be scalar, $\hat{\beta}$ is taken to be the ordinary least squares estimator and it is assumed that the sample sizes within clusters are all equal to m. The meff in (11) is defined as $\mathrm{var}_3(\hat{\beta})/E_3[v_2(\hat{\beta})]$, where E_3 and var_3 are moments with respect to the three-level model in (5) and $v_2(\hat{\beta})$ is a variance estimator based upon the two-level model in (2). Under (5) we obtain

$$\mathrm{var}_3(\hat{\beta}) = \left(\sum_{cit} x_{cit}^2\right)^{-2}\left(\sigma_\eta^2\sum_c x_{c++}^2 + \sigma_u^2\sum_{ci} x_{ci+}^2 + \sigma_v^2\sum_{cit} x_{cit}^2\right),$$

where + denotes summation across a suffix, $\sigma_\eta^2, \sigma_u^2$ and σ_v^2 are the respective variances of η_a, u_{ai} and v_{ait} and x_{cit} is centred at 0. We further suppose that $v_2(\hat{\beta})$ is defined so that $E[v_2(\hat{\beta})] \approx (\sum_{cit} x_{cit}^2)^{-2}[(\sigma_\eta^2+\sigma_u^2)\sum_{ci} x_{ci+}^2 + \sigma_v^2\sum_{cit} x_{cit}^2]$. After some algebra we may show that

$$\mathrm{meff} = 1+(m-1)\tilde{\tau}\tau_z\rho[1+(T-1)\tau_x]/[1+(T-1)\rho\tau_x], \quad (12)$$

where $\tilde{\tau} = \sigma_\eta^2/(\sigma_\eta^2+\sigma_u^2)$, $\rho = (\sigma_\eta^2+\sigma_u^2)/(\sigma_\eta^2+\sigma_u^2+\sigma_v^2)$, $\tau_x = \sigma_{xB}^2/\sigma_x^2$, $\sigma_x^2 = \sum_{cit} x_{cit}^2/(nT)$, $\sigma_{xB}^2 = [\sum_{ci}(x_{ci+}/T)^2/n - \sigma_x^2/T]/[1-1/T]$, $\tau_z = \sigma_{zB}^2/\sigma_z^2$, $\sigma_z^2 = \sum_{ci} z_{ci}^2/n$, $\sigma_{zB}^2 = [\sum_c (z_{c+}/m)^2/C - \sigma_z^2/m]/[1-1/m]$ and $n = Cm$ is the sample size. Note that $\tilde{\tau}\rho = \tau_1$ and, when $T = 1$, $\tau_z = \tau_x$ so that (12) reduces to (10). In general $\rho \le 1$ and (11) follows from (12). In fact, we estimate ρ as 0.59 in our application so the bound in (11) is not expected to be very tight.

Acknowledgements

The research of the second author was supported by grant 20.0286/01.3 from the Brazilian National Council for Scientific and Technological Development (CNPq).

References

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*. 2nd Ed. Chichester: John Wiley & Sons, Inc.

Berrington, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study. In *Meaning and Choice*: *Value Orientations and Life Course Decisions*, (Ed., R. Lesthaeghe) Brussels: NIDI.

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

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-92.

Diggle, P.J., Heagerty, P., Liang, K. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2nd Ed. Oxford: Oxford University Press.

Fan, P.-L., and Marini, M.M. (2000). Influences on gender-role attitudes during the transition to adulthood. *Social Science Research*, 29, 258-283.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*. Vol. 37, Series C, 117-132.

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

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

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 74, 430-431.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Ed. London: Arnold.

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

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

Lavange, L.M., Koch, G.G. and Schwartz, T.A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, 20, 2609-23.

Lavange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. and Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, 5, 311-329.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, Series B, 54, 3-40.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270-278.

Lynn, P., and Lievesley, D. (1991). *Drawing General Population Samples in Great Britain*. London: Social and Community Planning Research.

Maas, C.J.M., and Hox, J.J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427-440.

Morgan, S.P., and Waite, L.J. (1987). Parenthood and the attitudes of young adults. *Am. Sociological Review*, 52, 541-547.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-56.

Renard, D., and Molenberghs, G. (2002). Multilevel modelling of complex survey data. In *Topics in Modelling Clustered Data* (Eds., M. Aerts, H. Geys, G. Molenberghs and L.M. Ryan). Boca Raton: Chapman and Hall/CRC. 263-272.

Scott, A.J., and Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

Shah, B.V., Barnwell, B.G. and Bieler, G.S. (1997). SUDAAN User's manual, release 7.5. Research triangle park, NC: Research Triangle Institute.

Skinner, C.J. (1986). Design effects of two stage sampling. *Journal of the Royal Statistical Society*, Series B, 48, 89-99.

Skinner, C.J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt and T.M.F. Smith) Chichester: John Wiley & Sons, Inc. 23-58.

Skinner, C.J. (1989b). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt and T.M.F. Smith) Chichester: John Wiley & Sons, Inc. 59-87.

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

Taylor, M.F. ed, Brice, J., Buck, N. and Prentice-Lane, E. (2001). *British Household Panel Survey - User Manual - Volume A*: *Introduction*, *Technical Report and Appendices*. Colchester, University of Essex.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 13-22
Statistics Canada, Catalogue No. 12-001

Modelling durations of multiple spells from longitudinal survey data

Milorad S. Kovačević and Georgia Roberts [1]

Abstract

We investigate some modifications of the classical single-spell Cox model in order to handle multiple spells from the same individual when the data are collected in a longitudinal survey based on a complex sample design. One modification is the use of a design-based approach for the estimation of the model coefficients and their variances; in the variance estimation each individual is treated as a cluster of spells, bringing an extra stage of clustering into the survey design. Other modifications to the model allow a flexible specification of the baseline hazard to account for possibly differential dependence of hazard on the order and duration of successive spells, and also allow for differential effects of the covariates on the spells of different orders. These approaches are illustrated using data from the Canadian Survey of Labour and Income Dynamics (SLID).

Key Words: Cox regression; Design-based inference; Model-based inference; Spell order; SLID.

1. Introduction

The modelling problem addressed in this paper is known under different names such as correlated failure-time modelling, multivariate survival modelling, multiple spells modelling, or a recurrent events problem. It has been studied in the biomedical (*e.g.*, Lin 1994, Hougaard 1999), social (Blossfeld and Hamerle 1989, Hamerle 1989) and economic literature (Lancaster 1979, Heckman and Singer 1982). Generally this type of modelling is required to address issues that arise in time-to-event studies when two or more events occur to the same subject and where the research interest is to assess the effect of various covariates on the length of a spell. Failure times are correlated within a subject, and thus the assumption of independence of failure times conditional on given measured covariates, required by standard survival models, is likely to be violated. In studies of duration of spells (poverty, unemployment, *etc*.), the "failure" is equivalent to exiting from the state of interest. An additional property of many multiple spells, often ignored, is that the spells are ordered "events"; that is, the second spell cannot occur before the first, *etc*. This paper was motivated by a study of unemployment spells, discussed further in Section 5.

The dependence among the spells from the same individual arises from the fact that these spells share certain unobserved characteristics of the individual. The effect of these unobserved characteristics can be explicitly modelled as a random effect (*e.g.*, Clayton and Cuzick 1985). When this is done, it is assumed that the random effect follows a known statistical distribution. The gamma distribution with mean 1 and unknown variance is the distribution of choice in many applications. Then, estimates of random and fixed effects can be obtained by some suitable method (*e.g.*, two-stage likelihood (Lancaster 1979), using an EM algorithm (Klein 1992), *etc*.). This paper does not explore this approach.

Another approach that has been taken - and is the one that we will be using - is to take a semi-parametric approach where we do not explicitly model the dependence among multiple spells. We model the marginal distributions of the individual spells, with a possible utilization of the order of the spells in the model specification. In the non-survey context, Lin (1994) describes how it is sufficient just to modify the "naïve" covariance matrix of the estimated model coefficients obtained under the assumption of independence since the correlated durations need to be accounted for in the variance estimates but not in the estimates of coefficients per se.

In socio-economic studies of spell durations the data sources are frequently longitudinal surveys with complex sample designs that involve stratification, sampling in several stages, selection with unequal probabilities, stochastic adjustments for attrition and non-response, calibration to known parameters, *etc*. Consequently, it is necessary to account for the impact of the sample design on the distribution of the sample data when estimating model parameters and the variances of these estimates. Our approach when analyzing complex survey data is to model the marginal distributions of the multiple spells using single-spell methods, treating the dependence among the spells as a nuisance - both the dependence due to the correlation of spells from the same person and dependence among individuals due to the survey design - but taking account of the unequal selection probabilities through the survey weights. Based on the model chosen, finite population

1. Milorad S. Kovačević, Methodology Research Advisor, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: kovamil@statcan.ca; Georgia Roberts, Chief of the Data Analysis Resource Center at the Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: robertg@statcan.ca.

parameters are defined and estimated as in Binder (1992). Standard errors are estimated using an appropriate design-consistent linearization method under the assumption that the primary sampling units are sampled with replacement within strata. This assumption is viable when the sampling rates at the first stage are small, as is generally the case in socio-economic surveys. Also, for such samples, the difference between finite population and superpopulation inference (*i.e.*, the standard errors and the test statistics) has been found to be rather negligible (Lin 2000). Therefore, the results from inference based on our approach extend beyond the finite population under study.

In the next section we review single-spell modelling and some methods for robust estimation of variances when the model is misspecified - first under a model-based framework and then under a design-based one. Section 3 contains further discussion of robust variance estimation for multiple spells. In Section 4, we introduce three models for multiple spells and describe how to fit these models using design-based robust estimation methods. In Section 5, we fit these models to data from the Canadian Survey of Labour and Income Dynamics (SLID) and discuss the results. Finally, Section 6 contains some overall remarks.

2. Inference for the single-spell hazard rate model

The duration of a spell (or simply, a spell) experienced by an individual is a random variable denoted by T. We are particularly interested in the hazard function $h(t)$ of T at time t, defined as the instantaneous rate of spell completion at time t given that it has not been completed prior to time t, formally

$$h(t) = \lim_{dt \to 0} \frac{\text{Prob}\{t \le T < t + dt \mid T \ge t\}}{dt}.$$

The value of the hazard function at t is called the exit rate to emphasize that the completion of the spell is equivalent to exiting the state of interest. Duration models and analysis of duration in general are formulated and discussed in terms of the hazard function and its properties.

From a subject matter perspective, frequently the main interest is to study the impact of some key covariates on the distribution of T. A proportional hazards model is often chosen for such a study. Under the proportional hazards model, the hazard function of the spell T given a vector of possibly time-varying covariates $\mathbf{x}(t) = (x_1(t), ..., x_p(t))'$ is

$$h(t \mid \mathbf{x}(t)) = \lambda_0(t)\, e^{\mathbf{x}'(t)\boldsymbol{\beta}}. \qquad (1)$$

The function $\lambda_0(t)$ is an unspecified baseline hazard function and gives the shape of $h(t \mid \mathbf{x}(t))$. The baseline hazard describes the duration dependence, such as whether the hazard rate depends on time already spent in the spell. For example, negative dependence describes the situation where the longer the spell the smaller the probability of exit. If an individual has all $\mathbf{x}(t)$ variables set at 0, the value (level) of the hazard function is equal to the baseline hazard.

2.1 Model-based inference

The vector $\boldsymbol{\beta}$ contains the unknown regression parameters showing the dependence of the hazard on the $\mathbf{x}(t)$ vector, and may be estimated by maximizing the partial likelihood function (Cox 1975):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i'(T_i)\boldsymbol{\beta}}}{\sum_{j=1}^{n} Y_j(T_i)\, e^{\mathbf{x}_j'(T_i)\boldsymbol{\beta}}} \right]^{\delta_i}. \qquad (2)$$

Here $T_1, ..., T_n$ are n possibly right-censored durations; $\delta_i = 1$ if T_i is an observed duration and $\delta_i = 0$ otherwise; and $\mathbf{x}_i(t)$ is the corresponding covariate vector observed on $[0, T_i]$. The denominator sum is taken over the spells that are at risk of being completed at time T_i, *i.e.*, $Y_j = 1$ if $t \le T_j$, and is equal to 0 otherwise. The estimate $\hat{\boldsymbol{\beta}}$ of the model parameter $\boldsymbol{\beta}$ is obtained by solving the partial likelihood score equation

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^{n} u_{i0}(T_i, \boldsymbol{\beta}) = 0, \qquad (3)$$

where

$$u_{i0}(T_i, \boldsymbol{\beta}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right\}, \qquad (4)$$

$$S^{(0)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}, \qquad (5)$$

and

$$S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}. \qquad (6)$$

If model (1) is true and the durations are independent, the model-based variance matrix of the score function $U_0(\boldsymbol{\beta})$ is

$$J(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$$

$$= \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} - \frac{S^{(1)}(T_i, \boldsymbol{\beta})[S^{(1)}(T_i, \boldsymbol{\beta})]'}{[S^{(0)}(T_i, \boldsymbol{\beta})]^2} \right\},$$

where

$$S^{(2)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t)\, \mathbf{x}_i'(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}.$$

The approximate variance of $\hat{\boldsymbol{\beta}}$, obtained by linearization, is $J^{-1}(\hat{\boldsymbol{\beta}})$.

If the form of (1) is incorrect but observations are independent, Lin and Wei (1989) provide the robust variance estimator for $\hat{\boldsymbol{\beta}}$ as

$$J^{-1}(\hat{\boldsymbol{\beta}})\, G(\hat{\boldsymbol{\beta}})\, J^{-1}(\hat{\boldsymbol{\beta}}), \tag{7}$$

where

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta})\, g_i'(\boldsymbol{\beta})$$

and

$$g_i(\boldsymbol{\beta}) = u_{i0}(T_i, \boldsymbol{\beta})$$

$$-\sum_{j=1}^{n} \delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\boldsymbol{\beta}}}{nS^{(0)}(T_j, \boldsymbol{\beta})} \left\{ \mathbf{x}_i(T_j) - \frac{S^{(1)}(T_j, \boldsymbol{\beta})}{S^{(0)}(T_j, \boldsymbol{\beta})} \right\}. \tag{8}$$

2.2 Design-based inference

For observations from a survey with a complex sample design, Binder (1992) used a pseudo-likelihood method to estimate the parameters and their variances for a proportional hazards model in the case of a single spell per individual. In particular, he first defined the finite population parameter of interest as a solution of the partial likelihood score equation (3) calculated from the spells of the finite population targeted by the survey:

$$U_0(\mathbf{B}) = \sum_{i=1}^{N} u_{i0}(T_i, \mathbf{B}) = 0,$$

where $u_{i0}(T_i, \mathbf{B})$ is the score residual defined in the same way as $u_{i0}(T_i, \boldsymbol{\beta})$, except that the averages in the definitions of $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ extend over N observations rather than n. Note that if all members of the finite population targeted by the survey do not experience spells, N represents the size of the subpopulation that experiences spells, and the summation is over these N individuals.

An estimate $\hat{\mathbf{B}}$ of the parameter $\mathbf{B}$ is obtained as a solution to the partial pseudo-score estimating equation

$$\hat{U}_0(\hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

where $w_i(s) = w_i$, the survey weight, if $i \in s$, and 0 otherwise. Function $\hat{u}_{i0}(T_i, \hat{\mathbf{B}})$ takes the form

$$\hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

where

$$\hat{S}^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}},$$

and

$$\hat{S}^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, \mathbf{x}_i(t)\, e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}}.$$

Generally, the design-based variance of an estimate $\hat{\boldsymbol{\theta}}$ that satisfies an estimating equation of the form $\hat{U}(\hat{\boldsymbol{\theta}}) = \sum w_i\, u_i(\hat{\boldsymbol{\theta}}) = 0$ can be estimated, using linearization, as

$$\hat{J}^{-1}\, \hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))\, \hat{J}^{-1}, \tag{9}$$

where $\hat{J} = \partial \hat{U}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and $\hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))$ is the estimated variance of the estimated total $\hat{U}(\boldsymbol{\theta})$ obtained by some standard design-based variance estimation method (see for example Cochran (1977)) and evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) states that in order to use this approach to derive a consistent estimate of the variance, $\hat{U}(\hat{\boldsymbol{\theta}})$ must be expressed as a sum of independent random vectors. In the case of the proportional hazards model above, $\hat{U}_0(\hat{\mathbf{B}})$ does not satisfy this condition since each $\hat{u}_{i0}$ is a function of $\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})$, both of which include many individuals besides the i^{th} one. Thus, Binder (1992) found an alternative expression for $\hat{U}_0(\hat{\mathbf{B}})$ which conforms to these conditions, making it possible to obtain a design consistent estimate $\hat{V}(\hat{U}_0(\hat{\mathbf{B}}))$ by application of a design-based variance estimation method to the alternate expression and then evaluating this variance estimate at $\mathbf{B} = \hat{\mathbf{B}}$. If the design-based variance estimation method chosen is the linearization method, then the first step consists of calculating the following residual for each of the sampled individuals:

$$\hat{u}_i(T_i, \hat{\mathbf{B}}) = \hat{u}_{i0}(T_i, \hat{\mathbf{B}})$$

$$-\sum_{j=1}^{N} w_j(s)\, \delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\hat{\mathbf{B}}}}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \left\{ \mathbf{x}_i(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \right\}. \tag{10}$$

Each individual in the sample belongs to a particular PSU within a particular stratum. Thus, instead of identifying an individual by a single subscript i we will use a triple subscript hci where $h = 1, 2, \ldots, H$ identifies the stratum, $c = 1, 2, \ldots, c_h$ identifies the PSU within the stratum and $i = 1, 2, \ldots, n_{hc}$ identifies the individual within the PSU. Then

$$\hat{V}(\hat{U}_0(\hat{\mathbf{B}})) = \sum_{h=1}^{H} \frac{1}{c_h(c_h - 1)} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h)\, (t_{hc} - \bar{t}_h)',$$

where

$$t_{hc} = c_h \sum_{i=1}^{n_{hc}} w_{hci}\, \hat{u}_{hci} \quad \text{and} \quad \overline{t}_h = \sum_{c=1}^{c_h} t_{hc} \Big/ c_h.$$

3. Inference for multiple-spell hazard rate models

3.1 Model-based inference

If more than one spell is observed for an individual, it is realistic to assume that these spells are not independent. Thus, the partial likelihood function (2) is misspecified for multiple spells since it does not account for intra-individual correlation of the spells observed on the same individual. Following Lin and Wei (1989), it is sufficient to modify only the covariance matrix of the estimated model parameters since the correlated durations affect the variance while the model parameters can be estimated consistently without accounting for this correlation. Lin (1994) demonstrates how the covariance matrix of the estimated model parameters might be estimated when there is intra-individual correlation of spells, provided that spells from different individuals are independent.

3.2 Design-based inference

In a longitudinal survey with a multi-stage design, the multiple events can be correlated at different levels: the spells are clustered within an individual, and individuals are clustered within high-stage units. The positive intracluster correlation at any level adds extra variation to estimates calculated from such data, beyond what is expected under independence. The assumption of independence of observations when they are cluster-correlated leads to underestimating the true standard errors, which inflates the values of test statistics, and ultimately results in too-frequent rejection of null hypotheses. Thus, for multiple spells for individuals, where the data are from a longitudinal survey, accounting just for correlation within individuals is insufficient.

Design-based variance estimation for nested cluster-correlated data can be greatly simplified when it is reasonable to assume that individuals from different primary sampling units (PSU's) are uncorrelated. This is equivalent to assuming that the PSU's are sampled with replacement. This assumption also holds approximately when the first stage units are obtained by sampling without replacement, provided that the sampling rate at the first stage is very small. In such a case, an estimate of the between-PSU variability captures the variability among units in all subsequent stages, regardless of the dependence structure among observations within each PSU. For a recent summary of robust variance estimation for cluster-correlated data see Williams (2000). This implies that Binder's (1992) approach for robust variance estimation of the single-spell model in the case of a survey design having with-replacement sampling at the first stage can be directly applied to the multiple spell situation since it accounts for the impact of cluster-correlation at all levels within each PSU.

4. Three models for multiple spells

In order to allow the covariates to have different effects for spells of different orders, as well as to allow different time dependencies (baseline hazards), we are exploring three models for multiple spells. The models differ according to the definition of the risk set and the assumptions about the baseline hazard. Two of these models account for the order of the spells.

It should be noted, however, that in our work, spell order refers only to spells occurring in the observation period from which the data are collected and not to the entire history of an individual (unless these two time periods coincide). For example, by the first spell we mean a first spell in the observation period although it may be a spell of some higher absolute order over the person's lifetime. This limitation implies a careful interpretation of any impact that spell order may have on covariate effects or on time dependency.

Model 1: In the first model, the risk set is carefully defined to take spell order into account in the sense that an individual cannot be at risk of completing the second spell before he completes the first, *etc*. This model, known as the conditional risk set model, was proposed by Prentice, Williams and Peterson (1981) and was reviewed by Lin (1994). It was also discussed by Hamerle (1989) and Blossfeld and Hamerle (1989) in the context of modelling multi-episode processes. Generally, the conditional risk set at time t for the completion of a spell of order j consists of all individuals that are in their j^{th} spells. This model allows spell order to influence both the effect of covariates and the shape of the baseline hazard function.

The hazard function for the i^{th} individual for the spell of j^{th} order is

$$h_j(t \mid \mathbf{x}_{ij}(t)) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}_j},$$

where, for each spell order, a different baseline hazard function and a different coefficient vector are allowed. For this model and for other models that will be considered in this Section, time t is measured from the beginning of the j^{th} spell. Although spells within the same individual may not be independent, the following partial likelihood is still valid for estimation of the $\boldsymbol{\beta}_j$'s:

$$L(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[\frac{e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij}) e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}} \right]^{\delta_{ij}}, \quad (11)$$

Here, $T_{1j}, ..., T_{N_j j}$ are N_j durations of possibly right-censored j^{th} order spells, $\delta_{ij} = 1$ if T_{ij} is an observed duration and $\delta_{ij} = 0$ otherwise, and K is the highest order of spells to be included in the Cox model. The denominator sum is taken over the j^{th} spells that are at risk of being completed at time T_{ij}, *i.e.*, $Y_{rj}(t) = 1$ if $t \le T_{rj}$, and is equal to 0 otherwise. The corresponding covariate vector observed on $[0, T_{ij}]$ is $\mathbf{x}_{ij}(t)$. Partial likelihood (11) can be maximized separately for each j if there are no additional restrictions on the $\boldsymbol{\beta}_j$'s.

The corresponding score equations that define the finite population parameter $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, ..., \mathbf{B}'_K)'$ are:

$$U_0(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u_{ij0}(T_{ij}, \mathbf{B}_j) = 0, \quad (12)$$

with

$$u_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B}_j)}{S^{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

and with $S^{(0)}(t, \mathbf{B}_j)$ and $S^{(1)}(t, \mathbf{B}_j)$ having the form of (5) and (6) respectively, but with N_j replacing n and $\mathbf{B}_j$ replacing $\boldsymbol{\beta}$.

The design-based estimates of the parameters $\mathbf{B}_j$ are obtained by solving equations $\sum_{i=1}^{N_j} w_i(s) \hat{u}_{ij0}(T_{ij}, \hat{\mathbf{B}}_j) = 0$ separately for each j, where $\hat{u}_{ij0}$ has the form of u_{ij0} but with $S^{(0)}$ and $S^{(1)}$ replaced by $\hat{S}^{(0)}$ and $\hat{S}^{(1)}$ respectively. Note that the sampling weights correspond to individuals and not to spells. Similarly, estimation of the covariance matrix of each $\hat{\mathbf{B}}_j$ will be done separately using the design-based robust estimation approach described in Section 2.2. Technically, this is a set of analyses separated by spell order.

Model 2: The second model considered is the marginal model (Wei, Lin and Weissfeld 1989):

$$h_j(t \mid \mathbf{x}_{ij}(t)) = \lambda_{0j}(t) e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}},$$

where, for each spell order, we allow a different baseline hazard function while the covariate effects are kept the same over different spell orders. The corresponding partial likelihood function as well as the risk set, under the assumption that spells within the same individual are independent, is the same as for Model 1, with $\boldsymbol{\beta}$ replacing the $\boldsymbol{\beta}_j$'s. The corresponding score equation that defines the finite population parameter is

$$U_0^*(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u^*_{ij0}(T_{ij}, \mathbf{B}) = 0,$$

with

$$u^*_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B})}{S^{(0)}(T_{ij}, \mathbf{B})} \right\},$$

where $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ are defined by (5) and (6) respectively, but with N_j replacing n and $\mathbf{B}$ replacing $\boldsymbol{\beta}$.

The design-based estimate of the parameter $\mathbf{B}$ is obtained by solving the weighted score equations

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} w_i(s) \hat{u}^*_{ij0}(T_{ij}, \hat{\mathbf{B}}) = 0,$$

where $\hat{u}^*_{ij0}$ has the form of u^*_{ij0} but with $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ replaced by $\hat{S}^{(0)}(t, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(t, \hat{\mathbf{B}})$ respectively.

The estimation of the covariance matrix of $\hat{\mathbf{B}}$ will be done using the design-based robust estimation approach explained in Section 3.2.

Model 3: The last model considered is the following:

$$h_j(t \mid \mathbf{x}_{ij}) = \lambda_0(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}}.$$

In this model we assume that the baseline hazard functions and the effects of covariates are common for different orders of spells. The risk set at time T_{ij} is defined differently than for Models 1 and 2, and contains all spells with $t \le T_{ij}$, effectively assuming that all spells are from different individuals. Technically, this model is a single-spell model, so that estimation of coefficients and variances by a design-based robust method is straightforward.

5. Example of modelling multiple unemployment spells

5.1 The data

The data set that we use for illustration comes from the first six-year panel (1993-1998) of the Canadian Survey of Labour and Income Dynamics (SLID). In this panel, about 31,000 individuals from approximately 15,000 households were followed for six years through annual interviews. Some individuals dropped out of the sample over time for any number of reasons while a few others, after missing one or more interviews, resumed their participation. A complex weighting of the responding SLID individuals each year takes into account different types of attrition so that each respondent in a particular year is weighted against the

relevant reference population of 1993. This results in a separate longitudinal weight for each wave (*i.e.*, year) of data. For this analysis we used the longitudinal weights from the last year of the panel, *i.e.*, 1998, which meant that data just from those individuals who were respondents in the final wave of the panel were included in the analyses. A good summary of the sample design issues in SLID is given in Lavigne and Michaud (1998). A review of the issues related to studies of unemployment spells from SLID is given in Roberts and Kovačević (2001).

The state of interest is "being unemployed", defined in this case as the state between a permanent layoff from a full-time job and the commencement of another full-time job. A job is "full-time" if it requires at least 30 hours of work per week. The event of interest is "the exit from unemployment". Only spells beginning after January 1, 1993 were included since January 31, 1993 is the starting date for observations from the panel. Spells that were not completed by the end of the observation period (December 31, 1998) were considered censored. Sample counts of the number of individuals experiencing eligible spells and the number of spells according to their order are given in Table 1. In brief, there were17,880 spells from 8,401 longitudinal individuals. About half of the sampled individuals (4,260) who became unemployed during this period experienced two or more unemployment spells. There were 3,809 spells that remained uncompleted due to the termination of the panel.

From a long list of available covariates we chose only ten. The variable sex [SEX] of the longitudinal individual is the only variable that remains constant over different spells. Four variables have values recorded at the end of the year in which the spell commenced: education level [EDUCLEV] with 4 categories (low, low-medium, medium, high), marital status [MARST] with three categories (single, married/common law, other), family income per capita (in Canadian dollars) with 4 categories (<10K, 10K-20K, 20K-30K, 30K+), and age [AGE] (in years). Three variables have the values from the lay-off job preceding the spell: type of job ending [TYPJBEND] with two categories (fired and voluntary), occupation [OCCUPATION] with 6 categories (professional, administration, primary sector, manufacturing, construction, and others); and firm size [FIRMSIZE] with five categories (<20, 20-99, 100-499, 500-999, 1,000 + employees). Two binary variables represent the situation during the spell: having a part time job [PARTTJB], and attending school [ATSCH].

The data set was prepared in the "counting process" style where each individual with eligible spells is represented by a set of rows, and each row corresponds to a spell. Although a row contains time of entry to the spell t_1, and time of exit t_2 or time of censoring t_c, the duration time for analysis is always considered in the form $(0,\ t_2 - t_1)$ or $(0,\ t_c - t_1)$. The covariates under consideration are attached to each row. Also attached to each row are the 1998 longitudinal weight and the identifiers for the stratum and the PSU of the person whose spell is being described by that record.

Table 1 Counts of individuals in the six-year panel of SLID with unemployment spells beginning between January 1993 and December 1998, by the total number of spells and by order of spell (C-completed, U-uncompleted)

Individuals by number of spells		Spells by order									
		First		Second		Third		Fourth		5th +	
		C	U	C	U	C	U	C	U	C	U
1 spell	4,141	2,221	1,920	-	-	-	-	-	-	-	-
2 spells	1,915	1,915	-	1,154	761	-	-	-	-	-	-
3 spells	1,044	1,044	-	1,044	-	612	432	-	-	-	-
4 spells	629	629	-	629	-	629	-	348	281	-	-
5+ spells	672	672	-	672	-	672	-	672	-	1,158	415
Total	8,401	6,481	1,920	3,499	761	1,913	432	1,020	281	1,158	415

5.2 Analysis

For the purpose of this illustration we restricted the analysis to the first four spells, which means that all sampled individuals with eligible spells are included in the analysis but the spell records after the fourth spell are not considered due to their small number in the sample.

We estimated coefficients and their variances for the 3 models by the design-based methods described in Section 4 through the use of the "SURVIVAL" procedure in SUDAAN Version 8. For all three models, the survey design was specified to be stratified with with-replacement selection of PSU's (*i.e.*, DESIGN = WR). All three models were fit to the same number of spells (16,307). For each model, we then calculated the empirical cumulative baseline hazard functions using a product-limit approach (see Kalbfleisch and Prentice (2002), pages 114-116) as implemented in the SURVIVAL procedure in SUDAAN.

In the robust model-based approach for multiple spells described in Section 3.1, there is an adjustment in the variance estimates to account for the possible dependence among spells from the same individual, assuming independence of spells from different individuals; however, in this approach, no account is made for the unequal probabilities of selection of the sampled individuals - in either the coefficient estimates or the variance estimates. In order to do this, for models 1 and 2 we also used the SURVIVAL procedure in SUDAAN Version 8, to estimate the variances of the weighted coefficient estimates where we assumed independence of spells between individuals but allowed for possible correlation of spells from the same individual. We did this by specifying the sampling design to be unstratified and having with-replacement selection of clusters, and we specified that each individual formed his own cluster. The dependence assumptions are the same as those used by Lin (1994) but we accounted for the use of weights in the estimation of the coefficients and the variances. We will call these variance estimates "modified robust model-based variance estimates of weighted coefficient estimates".

5.3 Some descriptive statistics

The estimated mean duration of a completed spell is 33.3 weeks while the estimated mean duration of the observed portion of a censored (uncompleted) spell is 48.5 weeks.

Visual examination of estimated Kaplan-Meier survival functions (not shown) for spells of each order indicated that, as order increased, the value of the survivor function at any fixed time t decreased, indicating that first spells are the longest among completed spells, and that the higher the order of a multiple spell the shorter is its duration. This is likely to be a consequence of the limited life of the panel, in the sense that an individual with more spells in the given six-year time frame is likely to have shorter spells.

5.4 Model fits using a design-based approach

As noted earlier, our example is just an illustration of the design-based approach to fitting proportional hazards models to multiple-event data from a survey with a complex design. Thus, little time is spent in this article on discussing how to assess the adequacy of these models, such as the adequacy of the proportionality assumptions in all of the models or whether one type of model fits as well as another.

Estimated coefficients from fitting the three models to the SLID data are given in Table 2. Coefficients found significant at the 5% level, through the use of individual t tests, are shown in bold.

Model 1 is conditional on the spell order and involved fitting four models separately to the data from the four different spell orders. As seen in Table 2, SEX, AGE, and at least one category of the Family Income variable were significant for spells of all orders, although magnitudes of the estimated coefficients differed with the spell order. The estimated coefficients for AGE were negative but decreased in magnitude as the spell order increased, while there was no discernable pattern in the estimated coefficients for the other 2 variables. The variables EDUCLEV, PARTJB and ATSCH had significant coefficients for spells of order 1, 2, and 3, but not for spells of order 4. This can be at least partly attributed to the small sample size for the fourth spells. For each of the other three variables in the model (MARST, OCCUPATION, and FIRMSIZE), there was just one spell order for which a coefficient was significant.

For Model 2, the model coefficients are restricted to be the same for all spell orders. As seen in Table 2, numerically many - but not all - of the estimated coefficient values were situated between the estimates for the first and the second spells obtained for Model 1 which could be due to the fact that a high proportion of the sample corresponded to events of these orders. All but the OCCUPATION variable had a significant coefficient. Standard errors of coefficients were smaller for Model 2 than for Model 1.

Model 3 is a single-spell model with a single set of model coefficients and a single baseline hazard function. The estimated model coefficients are similar to the estimates obtained by Model 2.

Table 2 Estimated β coefficients for three models

	Model 1				Model 2	Model 3
	Order 1	**Order 2**	**Order 3**	**Order 4**		
SEX (F)						
M	**0.4417**	**0.3781**	**0.3299**	**0.4435**	**0.4049**	**0.4090**
EDUCLEV (H)						
L	**-0.4561**	**-0.5234**	**-0.3748**	-0.1065	**-0.4128**	**-0.4331**
LM	**-0.2330**	**-0.2700**	**-0.3310**	-0.1653	**-0.2436**	**-0.2474**
M	-0.0744	-0.1060	-0.1156	0.0668	-0.0684	-0.0671
MARST (M)						
Single	**-0.1142**	-0.1290	-0.0622	-0.1375	**-0.1357**	**-0.1330**
Other	0.0985	-0.0894	0.1124	-0.1072	0.0328	0.0401
TYPJBEND (Fired)						
Voluntary	0.0704	**0.2752**	**0.4207**	**0.3413**	**0.1579**	**0.1284**
OCCUPATION(Othrs)						
Professionals	0.1592	-0.1364	-0.1388	0.0903	0.0490	0.0485
Admin	-0.0265	**-0.2930**	-0.1769	0.0579	-0.0971	-0.0938
PrimSector	-0.0211	-0.2175	-0.1187	0.2032	-0.0410	-0.0201
Manufacture	-0.0003	-0.0994	-0.1295	0.2862	-0.0093	-0.0088
Construction	0.1290	-0.1862	-0.0879	0.2339	0.0490	0.0813
FIRMSIZE (1000+)						
<20	-0.0027	-0.0097	0.1005	**0.4403**	0.0441	0.0408
20-99	0.0358	0.0881	0.0815	0.3999	**0.0928**	**0.0951**
100-499	0.0436	-0.0905	0.0328	0.0257	0.0214	0.0278
500-999	-0.0006	0.0153	-0.0623	-0.0067	-0.0005	0.0020
PARTTJB (No)						
Yes	**-0.2903**	**-0.5414**	**-0.5109**	-0.1407	**-0.3693**	**-0.3743**
ATSCH (No)						
Yes	**-1.0832**	**-1.1516**	**-1.2956**	-1.3541	**-1.1205**	**-1.1266**
Family Income Per Capita (10K-)						
10K-20K	**0.1294**	**0.1802**	0.0692	**0.1117**	**0.1345**	**0.1330**
20K-30K	**0.1644**	**0.3611**	0.1572	**0.4900**	**0.2241**	**0.2141**
30K+	**0.1712**	**0.3916**	**0.3005**	**0.4241**	**0.2280**	**0.2115**
AGE	**-0.0491**	**-0.0311**	**-0.0269**	**-0.0207**	**-0.0424**	**-0.0435**
Spells in risk set	8,386	4,255	2,345	1,300	16,286	16,286
Censored	1,913	759	432	281	3,385	3,385
Completed	6,473	3,496	1,913	1,019	12,901	12,901

The values significant at a 5% level are bold.

The estimated cumulative baseline hazard functions for Models 1 to 3 are given in Figures 1 to 3 respectively. In all cases, for durations up to approximately 50 weeks, the functions have a concave shape, implying that there is a positive time dependence of the exit rate (*i.e.* the longer the spell, the higher the probability of exit). For durations longer than 50 weeks, the shapes become convex, suggesting negative time dependence for the longer spells. In Figure 1, positions of the estimated cumulative baseline hazard functions vary according to spell order, with the curve for spells of order 1 being the highest, and the curve for spells of order 4 being the lowest. In Figure 2, for Model 2, the positions of the different curves do not follow spell order. This observed difference between Figures 1 and 2 could serve as one visual diagnostic that further study is required in order to assess whether Model 1 or Model 2 is a better descriptor of the data, since estimated coefficients have an impact on the estimated baseline hazards.

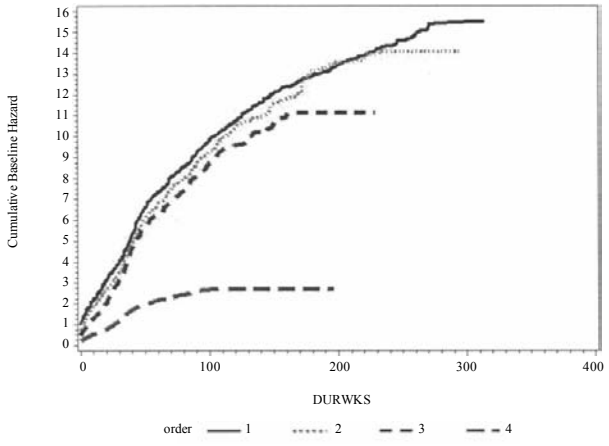

Figure 1 Cumulative Baseline Hazard – Model 1

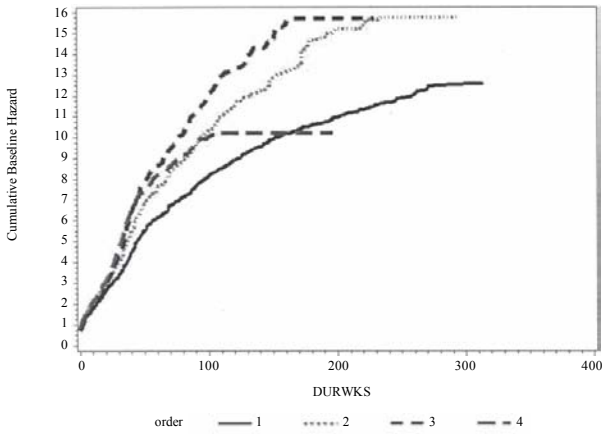

Figure 2 Cumulative Baseline Hazard – Model 2

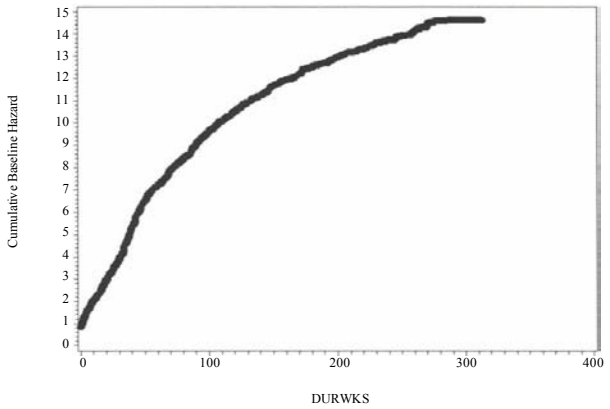

Figure 3 Cumulative Baseline Hazard – Model 3

5.5 Comparison to modified robust model-based variance estimates

As described in Section 5.2, the modified robust model-based variance estimates account for possible correlation among spells from the same individual, where independence among individuals is assumed. When, for Models 1 and 2, the standard error estimates obtained by this approach were compared to the design-based standard error estimates, only very minor differences were observed. This would seem to indicate that the design-based estimates are picking up any correlation among spells from the same individual and also that there does not appear to be additional dependence above the level of the individual for our particular example.

6. Concluding remarks

We explored the problem of analysis of multiple spells by considering two general approaches for dealing with the lack of independence among the exit times: a robust model-based approach and a design-based approach. The first approach estimates the model parameters assuming independence of the spells, and then corrects the naïve covariance matrix to account for within-individual dependencies postulated by the researcher. This approach does not account for the possible clustering between individuals (or, in fact, for any clustering that might occur at a level above the individual) nor for the unequal probabilities of selection of individuals (although, in our example, we showed how the method could be extended to include the survey weights). The second approach defines the model coefficients as finite population parameters. These parameters are then estimated accounting for possible unequal selection probabilities of individuals. A design-based variance estimation method that accounts for possible correlations between individuals in the same PSU automatically accounts for the unspecified dependencies of spells at levels below the PSU, such as dependencies within individuals. For large sample sizes this design-based inference extends directly to the super-population from which, hypothetically, the finite population was generated. The deficiency of the first approach is that it totally ignores the potential for clustering between individuals. A possible disadvantage of the second approach, as we applied it, is that it relies on the assumption of with-replacement sampling of PSU's of individuals. The two approaches coincide in the case of simple random sampling of individuals where, in the robust model-based approach, dependence among spells from the same individual is explicitly postulated and accounted for in the variance estimation formula and where, in the design-based approach, spells from the same individual are treated as a cluster in the design-based variance estimation.

We applied the design-based approach to three proportional-hazards-type models. One model allowed for differential unspecified baseline hazards and different coefficients for each spell order. The second model still allowed for differential unspecified baseline hazards for different spell orders but required the coefficients to be the same over orders. The third model was a simple single-spell model. We found that how information on the spell order was used affected the results of our model-fitting. A visual comparison of the coefficient estimates and the estimates of the cumulative baseline hazards for Models 1 and 2 indicated different results. A formal test for whether the coefficients actually differ by spell order (as allowed in Model 1), given baseline hazards that can differ by spell order, would be useful, as suggested by one of the referees. It is actually straightforward to produce such a test, and can be done as follows. Let $\boldsymbol{\gamma} = (\mathbf{B}_1', \mathbf{B}_2', ..., \mathbf{B}_K')'$ be the vector of all K coefficient vectors of Model 1, where each has length p, and let $\mathbf{z}_{ij}(t) = (0', 0', ..., \mathbf{x}_{ij}(t)', 0', ..., 0')'$ be the vector of length pK for the j^{th} spell of the i^{th} individual where the j^{th} component of this vector contains the vector of covariates $\mathbf{x}_{ij}(t)$. Then, Model 1 can be expressed as

$$h_j(t \mid \mathbf{z}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{z}_{ij}'(t)\boldsymbol{\gamma}},$$

which has the general form of baseline hazards varying with spell order but a fixed coefficient vector. A test for constancy of the coefficients pertaining to each spell order, *i.e.*, $H_0 : \mathbf{B}_1 = \mathbf{B}_2 = ... \mathbf{B}_K$ is equivalent to testing $H_0 : \mathbf{C}\boldsymbol{\gamma} = 0$ where $\mathbf{C}$ is the $(K-1)p \times Kp$ matrix $\mathbf{C} = I_p \otimes [1_{K-1} \; -I_{K-1}]$. Given an estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ and an estimate $\hat{V}(\hat{\boldsymbol{\gamma}})$ of the covariance matrix of $\hat{\boldsymbol{\gamma}}$, obtained as described in Section 4 for Model 2, a Wald statistic may be calculated in order to test the hypothesis. If the hypothesis is not rejected, it may be concluded that a model with constant coefficients over spell order (but baseline hazard varying with spell order) appears to fit the data as well as a model where both baseline hazard and coefficients vary with spell order. Other measures for model adequacy should also be straightforward to develop under the design-based framework.

We also visually compared, for our example, coefficient standard error estimates obtained under the design-based approach (accounting for clustering at the PSU level and lower) and obtained under a modification of the robust model-based approach (accounting for clustering at the individual level and lower) for Models 1 and 2. We found only minor differences, which indicated no clustering effects above the individual level for these particular data. We also calculated standard error estimates assuming independence even between spells from the same person and again found only minor differences with those obtained from the design-based approach. It thus seems that, for this particular example, there is little inter-spell dependence. However, in general, we feel that a design-based approach guards against missing any unpostulated dependencies at the PSU level and lower in the variance estimates.

Acknowledgements

We are grateful to Normand Laniel and Xuelin Zhang for their useful comments to an earlier version of this manuscript. We also thank the associate editor and the referees for comments and suggestions improving greatly the readability of the manuscript.

References

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-291.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika,* 79, 139-147.

Blossfeld, H.-P., and Hamerle, A. (1989). Using Cox models to study multiepisode processes. *Sociological Methods and Research*, 17, 4, 432-448.

Clayton, D., and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society*, Series A, 1985, 148, 82-117.

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

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Hamerle, A. (1989). Multiple-spell regression models for duration data. *Applied Statistics*, 38, 1, 127-138.

Heckman, J., and Singer, B. (1982). Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*, (Eds., K. Land and A. Rogers), New York: Academic Press, 567-599.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55, 1, 13-22.

Kalbfleisch, J.D., and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Edition, New York: John Wiley & Sons, Inc.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.

Lavigne, M., and Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Working Paper, Statistics Canada, 75F0002M No. 98-05.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.

Lin, D.Y. (2000). On Fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., and Wei, L.J. (1989). The robust Inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.

Roberts, G., and Kovačević, M. (2001). New research problems in analysis of duration data arising from complexities of longitudinal surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada,* 111-116.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 23-34
Statistics Canada, Catalogue No. 12-001

Bayesian weight trimming for generalized linear regression models

Michael R. Elliott [1]

Abstract

In sample surveys where units have unequal probabilities of inclusion in the sample, associations between the probability of inclusion and the statistic of interest can induce bias. Weights equal to the inverse of the probability of inclusion are often used to counteract this bias. Highly disproportional sample designs have large weights, which can introduce undesirable variability in statistics such as the population mean estimator or population regression estimator. Weight trimming reduces large weights to a fixed cutpoint value and adjusts weights below this value to maintain the untrimmed weight sum, reducing variability at the cost of introducing some bias. Most standard approaches are ad-hoc in that they do not use the data to optimize bias-variance tradeoffs. Approaches described in the literature that are data-driven are a little more efficient than fully-weighted estimators. This paper develops Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. An application to estimate injury risk of children rear-seated in compact extended-cab pickup trucks using the Partners for Child Passenger Safety surveillance survey is considered.

Key Words: Sample survey; Sampling weights; Weight Winsorization; Bayesian population inference; Weight smoothing; Generalized linear mixed models.

1. Introduction

Analysis of data from samples with differential probabilities of inclusion typically use case weights equal to the inverse of the probability of inclusion to reduce or remove bias in the estimators of population quantities of interest. Replacing implicit means and totals in statistics with their case-weighted equivalents yields unbiased linear estimators and asymptotically unbiased non-linear estimators of population values (Binder 1983). Case weights may also incorporate non-response adjustments, which typically are equal to the inverse of the estimated probability of response (Gelman and Carlin 2002, Oh and Scheuren 1983), or calibration adjustments, which constrain case weights to equal known population totals, either jointly, as in poststratification or generalized regression estimation, or marginally, as in generalized raking estimation (Deville and Särndal 1992, Isaki and Fuller 1982).

There is little debate that sampling weights be utilized when considering descriptive statistics such as means and totals obtained from unequal probability-of-selection designs. However, when estimating "analytical" quantities (Cochran 1977, page 4) that focus on associations between, *e.g.*, risk factors and health outcomes estimated via linear and generalized linear models, the decision to use sampling weights is less definitive (*cf* Korn and Graubard 1999, pages 180-182). In a regression setting, discrepancies between weighted and unweighted regression slope estimators can occur either because the data model is misspecified or there is an association between the residual errors and/or the probability of inclusion (sampling is informative). When the data model is misspecified, one option is to improve the model specification. However, it may be difficult to determine the exact functional form; or it may be that the degree of misspecification is very modest but is magnified by the sample design; or it may be that an approximation to the true model is desired to simplify explanation (linearly approximating a quadratic trend). In the case of informative or non-ignorable sampling, design weights may be required to obtain consistent estimators of regression parameters (Korn and Graubard 1995). More formally, fully-weighted estimators of regression parameters are "pseudo-maximum likelihood" estimators (PMLEs) (Binder 1983, Pfeffermann 1993) in that they are "design consistent" for MLEs that would solve the score equations for the regression parameters under the assumed superpopulation regression model if we had observed data for the entire population. Design consistency implies that the difference between the population target quantity and the estimate derived from the sample tends to zero as the sample size and population size jointly increase, or that these differences will on average tend to 0 from repeated sampling of the population, where samples are selected in an identical fashion from $t \to \infty$ replicates of the population: see Särndal (1980) or Isaki and Fuller (1982). If observations are clustered, more care must be taken to develop design consistent estimators of PLMEs, although nested multi-stage designs allow for the census log-likelihood estimates to be approximated using weighted score equations if care is taken to account for the fact that the within-cluster sample sizes typically are small and remain so even if the number of clusters increases

1. Michael R. Elliott is Assistant Professor, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI. E-mail: mrelliot@umich.edu.

(Pfeffermann, Skinner, Holmes, Goldstein and Rabash 1998, Korn and Graubard 2003).

Although PMLEs are popular because of design consistency, this property is purchased at the cost of increased variance. This increase can overwhelm the reduction in bias, so that the MSE actually increases under a weighted analysis. This is particularly likely if a) the sample size is small, b) the differences in the inclusion probabilities are large, or c) the model is approximately correctly specified and the sampling is approximately noninformative. Perhaps the most common approach to dealing with this problem is *weight trimming* (Potter 1990, Kish 1992, Alexander, Dahl and Weidman 1997), in which weights larger than some value w_0 are fixed as w_0. Typically w_0 is chosen in an *ad hoc* manner - say 3 or 6 times the mean weight - without regard to whether the chosen cutpoint is optimal with respect to MSE. Thus bias is introduced to reduce variance, with the goal of an overall reduction in MSE.

Other design-based methods have been considered in the literature. Potter (1990) discusses systematic methods for choosing w_0, including weight distribution and MSE trimming procedures. The weight distribution technique assumes that the weights follow an inverted and scaled beta distribution; the parameters of the inverse-beta distribution are estimated by method-of-moment estimators, and weights from the upper tail of the distribution, say where $1 - F(w_i) < 0.01$, are trimmed to w_0 such that $1 - F(w_0) = 0.01$. The MSE trimming procedure determines the empirical MSE at trimming level w_t, where the trimmed weight $w_i^* = w_t I(w_i \geq w_t) + w_i I(w_i < w_t)$, $i = 1, ..., n$ under the assumption that the fully weighted estimate is unbiased for the true mean. In practice, one considers a variety of trimming levels $t = 1, ..., T$, where $t = 1$ corresponds to the unweighted data $(w_1 = \min_i(w_i))$ and $t = T$ to the fully-weighted data $(w_T = \max_i(w_i))$, and $\hat{\theta}_t$ is the value of the statistic using the trimmed weights at level t. The trimming level chosen is then given by $w_0 = w_{t^*}$, where $t^* = \text{argmin}_t(\text{MSE}_t)$ for $\text{MSE}_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\theta_t)$.

In the calibration literature, techniques have been developed that allow generalized poststratification or raking adjustments to be bounded to prevent the construction of extreme weights (Deville and Särndal 1992, Folsom and Singh 2000). Beaumont and Alavi (2004) extend this idea to develop estimators that focus on trimming large weights of highly influential or outlying observations. While these bounds trim extreme weights to a fixed cutpoint value, the choice of this cutpoint remains arbitrary.

An alternative approach to the direct weight trimming procedures has been developed in the Bayesian finite population inference literature (Elliott and Little 2000, Holt and Smith 1979, Ghosh and Meeden 1986, Little 1991, 1993, Lazzeroni and Little 1998, Rizzo 1992). These approaches account for unequal probabilities of inclusion by considering the case weights as stratifying variables within strata defined by the probability of inclusion. These "inclusion strata" may correspond to formal strata from a disproportional stratified sample design, or may be "pseudo-strata" based on collapsed or pooled weights derived from selection, poststratification, and/or non-response adjustments. Standard weighted estimates are then obtained when the weight stratum means of survey outcomes are treated as fixed effects, and trimming of the weights is achieved by treating the underlying weight stratum means as random effects. These methods allow for the possibility of "partially-weighted" data that uses the data itself to appropriately modulate the bias-variance tradeoff, and also allows estimation and inference from data collected under unequal probability-of-inclusion sample designs to be based on models common to other fields of statistical estimation and inference.

This paper extends these random-effects models, which we term "weight smoothing" models, to include estimation of population parameters in linear and generalized linear models. Section 2 briefly reviews Bayesian finite population inference, formalizes the concept of ignorable and non-ignorable sampling mechanisms, and develops the weight smoothing models for linear and generalized linear regression models in a fully Bayesian setting. Section 3 provides simulation results to consider the repeated sampling properties of the weight smoothing estimators of linear and logistic regression parameters in a disproportional-stratified sample design and compares them with standard design-based estimators. Section 4 illustrates the use of the weight smoothing estimators in an analysis of risk of injury to children in passenger vehicle crashes. Section 5 summarizes the results of the simulations and considers extensions to more complex sample designs.

2. Bayesian finite population inference

Let the population data for a population with $i = 1, ..., N$ units be given by $Y = (y_1, ..., y_N)$, with associated covariate vectors $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ and sampling indicator variable $I = (I_1, ..., I_N)$, where $I_i = 1$ if the i^{th} element is sampled and 0 otherwise. As in design-based population inference, Bayesian population inference focuses on population quantities of interest $Q(Y)$, such as population means $Q(Y) = \bar{Y}$ or population least-squares regression parameters $Q(Y, X) = \min_{B_0, B_1} \sum_{i=1}^{N} (y_i - B_0 - B_1 x_i)^2$. In contrast to design-based inference, but consistent with most other areas of statistics, one posits a model for the population data Y as a function of parameters θ:

$Y \sim f(Y|\theta)$. Inference about $Q(Y)$ is made based on the posterior predictive distribution of $p(Y_{\text{nob}} | Y_{\text{obs}}, I)$, where Y_{nob} consists of the elements of Y_i for which $I_i = 0$:

$$p(Y_{\text{nob}} | Y_{\text{obs}}, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}} \quad (1)$$

where $p(I | Y, \theta, \phi)$ models the inclusion indicator.

If we assume that ϕ and θ are *a priori* independent and if the distribution of sampling indicator I is independent of Y, the sampling design is said to be "unconfounded" or "noninformative"; if the distribution of I depends only on Y_{obs}, then the sampling mechanism is said to be "ignorable" (Rubin 1987), equivalent to the standard missing data terminology (the unobserved elements of the population can be thought of as missing by design). Under ignorable sampling designs, $p(\theta, \phi) = p(\theta)p(\phi)$ and $p(I | Y, \theta, \phi) = p(I | Y_{\text{obs}}, \phi)$, and thus (1) reduces to

$$\frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta dY_{\text{nob}}} = p(Y_{\text{nob}} | Y_{\text{obs}}), \quad (2)$$

allowing inference about $Q(Y)$ to be made without explicitly modeling the sampling inclusion parameter I (Ericson 1969, Holt and Smith 1979, Little 1993, Rubin 1987, Skinner, Holt and Smith 1989). Noninformative sample designs are a special case of ignorable sample designs, equivalent to missing completely at random mechanisms being a special case of missing at random mechanisms.

In the regression setting, where inference is desired about parameters that govern the distribution of Y conditional on fixed and known covariates X, (1) becomes

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}}$$

which reduces to

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X) = \frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta dY_{\text{nob}}}$$

if and only if I depends only on (Y_{obs}, X), of which dependence on X only is a special case. Thus if inference is desired about a regression parameter $Q(Y, X)$, then a noninformative or more generally ignorable sample design can allow inclusion probabilities to be a function of the fixed covariates.

2.1 Accommodating unequal probabilities of inclusion

Maintaining the ignorability assumption for the sampling mechanism often requires accounting for the sample design in both the likelihood and prior model structure. In the case of the unequal probability-of-inclusion sample designs, this can be accomplished by developing an index $h = 1, ..., H$ of the probability of inclusion (Little 1983, 1991); this could either be a one-to-one mapping of the case weight order statistics to their rankings, or a preliminary "pooling" of the case weights using, *e.g.*, the $100/H$ percentiles of the case weights. The data are then modeled by

$$y_{hi} | \theta_h \sim f(y_{hi}; \theta_h), \; i = 1, ..., N_h$$

for all elements in the h^{th} inclusion stratum, where θ_h allows for an interaction between the model parameter(s) θ and the inclusion stratum h. Putting a noninformative prior distribution on θ_h then reproduces a fully-weighted analysis with respect to the expectation of the posterior predictive distribution of $Q(Y)$.

To make this concrete, assume we are interested in estimating a population mean $Q(Y) = \bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ from a unequal probability-of-inclusion sample with a simple random sample within inclusion strata. Rewriting as $Q(Y) = \sum_h P_h \bar{Y}_h$ where $\bar{Y}_h = N_h^{-1} y_{hi}$ is the population inclusion stratum mean and $P_h = N_h/N$, we have

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h \{ n_h \bar{y}_{h,\text{obs}} + (N_h - n_h) E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) \}$$

where $\bar{Y}_h$ is decomposed into the observed inclusion stratum mean $\bar{y}_{h,\text{obs}} = n_h^{-1} \sum_{i=1}^{N_h} I_{hi} y_{hi}$ and the unobserved inclusion stratum mean $\bar{Y}_{h,\text{nob}} = (N_h - n_h)^{-1} \sum_{i=1}^{N_h} (1 - I_{hi}) y_{hi}$. If we assume

$$y_{hi} | \mu_h, \sigma_h^2 \overset{\text{ind}}{\sim} N(\mu_h, \sigma_h^2)$$

$$p(\mu_h, \sigma_h^2) \propto 1$$

then

$$E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) = E(E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) | Y_{\text{obs}}, \mu_h, \sigma_h^2) = E(\mu_h | Y_{\text{obs}}) = \bar{y}_{h,\text{obs}}.$$

and the posterior predictive mean of the population mean is given by the weighted sample mean:

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h N_h \bar{y}_{h,\text{obs}} = N^{-1} \sum_h \sum_{i=1}^{N_h} I_{hi} w_h y_{hi}$$

where $w_{hi} \equiv w_h = N_n/n_h$ for all the observed elements in inclusion stratum h. Further, the weighted mean will be the posterior predictive expectation of the population mean for any assumed distribution of Y as long as $E(y_{hi} \mid \mu_h) = \mu_h$. In contrast, a simple exchangeable model for the data

$$y_i \mid \mu, \sigma^2 \overset{\text{ind}}{\sim} N(\mu, \sigma^2)$$

$$p(\mu, \sigma^2) \propto 1$$

yields $E(\bar{Y} \mid Y_{\text{obs}}) = n^{-1}\sum_{i=1}^{N} I_i y_i$, the unweighted estimator of the mean, which may be badly biased if exchangeability fails to hold, as would be the case if there is an association between the probability of inclusion and Y.

2.2 Weight smoothing models

In its general form, our proposed "weight smoothing method" stratifies the data by the probability of inclusion and then uses a hierarchical model to effect trimming via shrinkage. A general description of such a model is given by

$$y_{hi} \mid \theta_h \sim f(y_{hi}; \theta_h) \tag{3}$$

$$\theta_h \mid M_h, \mu, R \sim N(\hat{y}_h, R), \hat{y}_h = g(M_h, \mu)$$

$$\mu, R \mid M_h \sim \Pi.$$

where $h = 1, ..., H$ indexes the probability of inclusion from the highest to the lowest probabilities, $g(M_h, \mu)$ is a function linking information M_h from the inclusion probability stratum and a smoothing parameter μ to the data distribution parameter θ_h indexed by the inclusion stratum, and Π is a flat or weakly informative hyperparameter distribution (Little 2004).

The particulars of the likelihood and prior specifications will depend on the population parameter of interest, the sample design, distributional assumptions about y, and efficiency-robustness tradeoffs. Positing an exchangeable model on the inclusion stratum means from the previous example yields (Lazzeroni and Little 1998, Elliott and Little 2000)

$$y_{hi} \mid \theta_h \overset{\text{ind}}{\sim} N(\theta_h, \sigma^2)$$

$$\theta_h \overset{\text{ind}}{\sim} N(\mu, \tau^2).$$

Assuming for the moment σ^2 and τ^2 known, we have

$$E(\bar{Y} \mid Y_{\text{obs}}) = N^{-1}\sum_h \{n_h \bar{y}_{h,\text{obs}} + (N_h - n_h)E(\mu_h \mid Y_{\text{obs}})\}$$

where $E(\mu_h \mid Y_{\text{obs}}) = w_h\bar{y}_h + (1 - w_h)\tilde{y}$ for $w_h = \tau^2 n_h/(\tau^2 n_h + \sigma^2)$ and $\tilde{y} = (\sum_h n_h/(n_h\tau^2 + \sigma^2))^{-1}\sum_h n_h/(n_h\tau^2 + \sigma^2)\bar{y}_h$. As $\tau^2 \to \infty$, $w_h \to 1$ so that $E(\bar{Y} \mid Y_{\text{obs}}) = \sum_h P_h \bar{y}_h$; thus a flat prior recovers the fully-weighted estimator, as we showed previously. On the other hand, as $\tau^2 \to 0$, $w_h \to 0$ so that $E(\mu_h \mid Y_{\text{obs}}) \to \tilde{y} \mid_{\tau^2=0} = \bar{y}$, the unweighted mean; thus the excluded units of the sample are estimated at the pooled mean since the model assumes that all y_{hi} are drawn from a common mean. Hence this weight smoothing model allows compromise between the design-consistent estimator which may be highly inefficient, and the unweighted estimator that is fully efficient under the strong assumption that the inclusion probability and mean of Y are independent. By assuming a weak hyperprior distribution on τ^2, the degree of compromise between the weighted and unweighted mean will be "data-driven," albeit under the modeling assumptions.

2.3 Weight smoothing for linear and generalized linear regression models

Generalized linear regression models (McCullagh and Nelder 1989) postulate a likelihood for y_i of the form

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \tag{4}$$

where $a_i(\phi)$ involves a known constant and a (nuisance) scale parameter ϕ, and the mean of y_i is related to a linear combination of fixed covariates $\boldsymbol{x}_i$ through a link function $g(\cdot)$: $E(y_i \mid \theta_i) = \mu_i$, where $g(\mu_i) = g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. We also have $\text{Var}(y_i \mid \theta_i) = a_i(\phi)V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$. The link is canonical if $\theta_i = \eta_i$, in which case $g'(\mu_i) = V^{-1}(\mu_i)$. Well-known examples are the normal distribution, where $a_i(\phi) = \sigma^2$ and the canonical link is $g(\mu_i) = \mu_i$; the binomial distribution, where $a_i(\phi) = n_i^{-1}$ and the canonical link is $g(\mu_i) = \log(\mu_i/(1 - \mu_i))$; and the Poisson distribution, where $a_i(\phi) = 1$ and the canonical link is $g(\mu_i) = \log(\mu_i)$.

Indexing the inclusion stratum by h, we have $g(E[y_{hi} \mid \boldsymbol{\beta}_h]) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h$. We assume a hierarchical model of the form

$$(\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_H^T)^T \mid \boldsymbol{\beta}^*, G \sim N_{Hp}(\boldsymbol{\beta}^*, G). \tag{5}$$

where $\boldsymbol{\beta}^*$ is an unknown vector of mean values for the regression coefficients and G is an unknown covariance matrix.

We consider the target population quantity of interest $\boldsymbol{B} = (B_1, ..., B_p)^T$ to be the slope that solves the population score equation $U_N(\boldsymbol{B}) = 0$ where

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i; \boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{(y_{hi} - g^{-1}(\mu_i(\boldsymbol{\beta})))x_{hi}}{V(\mu_{hi}(\boldsymbol{\beta}))g'(\mu_{hi}(\boldsymbol{\beta}))}. \quad (6)$$

Note that the quantity $\boldsymbol{B}$ such that $U(\boldsymbol{B}) = 0$ is always a meaningful population quantity of interest even if the model is misspecificed (*i.e.*, η_i is not exactly linear with respect to the covariates), since it is the linear approximation of $\boldsymbol{x}_i$ to $\eta_i = g(\mu_i)$. Under the model given by (4) and (5), a first-order approximation (assuming a negligible sampling fraction) to $E(\boldsymbol{B} \mid y, X)$ is given by $\hat{B}$ where

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{\boldsymbol{B}})))\boldsymbol{x}_{hi}}{V(\mu_{hi}(\hat{\boldsymbol{B}}))g'(\mu_{hi}(\hat{\boldsymbol{B}}))} = 0 \quad (7)$$

where $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}(\boldsymbol{x}_{hi}^T\hat{\boldsymbol{\beta}}_h)$, and $\hat{\boldsymbol{\beta}}_h = E(\boldsymbol{\beta}_h \mid y, X)$, as determined by the form of (5). (If N_h is unknown, it can be replaced with $\hat{N}_h = \sum_{i \in h} w_{hi}$, and the $\hat{N}_1, ..., \hat{N}_H$ treated as a multinomial distribution of size N parameterized by unknown inclusion stratum probabilities $q_1, ..., q_H$ with, *e.g.*, a Dirichlet prior.) Thus, in the example of linear regression, where $V(\mu_i) = \sigma^2$ and $g'(\mu_i) = 1$, (7) resolves to

$$\hat{\boldsymbol{B}} = E(\boldsymbol{B} \mid y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right]^{-1}\left[\sum_h W_h \left(\sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right)\hat{\boldsymbol{\beta}}_h\right]. \quad (8)$$

In the example of logistic regression, where $V(\mu_i) = \mu_i(1-\mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1-\mu_i)^{-1}$, $E(\boldsymbol{B} \mid y, X)$ is given by solving for the population regression parameters B_j, $j = 1, ..., p$

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij}B_j)}{1 + \exp(x_{hij}B_j)} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij}\hat{\beta}_{hj})}{1 + \exp(x_{hij}\hat{\beta}_{hj})}. \quad (9)$$

This can be accomplished via simple root-finding numerical methods such as Newton's Method.

We consider four forms of $\boldsymbol{\beta}^*$ and G in (5) in this paper:

1. Exchangeable Random Slope (XRS):
   $\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*)$ for all h, $G = I_H \otimes \sum$. (10)
2. Autoregressive Random Slope (ARS):
   $\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*)$ for all h,
   $G = A \otimes \sum$, $A_{jk} = \rho^{|j-k|}$, $j, k = 1, ..., H$.
3. Linear Random Slope (LRS):
   $\boldsymbol{\beta}_h^* = (\beta_{00}^* + \beta_{01}^* h, ..., \beta_{p0}^* + \beta_{p1}^* h)$,
   $G = I_H \otimes \Lambda$.
4. Nonparametric Random Slope (NPRS):
   $\boldsymbol{\beta}_h^* = (f_0(h), ..., f_p(h))$, $G = 0$.

$$\left\{ f_j : f_j^{v} \text{absolutely continuous, } v = 0,1, \int (f_j^{(2)}(u))^2 du < \infty, \min_{f_j} \sum_h (\beta_{hj}^* - f_j(h))^2 + \lambda_j \int (f_j^{(2)}(u))^2 du \right\}$$

where h again indexes the probability of inclusion, I_H is an $H \times H$ identity matrix, ρ is an autocorrelation parameter that controls the degree of shrinkage across the weight strata, $\sum$ is an unconstrained $p \times p$ covariance matrix, Λ is a $p \times p$ diagonal matrix, and $f_j(h)$ is a twice differentiable smooth function of h that minimizes the residual sum of squares plus a roughness penalty parameterized by λ_j (Wahba 1978, Hastie and Tibshirani 1990). Reformulating the NPRS model as in Wang (1998) we have

$$y_{hi} \mid \boldsymbol{\beta}_h \overset{\text{ind}}{\sim} N(\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h, \sigma^2)$$

$$\beta_{hj} = \beta_{j0}^* + \beta_{j1}^* h + \boldsymbol{\omega}_h \boldsymbol{u}_j$$

$$\boldsymbol{u}_j \overset{\text{ind}}{\sim} N_{H-1}(0, I\tau_j^2), \tau_j^2 = \sigma^2/(H\lambda_j)\ j = 0, ..., p$$

where $\boldsymbol{\omega}_h$ is the h^{th} row of Choleski decomposition of the cubic spline basis matrix Ω where $\Omega_{hk} = \int_0^1 ((h-1)/(H-1) - t)_+((k-1)/(H-1) - t)_+ dt$, $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$, $h, k = 1, ..., H$. The NPRS model can be extended into the generalized linear model form as in Lin and Zhang (1999), where the first-stage normality assumption is replaced with a link function that is linear in the covariates: $g(E(y_{hi} \mid \boldsymbol{\beta}_h)) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h$, for $g(\cdot)$ as in (4).

Assuming for the moment that the second stage parameters are known, we see that, in the case of the XRS model with normal data, as $|G| \to \infty$, sharing of information across inclusion strata ceases, and $\hat{\boldsymbol{\beta}}_h \approx (\boldsymbol{x}_h^T\boldsymbol{x}_h)^{-1}\boldsymbol{x}_h^T\boldsymbol{y}_h$, the regression estimator within the inclusion stratum. Replacing this into (8) yields $\hat{\boldsymbol{B}} \approx \hat{\boldsymbol{B}}^w$, the fully weighted estimator of the population slope. Similarly, as $|G| \to 0$, the within-inclusion-stratum slopes $\hat{\boldsymbol{\beta}}_h \approx \boldsymbol{\beta}^*$ the common prior slope, yielding $\hat{\boldsymbol{B}} \approx \boldsymbol{\beta}^*$ when replaced in (8), or $\hat{\boldsymbol{B}}^u$ if a non-informative hyperprior distribution is placed on $\boldsymbol{\beta}^*$ and its posterior mean obtained as $(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$. Empirical or fully Bayesian methods that allow the data to estimate the second stage parameters thus allow for data-driven "weight smoothing," compromising between the unweighted and fully-weighted estimators.

In practice, of course, the second-stage mean and variance components are usually not known; hence we

complete the model specification by postulating a hyperprior distribution for the second-stage parameters:

$$p(\phi, \boldsymbol{\beta}^*, G) \propto p(\boldsymbol{\zeta}).$$

Typically the hyperprior distribution $p(\boldsymbol{\zeta})$ is either weakly informative or non-informative. Gibbs sampling (Gelfand and Smith 1990; Gelman and Rubin 1992) can then be utilized to obtained draws from the full joint posterior of $(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \phi, G)^T \mid \boldsymbol{y}, X$. In the XRS model, we consider $p(\sigma, \boldsymbol{\beta}^*, \Sigma) \propto \sigma^{-2} \mid \Sigma \mid^{-(p+1/2)} \exp(-1/2 \text{ tr } \{r\Sigma^{-1}\})$, that is, non-informative prior distributions on the scale and prior mean parameters and an independent inverse-Wishart hyperprior distribution on the prior variance G centered at the identity matrix scaled by r with p degree of freedom. The same prior distribution is used for the ARS model, with the additional assumption that $\rho \overset{\text{ind}}{\sim} U(0,1)$ (non-negative autocorrelation between inclusion strata). In the LRS and NPRS models, $p(\sigma, \boldsymbol{\beta}^*, \Lambda) \propto \sigma^{-2}$ and $p(\sigma, \boldsymbol{\beta}^*, \boldsymbol{\tau}) \propto \sigma^{-2}$ (standard non-informative scale prior distribution and hyperprior distribution). Description of the conditional draws of the Gibbs sampler are available at http://www.sph.umich.edu/mrelliot/trim/meth2.pdf.

The degree of compromise is a function of the mean and variance structure of the chosen model. The XRS and ARS models assume exchangeable slope means; the ARS model is more flexible in that its variance structure allows units with more nearly equal probabilities of inclusion to be smoothed more heavily than units with very unequal probabilities of inclusion. The LRS model assumes an underlying linear trend in slopes, whereas the NPRS model assumes only an underlying trend smooth up to its second derivative. Note that, in the LRS and NPRS models, we assume *a priori* independence for the regression parameters associated with a given covariate, *i.e.*, $(\beta_{1j}, \ldots \beta_{Hj}) \perp (\beta_{1j'}, \ldots \beta_{Hj'})$, $j \neq j'$. This is because we model trends in these parameters across the inclusion stratum, and do not wish to "link up" these trends across the covariates.

Shrinkage will be greatest, corresponding to the most severe weight trimming, when the weight stratum slopes have little variability, or when the lowest probability-of-inclusion stratum are poorly estimated. Little shrinkage should occur when weight stratum slopes are precisely estimated and when they are systematically associated with their probability of inclusion. Based on Elliott and Little (2000), we would expect the XRS model to be the most efficient when large amounts of weight trimming are required to minimize MSE, but to be the most vulnerable to "overshrinking" when bias correction is most important. Increasing structure, particularly in the mean portion of the model as in LRS and NPRS, will provide more robust estimation in the sense that overshrinkage will occur only in near-pathological situations (*e.g.*, when mean trends are non-monotonic and highly discontinuous), and even then may only lead to slightly less bias correction than the data warrant. The price to be paid for this robustness, however, will be a reduction in efficiency relative to the exchangeable models.

3. Simulation results

Because we desire models that are simultaneously more efficient than design based estimators yet reasonably robust to model misspecification - and in general we feel that even Bayesian models should have good frequentist properties - we evaluate our proposed models in a repeated sampling context. We consider linear and logistic regression, under a misspecified model with a non-informative sampling design.

3.1 Linear regression

For the linear regression model in the presence of model misspecification, we generated population data as follows:

$$Y_i \mid X_i, \sigma^2 \sim N(\alpha X_i + \beta X_i^2, \sigma^2), \qquad (11)$$

$$X_i \sim U(0, 10), i = 1, \ldots, N = 20,000.$$

A noninformative, disproportionally stratified sampling scheme sampled elements as a function of X_i (I_i equals 1 if sampled and 0 otherwise):

$$h_i = \lceil X_i \rceil$$

$$P(I_i = 1 \mid h_i) = \pi_i \propto (1 + h_i/2.5) h_i$$

This created 10 strata, defined by the integer portions of the X_i values. Elements (Y_i, X_i) had $\approx 1/36^{\text{th}}$ the selection probability when $0 < X_i \leq 1$ as when $9 < X_i < 10$. We sampled $n = 500$ elements without replacement for each simulation. The object of the analysis is to obtain the population slope $B_1 = \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^N (X_i - \bar{X})^2$. We fixed $\alpha = \beta = 1$, yielding a positive bias in the estimate of B_1, and varied σ^2. The effect of model misspecification increases as $\sigma^2 \to 0$ as the bias of the estimators becomes larger relative to the variance, and conversely decreases as $\sigma^2 \to \infty$. We considered values of $\sigma^2 = 10^l$, $l = 1, \ldots, 5$; 200 simulations were generated for each value of σ^2.

Here and below we utilized an inverse-Wishart hyperprior distribution on the prior variance G, centered at the identity matrix with 2 degree of freedom.

In addition to the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) models

discussed in Section 2.3, we consider the standard designed-based (fully weighted) estimator, as well as trimmed weight and unweighted estimators. For the fully-weighted (FWT) estimator, we use the PMLE $\boldsymbol{B}_w = (X'WX)^{-1}X'W\boldsymbol{y}$ where, denoting by lower case the sampled elements ($I_i = 1$), $w_{hi} \equiv w_h$ for $h = 1, ..., H$, $i = 1, ..., n_h$, $W = \text{diag}(w_{hi})$, $\boldsymbol{x}_{hi} = (1\ x_{hi})'$, X_h contains the stacked rows of $\boldsymbol{x}'_{hi}$ and X contain the stacked matrices X_h. We obtained inference about $\hat{\boldsymbol{B}}_w$ via the standard Taylor Series approximation (Binder 1983):

$$\text{Var}(\hat{\boldsymbol{B}}_w) = \hat{S}_{XX}^{-1}\hat{\sum}(\hat{\boldsymbol{B}}_w)\hat{S}_{XX}^{-1}$$

where $\hat{S}$ is a design-consistent estimator of the population total $\sum_{i=1}^{N}\boldsymbol{x}'_i\boldsymbol{x}_i$ given by $X'WX$ and $\hat{\sum}(\hat{\boldsymbol{B}}_w)$ is a design-consistent estimate of the variance of the total $\sum_{i=1}^{N}\epsilon_i\boldsymbol{x}_i$ where $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{B}$ is the difference between the value of y_i and its estimated value under the true population slope $\boldsymbol{B}$: $\hat{\sum}(\hat{\boldsymbol{B}}_w) = \sum_h n_h/(n_h-1)\sum_{i=1}^{n_h}(\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)'(\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)$, where $\tilde{\boldsymbol{x}}_{hi} = w_{hi}e_{hi}\boldsymbol{x}_{hi}$ for $e_{hi} = y_{hi} - \boldsymbol{x}_{hi}\hat{\boldsymbol{B}}_w$. We also consider the trimmed (TWT) estimator obtained by replacing the weights w_{hi} with trimmed values w^t_{hi} that set the maximum normalized value to 3: $w^t_{hi} = N\tilde{w}^t_{hi}/\sum_{h=1}^{H}n_h\tilde{w}^t_h$, where $\tilde{w}^t_{hi} = \min(w_{hi}, 3N/n)$, and the unweighted (UNWT) estimator obtained by fixing $w_{hi} = N/n$ for all h, i.

Table 1 shows the relative bias, root mean square error (RMSE), and nominal 95% coverage for the three design-based and four model-based estimators of the population slope (second component of $\hat{\boldsymbol{B}}$) under consideration, as a function of the variance σ^2.

The fully-weighted estimator of the population slope is essentially design-unbiased under model misspecification; the unweighted and trimmed estimators are biased. The biases of the exchangeable and autoregressive models increase as variance increases, as these models trade unbiasedness of the fully-weighted estimator for the reduced variance of the unweighted estimator. The linear and nonparametric model were approximately unbiased.

The unweighted and trimmed weight estimators perform poorly with respect to MSE for small values of σ^2, where the bias due to model misspecification is critical, and well for larger values of σ^2, where the instability of the fully-weighted estimator is more important than bias reduction. The exchangeable model-based estimator has good RMSE properties for small and large values of σ^2, with MSE reductions of over 30%, but oversmooths for intermediate degrees of model specification. The autoregressive model performance equals that of the exchangeable model for small and large values of σ^2, but is largely protected against the oversmoothing of the exchangeable models at intermediate levels. The linear and nonparametric models essentially dominated the fully weighted estimators with respect to MSE under all of the simulations considered, although MSE reductions were only on the order of 10%.

The unweighted and trimmed estimators have poor coverage except when model misspecification in nearly absent. The failure of the bias-variance tradeoff for the exchangeable estimator in the presence of model misspecification is evident in the poor coverage of the estimator for intermediate values of σ^2; this effect is ameliorated, but not completely removed, for the autoregressive estimator. The linear and non-parametric estimators have good coverage when model misspecification is less important but undercover to some degree when model misspecification is more important.

Table 1
Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population linear regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface

	Relative bias (%) Variance $\log_{10}$					RMSE relative to FWT Variance $\log_{10}$					True Coverage Variance $\log_{10}$				
Estimator	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
UNWT	21.5	21.8	22.2	20.8	22.3	12.1	4.57	1.76	**0.75**	**0.67**	0	0	6	78	92
FWT	0.0	0.1	1.4	1.6	-0.2	1	1	1	1	1	94	95	96	94	96
TWT	8.3	8.4	9.6	8.8	7.8	4.74	1.88	1.02	**0.71**	**0.75**	0	13	78	95	96
XRS	0.2	2.2	11.4	15.1	18.3	1.00	1.17	1.18	**0.73**	**0.68**	87	86	64	91	96
ARS	0.1	1.4	9.6	14.5	17.4	1.00	1.03	1.11	**0.74**	**0.69**	87	89	78	90	96
LRS	-0.2	-0.4	1.1	1.6	-0.3	**0.99**	**0.91**	**0.91**	**0.91**	**0.93**	85	91	96	95	94
NPRS	-0.1	-0.3	0.9	1.5	-0.4	**0.89**	**0.90**	**0.95**	**0.90**	**0.95**	86	92	96	94	94

3.2 Logistic regression

For the logistic regression model, we generated population data as follows:

$$P(Y_i = 1 \mid X_i) \sim B(\text{expit}(3.25 - 0.75X_i + \gamma X_i^2)), \quad (12)$$

$$X_i \sim U(0, 10), i = 1, ..., N = 20,000.$$

where $B(p)$ is a Bernoulli distribution with probability of "success" p, $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. The object of the analysis is to obtain the logistic population regression slope, defined as the value B_1 in the equation $\sum_i^N (y_i - \text{expit}(B_0 + B_1 x_i))\begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$. An unequal probability of selection sampling scheme was implemented as described in the linear regression simulations. We consider values of $\gamma = 0, 0.0158, 0.0273, 0.0368, 0.0454,$ corresponding to curvature measures of $K = 0, 0.02, 0.04, 0.06, 0.08$ at the midpoint 5 of the support for X, where $K(X; \gamma) = |2\gamma/[1 + (2\gamma X - 0.75)^2]^{3/2}|$; 200 simulations were generated for each value of γ. As in the linear regression simulations, elements were sampled without replacement with probability proportional to $(1 + h_i/2.5)h_i$; a total of 1,000 elements were sampled for each simulation. We again considered the PMLE-based the fully weighted (FTW), unweighted (UNWT), and trimmed weight estimator (TWT), along with the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators. Inference about the PMLE estimators is obtained via Taylor Series approximations (Binder 1983), as discussed in the previous section.

Table 2 shows the relative bias, RMSE relative to the RMSE of the fully-weighted estimator, and true coverage of the nominal 95% CIs or PPIs associated with each of the seven estimators of the population slope (B) for different values of curvature K, corresponding to increased degrees of misspecification.

The undersampling of small values of X meant that the maximum likelihood estimator of B in the model misspecification setting was unbiased for $K = 0$ and biased downward for $K = 0.02, 0.04, 0.06,$ and 0.08 unless the sample design was accounted for. The trimmed estimator's bias was intermediate between the unweighted and fully weighted estimator. The exchangeable estimator's bias was between the trimmed weight estimator and fully weighted estimator; the autoregressive estimator's bias between that of the exchangeable and fully weighted estimator; while the linear and nonparametric estimators were essentially unbiased.

The unweighted estimator had substantially improved MSE (40% reduction) when the linear slope model was approximately correctly specified, but failed with moderate to large degree of misspecification. The trimmed weight, autoregressive, and nonparametric estimators all dominated the standard fully-weighted estimator, and the exchangeable and linear estimators nearly so, over the range of simulations considered. The crude trimming estimator yielded up to 30% reduction in MSE, the nonparametric, exchangeable and autoregressive estimators reductions of up to 20-25%, and the linear estimator reductions of only 10% or less.

The unweighted estimator had poor coverage except when the linear slope model was correctly specified, or nearly so. The model-based estimators had generally good coverage properties when the linear model was correctly specified, with slight reductions in coverage when curvature was substantial.

Table 2
Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population logistic regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface

	Relative bias (%)					RMSE relative to FWT					True Coverage				
	Curvature K					Curvature K					Curvature K				
Estimator	0	0.02	0.04	0.06	0.08	0	0.02	0.04	0.06	0.08	0	0.02	0.04	0.06	0.08
UNWT	1.0	-4.9	-11.9	-21.6	-34.6	**0.57**	**0.73**	**0.88**	1.19	1.61	96	89	66	32	17
FWT	1.1	2.2	1.3	-0.3	1.6	1	1	1	1	1	95	94	90	94	94
TWT	0.5	-1.0	-3.5	-7.2	-12.1	**0.70**	**0.77**	**0.77**	**0.78**	**0.95**	98	97	94	84	92
XRS	1.3	-0.8	-1.9	-5.6	-8.7	**0.75**	**0.82**	**0.85**	**0.88**	1.02	97	94	92	91	90
ARS	1.3	-0.5	-2.2	-4.8	-7.5	**0.78**	**0.85**	**0.84**	**0.84**	**0.95**	94	92	90	92	90
LRS	0.8	1.7	1.5	-0.4	1.1	**0.89**	**0.97**	**0.94**	**0.91**	1.02	95	91	88	92	89
NPRS	0.3	1.5	1.1	0.9	0.5	**0.87**	**0.88**	**0.87**	**0.80**	**0.90**	95	92	88	94	96

4. Application: Estimation of injuries to children in compact extended-cab pickup trucks

The Partners for Child Passenger Safety dataset consists of the disproportionate, known-probability sample from all State Farm claims since December 1998 involving at least one child occupant $\leq$ 15 years of age riding in a model year 1990 or newer State Farm-insured vehicle (Durbin, Bhatia, Holmes, Shaw, Werner, Sorenson and Winston 2001). Because injuries, and especially "consequential" injuries defined as facial lacerations or other injuries rated 2 or more on the Abbreviated Injury Scale (AIS) (Association for the Advancement of Automotive Medicine 1990), are relatively rare even among children in the population of crash-related vehicle damage claims, a disproportional stratified cluster sample is used to select vehicles (the unit of sampling) for the conduct of a telephone survey with the driver. Vehicles containing children who received medical treatment following the crash were over-sampled so that the majority of injured children would be selected while maintaining the representativeness of the overall population. (Medical treatment is defined as treatment by paramedics, treatment at a physician's office or emergency room, or hospitalization.) If a vehicle was sampled, all child occupants in that vehicle were included in the survey. Drivers of sampled vehicles were contacted by phone and, if medical treatment had been received by a passenger, screened via an abbreviated survey to verify the presence of at least one child occupant with an injury. All vehicles with at least one child who screened positive for injury and a 10% random sample of vehicles in which all child occupants who were reported to receive medical treatment but screened negative for injury were selected for a full interview; a 2% (later 2.5%) sample of crashes where no medical treatment was received were also selected. Because the treatment stratification is imperfectly associated with risk of injury (more than 15% of the population with consequential injuries are estimated to be in the lowest probability-of-selection category and nearly 20% of those without consequential injuries are in the highest probability-of-selection category), the sampling design is informative, with unweighted odds ratios biased toward the null (Korn and Graubard 1995). In addition, the weights for this dataset are quite variable: $1 \leq w_i \leq 50$, where 9% of the weights have normalized values greater than 3.

Winston, Kallan, Elliott, Menon and Durbin (2002) determined that children rear-seated in compacted extended cab pickups are at greater risk of consequential injuries than children rear-seated in other vehicles. However, quantifying degree of excess risk, and thus the size of the public health problem, was problematic. The unweighted odds ratio (OR) of consequential injury for children riding in compacted extended cab pickups versus other vehicles was 3.54 (95% CI 2.01, 6.23), versus the fully-weighted estimator of 11.32 (95% CI 2.67, 48.03). Because both injury risk and compacted extended cab pickup use were associated with child age, crash severity (passenger compartment intrusion and drivability), direction of impact, and vehicle weight, a multivariate logistic regression model that adjusted for these factors was also considered. The unweighted and fully-weighted adjusted ORs for injury risk in rear seated children in compacted extended cab pickups versus other vehicles are 3.50 (95% CI 1.88, 6.53) and 14.56 (95% CI 3.45, 61.40) respectively. Utilizing the unweighted estimator was problematic because of bias toward the null induced by the informative sample design; however the fully weighted estimator appeared to be highly unstable. in part because of the presence of one consequential-injured child in the compact extended cab pickups had a very low probability of selection (0.025). In Winston *et al.* (2002), this child was removed before conducting the analysis.

Table 3 shows the results for the unadjusted and adjusted odds ratios of consequential injury risk using the unweighted, fully-weighted, and trimmed-weight design-based estimators, along with the model-based exchangeable, autoregressive, and linear regression slope models. (Results for the model-based estimators from 250,000 draws of a single chain after a 50,000 draw burn-in; convergence was assessed via Geweke (1992).) For the XRS and ARS models, $p(\sum) \sim \text{INVERSE-WISHART}(p, 0.1I)$, where $p = 2$ for the unadjusted model and $p = 13$ for the adjusted model. In the unadjusted results, the XRS and ARS estimators are intermediate between the unweighted and fully-weighted estimator, while the linear and nonparametric estimators tends to track the fully-weighted estimator. In the adjusted analysis, all three model-based estimators are intermediate between the unweighted and fully-weighted estimators, with the XRS estimator closest to the unweighted estimator and the LRS estimator closest to the fully-weighted estimator. Based on the results of the simulation, it appears that the ARS estimator, which suggest relative risks of injury on the order of 7 for children in compact extended cab pickups relative to other vehicles, may be a better estimator of relative risk than either the unweighted or fully weighted estimator. (As a "sanity check" of sorts, we note that an additional two years of data, not available at the time of Winston *et al.* (2002), which included an additional 4,091 rear-seated children in passenger vehicles [44 in compact extended-cab pickup trucks], provided a fully-weighted unadjusted odds ratio for injury for children in compact-extended cab pickups of 6.3, and an adjusted OR of 7.0.)

Table 3
Estimated odds ratio of injury for children rear-seated in compacted extended cab pickups ($n = 60$) versus rear-seated in other vehicles ($n = 8,060$), using unweighted (UNWT), fully-weighted (FWT), weights trimmed to a normalized value of 3 (TWT), exchangeable random slope (XRS), autoregression random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators; unadjusted and adjusted for child age, crash severity, direction of impact, and vehicle weight. Point estimates for XRS, ARS, and LRS models from posterior median. 95% confidence interval or posterior predictive interval in subscript. Data from Partners for Child Passenger Safety

	UNWT	FWT	TWT	
Unadj.	$3.54_{(2.01, 6.23)}$	$11.32_{(2.67, 48.02)}$	$9.15_{(2.65, 31.57)}$	
Adj.	$3.50_{(1.88, 6.53)}$	$14.56_{(3.45, 61.40)}$	$10.99_{(2.97, 34.64)}$	
	XRS	**ARS**	**LRS**	**NPRS**
Unadj.	$6.70_{(2.51, 20.92)}$	$6.69_{(2.64, 21.05)}$	$11.17_{(3.21, 24.94)}$	$10.34_{(3.27, 24.62)}$
Adj.	$4.45_{(2.39, 8.67)}$	$6.67_{(3.56, 11.94)}$	$11.87_{(3.33, 36.93)}$	$10.23_{(3.02, 37.93)}$

5. Discussion

The models discussed in this paper generalize the work of Lazzeroni and Little (1998) and Elliott and Little (2000), where population inference was restricted to population means under Gaussian distributional assumptions. Viewing weighting as an interaction between inclusion probability and model parameters opens up an alternative paradigm for weight trimming as a random effects model that smoothes model parameters of interest across inclusion classes. Models with exchangeable mean structures offer the largest degree of shrinkage or trimming but the most sensitivity to model misspecification; models with highly structured means are potentially less efficient but are more robust to model misspecification. This robustness property may be particularly important in light of the fact that elements of the large inclusion strata provide the largest degree of potential variance reduction in the model-based setting but are also subject to the largest degree of model bias and variance due to extrapolation.

We consider simulations under varying degrees of model misspecification and informative sampling for both linear and logistic regression models. The linear and non-parametric smoothing models nearly dominated fully-weighted estimators with respect to squared error loss in the simulations considered. The exchangeable model showed some tendency to oversmooth, favoring variance reduction over bias correction, especially in the linear regression setting. All of the weight smoothing estimators tended to have less than nominal coverage when models were highly misspecified, although in no case was the nominal coverage catastrophically low. The autoregressive smoothing model, which allows for differential degrees of local smoothing across weight strata, appeared to provide non-trivial increases in efficiency with limited risk of severe over-smoothing or undercoverage.

Applying the methods to the Partners for Child Passenger Safety data to determine the excess risk of injury in a crash to rear-seated children in compacted extended-cab pickups relative to rear-seated children in other passenger vehicles, it appears that the decision in Winston *et al.* (2002) to eliminate a low probability-of-selection child from the analysis to stabilize the estimates was indeed conservative. Indeed, the ARS estimator, favored by MSE in simulations, suggests an adjusted excess risk of 6.7 with a 95% PPI of (3.6, 11.9), versus the 14.6 with 95% CI of (3.4, 61.4) of the fully-weighted estimator.

Although this paper utilizes a fully Bayesian approach to inference about the posterior predictive distribution of the population regression slope, empirical Bayes (EB) estimates can also be obtained via ML or REML estimation using standard linear or generalized linear mixed model methods. In the Gaussian setting, the EB estimates of G and σ^2 can be "plugged into" the closed-form expressions for $E(\boldsymbol{B} \mid y, X)$ and $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. The general exponential setting is more problematic. The plug-in estimates can be used to determine $E(\boldsymbol{B} \mid y, X)$ via root-finding methods; the lack of a closed form for $E(\boldsymbol{B} \mid y, X)$ makes it difficult to obtain model-based Empirical Bayes estimators for $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. Also, standard Empirical Bayes estimators do not account for the uncertainty in the estimation of G.

We also note that, while computation of the actual trimming values of the case weights is unnecessary in this approach, it is possible to determine the revised design weights implied by the shrinkage. In the linear model setting, these can be obtained via a iterative application of a calibration weighting scheme such as generalized regression estimators or GREG (Deville and Särndal 1992). The

general exponential setting required embedding the calibration weight algorithm within the iterative reweighted least squares (IRWLS) algorithm used to fit a generalized linear model.

When sampling weights are used to account for misspecification of the mean in a regression setting, it could be argued that the correct approach is to correctly specify the mean to eliminate discrepancies between the fully-weighted and unweighted estimates of the regression parameters. However, perfect specification is an unattainable goal, and even good approximations might be highly biased if case weights are ignored when the sampling probabilities are highly variable. In the informative sampling setting, it may be impossible to determine whether discrepancies between weighted and unweighted estimates are due to model misspecification or to the sample design itself. Finally, even misspecified regression models have the attractive feature in the finite population setting of yielding a unique target population quantity. Consequently accounting for the probability of inclusion in linear and generalized linear model settings continues to be advised, and methods that balance between a low-bias, high variance fully-weighted analysis and a high bias, low variance unweighted analysis remain useful.

The methods discussed in this paper show the promise of adapting model-based methods to attack problems in survey data analysis. Our goal is not to develop a single hierarchical Bayesian model finely-tuned to a specific or question dataset at hand, but to develop robust yet efficient methods that can be applied in a fast-paced "automated" setting that many applied survey research analysts must sometimes work. Although computationally intensive, the methods considered are applications or extensions of the existing random-effect model "toolbox," and can either be implemented in existing statistical packages or executed with relatively simple MCMC methods. Our approach retains a design-based flavor in that we attempt to develop "automated" Bayesian model-based estimation techniques that yield robust inference in a repeated-sampling setting when the model itself is misspecified. However, because these models rely on stratifying the data by probability of selection as a prelude to using pooling or shrinkage techniques to induce data-driven weight trimming, there is a natural correspondence between this methodology and (post)stratified sample designs in which strata correspond to unequal probabilities of inclusion. Developing methods that accommodate a more general class of complex sample designs that include single or multi-stage cluster samples and/or strata that "cross" the inclusion strata remains an important area for future work.

Acknowledgements

The author thanks Roderick J.A. Little, along with the Editor, Associate Editor, and two anonymous reviewers, for their review and comments. The author also thanks Drs. Dennis Durbin and Flaura Winston of the Partners for Child Passenger Safety project for their assistance, as well as State Farm Insurance Companies for their support of the Partners for Child Passenger Safety project. This research was supported by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01.

References

Alexander, C.H., Dahl, S. and Weidman, L. (1997). Making estimates from the American Community Survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 2000, 88-97.

Association for the Advancement of Automotive Medicine (1990). *The Abbreviated Injury Scale*, *1990 Revision*. Association for the Advancement of Automotive Medicine, Des Plaines, Illinois.

Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195-208.

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

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

Durbin, D.R., Bhatia, E., Holmes, J.H., Shaw, K.N., Werner, J.V., Sorenson, W. and Winston, F.K. (2001). Partners for child passenger safety: A unique child-specific crash surveillance system. *Accident Analysis and Prevention*, 33, 407-412.

Elliott, M.R., and Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Ericson, W.A. (1969). Subjective bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 31, 195-234.

Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2000, 598-603.

Gelfand, A.E., and Smith, A.M.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389-409.

Gelman, A., and Carlin, J.B. (2002). Poststratification and weighting adjustments. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), 289-302.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4*, *Proceedings of the Fourth Valencia International Meeting*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), 89-193.

Ghosh, M., and Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*, London: Chapman and Hall.

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

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

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

Korn, E.L., and Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society*, Series B, 65, 175-190.

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

Lin, X., and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, Series B, 61, 381-400.

Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

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

Little, R.J.A. (2004). To model or not model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition. CRC Press: Boca Raton, Florida.

Oh, H.L., and Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse, *Incomplete Data in Sample Surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184.

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

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

Potter, F. (1990). A study of procedures to identify and trim extreme sample weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1990, 225-230.

Rizzo, L. (1992). Conditionally consistent estimators using only probabilities of selection in complex sample surveys. *Journal of the American Statistical Association*, 87, 1166-1173.

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

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

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

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, Series B, 40, 364-372.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93, 341-348.

Winston, F.K., Kallan, M.K., Elliott, M.R., Menon, R.A. and Durbin, D.R. (2002). Risk of injury to child passengers in compact extended pick-up trucks. *Journal of the American Medical Association*, 287, 1147-1152.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 35-44
Statistics Canada, Catalogue No. 12-001

Semiparametric model-assisted estimation for natural resource surveys

F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson and M. Giovanna Ranalli [1]

Abstract

Auxiliary information is often used to improve the precision of survey estimators of finite population means and totals through ratio or linear regression estimation techniques. Resulting estimators have good theoretical and practical properties, including invariance, calibration and design consistency. However, it is not always clear that ratio or linear models are good approximations to the true relationship between the auxiliary variables and the variable of interest in the survey, resulting in efficiency loss when the model is not appropriate. In this article, we explain how regression estimation can be extended to incorporate semiparametric regression models, in both simple and more complicated designs. While maintaining the good theoretical and practical properties of the linear models, semiparametric models are better able to capture complicated relationships between variables. This often results in substantial gains in efficiency. The applicability of the approach for complex designs using multiple types of auxiliary variables will be illustrated by estimating several acidification-related characteristics for a survey of lakes in the Northeastern US.

Key Words: Regression estimation; Smoothing; Kernel regression; Lake chemistry.

1. Introduction

Post-stratification, calibration and regression estimation are different design-based approaches that can be used to improve the precision of estimators when auxiliary information is available at the estimation stage. *Model-assisted estimation* (Särndal, Swensson and Wretman 1992) provides a convenient framework in which to develop these and related survey estimators. Under that framework, a superpopulation model describes the relationship between the variable of interest and the auxiliary variables. This model is then used to construct sample-based estimators that have improved precision when the model is correct, but maintain key design properties such as consistency and an estimable variance when the model is incorrect.

Until recently, the superpopulation models used in this context were formulated as parametric models, most often ratio or linear models. While reasonable in many practical applications, there are also many situations in which such relatively simple models are not good representations of the relationship between the variable of interest and the auxiliary variables. In Breidt and Opsomer (2000), a nonparametric model-assisted estimator was proposed based on local polynomial regression, which generalized the well-established parametric regression estimators. With this estimator, the superpopulation is no longer required to follow a pre-specified parametric shape. Instead, the relationship between the the variable(s) of interest in the survey and the auxiliary variable is required to be smooth (continuous), but is otherwise left completely unspecified.

In the current paper, we formally extend the theory of Breidt and Opsomer (2000) to the semiparametric regression context, in which some variables are incorporated linearly, and others are incorporated through smooth additive terms. This extension makes their results more useful in practice, since auxiliary information is very often multidimensional in nature, and almost always contains categorical variables that need to enter the regression model parametrically (through the use of indicator variables). An illustration of this is provided by a survey of lakes in the Northeastern states of the U.S. conducted by the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency. In that survey, 334 lakes were sampled from a population of 21,026 lakes between 1991 and 1996. We will apply the semiparametric model-assisted estimator to produce estimates of the mean and distribution function of the *acid neutralizing capacity* and other chemistry variables of interest. In this application, we will include in the model both categorical and continuous variables linearly and a continuous variable as a smooth additive term.

In Opsomer, Breidt, Moisen and Kauermann (2007), the nonparametric model-assisted estimation principle was extended to generalized additive models (GAMs) and applied in an interaction model for the estimation of variables from Forest Inventory and Analysis surveys. While GAMs also contained a mixture of categorical (parametric) and nonparametric terms, a complete theoretical development is not possible in the case of GAMs, and was therefore not provided there. The semiparametric model considered in this article can be viewed as a special case of a GAM with an identity link function. Unlike the

1. F. Jay Breidt, Department of Statistics, Colorado State University, Fort Collins CO 80523, U.S.A.; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-Mail: jopsomer@iastate.edu; Alicia A. Johnson, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis MN 55455, U.S.A.; M. Giovanna Ranalli, Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli, 06123 Perugia, Italy.

"general" GAM, the semiparametric model allows for formal derivation of the statistical properties of the model-assisted estimator.

The remainder of the article is structured as follows. In Section 2, the semiparametric model-assisted estimator is defined. Section 3 states and proves the design properties of the estimator. Section 4 describes the application of semiparametric model-assisted estimation to the Northeastern Lakes data. Section 5 provides a conclusion.

2. Semiparametric model-assisted estimation

We begin by considering the superpopulation model with a single univariate nonparametric term and a parametric component; extension to several nonparametric terms is addressed in Section 3.2. The parametric component can be composed of an arbitrary number of linear terms. This model is the semiparametric model studied by Speckman (1988), among others. This superpopulation model, which we denote by ξ, can be written down as

$$\begin{aligned} \mathrm{E}_\xi(y_k) &= g(x_k,\ \boldsymbol{z}_k) = m(x_k) + \boldsymbol{z}_k\boldsymbol{\beta} \\ \mathrm{Var}_\xi(y_k) &= v(x_k,\ \boldsymbol{z}_k) \end{aligned} \qquad (1)$$

with x_k a continuous auxiliary variable to be modelled nonparametrically and $\boldsymbol{z}_k = (z_{1k}, ..., z_{Dk})$ a vector of D categorical or continuous auxiliary variables that are parametrically specified. The functions $m(\cdot)$ and $v(\cdot, \cdot)$ and the parameter vector $\boldsymbol{\beta}$ are unknown. For identifiability purposes, we will assume that the vector $\boldsymbol{z}_k$ contains an intercept term, and that the function $m(\cdot)$ is centered around 0 with respect to the distribution of the x_k. We will derive the model-assisted estimator that uses model (1) by first defining population-level estimators for the unknown functions and parameters, and then constructing sample-based estimators. This is the same approach used for the parametric case in Särndal *et al.* (1992, Chapter 6).

Let $U = \{1, 2, ..., N\}$ represent the ordered labels for a finite population of interest. As the population estimator for $g(x_k, \boldsymbol{z}_k)$, we will use the *backfitting estimator* described in Opsomer and Ruppert (1999). We first introduce the required notation. Let $K(\cdot)$ represent a kernel function used to define the neighborhoods in which the local polynomials will be fitted (assumptions on K are specified in the Appendix). The population *smoother vector* for local polynomial regression of degree p at x_k is defined as

$$\boldsymbol{s}_{Uk}^T = \boldsymbol{e}_1^T (\boldsymbol{X}_{Uk}^T \boldsymbol{W}_{Uk} \boldsymbol{X}_{Uk})^{-1} \boldsymbol{X}_{Uk}^T \boldsymbol{W}_{Uk}$$

with $\boldsymbol{e}_1$ a vector of length $p+1$ with a 1 in the first position and 0s elsewhere, $\boldsymbol{W}_{Uk} = \mathrm{diag}\{h^{-1}K((x_1 - x_k)/h), ..., h^{-1}K((x_N - x_k)/h)\}$ and

$$\boldsymbol{X}_{Uk} = \begin{bmatrix} 1 & x_1 - x_k & \dots & (x_1 - x_k)^p \\ \vdots & & \ddots & \vdots \\ 1 & x_N - x_k & \dots & (x_N - x_k)^p \end{bmatrix}.$$

The smoother $\boldsymbol{s}_{Uk}$ can be applied to the vector $\boldsymbol{Y}_U = (y_1, ..., y_N)^T$ to produce the nonparametric regression fit with respect to the variable x at observation x_k. It can also be applied to any of the columns of $\boldsymbol{Z}_U = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_N^T)^T$ to smooth those with respect to x. This will be done in the derivation of the properties of the semiparametric estimator (Section 3).

In addition to the smoother vector at x_k, $\boldsymbol{s}_{Uk}^T$, we also need to define the *smoother matrix* at all the observation points $x_1, ..., x_N$,

$$\boldsymbol{S}_U = \begin{bmatrix} \boldsymbol{s}_{U1}^T \\ \vdots \\ \boldsymbol{s}_{UN}^T \end{bmatrix},$$

and the *centered smoother matrix* $\boldsymbol{S}_U^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/N)\boldsymbol{S}_U$. When the smoother matrix is applied to $\boldsymbol{Y}_U$, it produces the vector of nonparametric regression fits at all the observation points. The centered smoother matrix $\boldsymbol{S}_U^*$ produces centered fits, *i.e.*, the overall mean of the fitted values is subtracted from each fitted value. The centering is used to maintain identifiability of the estimators, as explained in Opsomer and Ruppert (1999).

For any observation x_k, a possible estimator of $m(x_k)$ could be defined as $\boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U$, with or without a centering adjustment. This estimator would generally be poor, since it does not take into account the fact that the y_k contain a parametric component that depends on the $\boldsymbol{z}_k$. A more efficient estimator is provided by jointly estimating both $m(\cdot)$ and $\boldsymbol{\beta}$, as is done by the following set of estimators

$$\begin{aligned} \boldsymbol{B} &= (\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Z}_U)^{-1}\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Y}_U \\ m_k &= \boldsymbol{s}_{Uk}^T(\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}) \quad k = 1, ..., N. \end{aligned} \qquad (2)$$

In these estimators, $\boldsymbol{B}$ is calculated first, and then the "residual vector" $\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}$ is smoothed with respect to x. The estimators in (2) are identical to the *backfitting estimators* for additive models described in Hastie and Tibshirani (1990) and implemented in `gam` in S-Plus, R or SAS. As a population estimator for $\mathrm{E}_\xi(y_k) = g(x \mid k, \boldsymbol{z}_k)$, we use

$$g_k = m_k + \boldsymbol{z}_k\boldsymbol{B}.$$

We now explain how to construct a model-assisted estimator based on the semiparametric regression approach. Let $A \subset U$ be a sample of size n drawn from U according to sampling design $p(A)$ with one-way and two-way

inclusion probabilities $\pi_k = \sum_{A \ni k} p(A)$, $\pi_{kl} = \sum_{A \ni k,l} p(A)$, respectively. If the g_k, $k = 1, ..., N$ were available, it would be possible to construct a *difference estimator* for the population mean of the y_k, $\bar{y}_N = \sum_U y_k / N$, as

$$\hat{y}_{\text{dif}} = \frac{1}{N}\sum_U g_k + \frac{1}{N}\sum_A \frac{y_k - g_k}{\pi_k}, \quad (3)$$

which is design unbiased and has design variance

$$\text{Var}_p(\hat{y}_{\text{dif}}) = \frac{1}{N^2}\sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l}$$

(Särndal *et al.* 1992, page 221). The design variance is small if the deviations between y_k and g_k are small. This estimator is not feasible, since it requires knowledge of all the x_k, $\boldsymbol{z}_k$ and y_k for the population to calculate. Instead, we will construct a feasible estimator by replacing the g_k by sample-based estimators. The sample-based estimators corresponding to the population estimators in (2) are constructed as follows. The design-weighted local polynomial smoother vector is

$$\boldsymbol{s}_{Ak}^{0T} = \boldsymbol{e}_1^T (\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak})^{-1} \boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak}, \quad (4)$$

with $\boldsymbol{X}_{Ak}$ containing the rows of $\boldsymbol{X}_{Uk}$ corresponding to the $k \in A$ and

$$\boldsymbol{W}_{Ak} = \text{diag}\left\{\frac{1}{\pi_j h} K\left(\frac{x_j - x_k}{h}\right) : j \in A\right\}.$$

The matrix $\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak}$ in (4) will be singular if, for some sample A, there are less than $p+1$ observations in the support of the kernel at some x_k. This issue can be avoided in practice by selecting a bandwidth large enough to make that matrix invertible. However, this situation cannot be excluded in general and we need an estimator that exists for every sample A for the theoretical derivations of Section 3. Hence, we will consider the following adjusted sample smoother vector

$$\boldsymbol{s}_{Ak}^T = \boldsymbol{e}_1^T (\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak} + \text{diag}(\delta N^{-2}))^{-1} \boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak}, \quad (5)$$

for some small $\delta > 0$, as done in Breidt and Opsomer (2000). The sample smoother matrix and its centered version are

$$\boldsymbol{S}_A = [\boldsymbol{s}_{Ak}^T : k \in A] \quad \boldsymbol{S}_A^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{\Pi}_A^{-1}/N)\boldsymbol{S}_A$$

with $\boldsymbol{\Pi}_A = \text{diag}\{\pi_k : k \in A\}$. The design-weighted estimators for $\boldsymbol{B}$ and the m_k are

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Z}_A)^{-1} \boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Y}_A \quad (6)$$

$$\hat{m}_k = \boldsymbol{s}_{Ak}^T(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}}), \quad (7)$$

where $\boldsymbol{Z}_A$ and $\boldsymbol{Y}_A$ denote the sample versions of $\boldsymbol{Z}_U$ and $\boldsymbol{Y}_U$, respectively. Note that the estimator $\hat{m}_k$ is defined for any x_k in the population, not only those appearing in the sample. As for the population estimators, these estimators can again be written as the solution to backfitting equations, so that they can be calculated by appropriately weighted versions of the existing algorithms.The estimator for g_k is

$$\hat{g}_k = \hat{m}_k + \boldsymbol{z}_k \hat{\boldsymbol{B}}.$$

The semiparametric model-assisted estimator is then constructed by replacing the g_k in (3) by the $\hat{g}_k$:

$$\hat{y}_{\text{reg}} = \frac{1}{N}\sum_U \hat{g}_k + \frac{1}{N}\sum_A \frac{y_k - \hat{g}_k}{\pi_k}. \quad (8)$$

Defining $\bar{y}_\pi = \sum_A y_k/\pi_k$ and similarly for $\bar{\boldsymbol{z}}_\pi$, an equivalent expression for $\hat{y}_{\text{reg}}$ is given by

$$\hat{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}}_\pi)\hat{\boldsymbol{B}} + \frac{1}{N}\sum_U \hat{m}_k - \frac{1}{N}\sum_A \frac{\hat{m}_k}{\pi_k}, \quad (9)$$

which shows that the semiparametric estimator can be interpreted as a "traditional" linear regression survey estimator using the parametric model component $\boldsymbol{z}\boldsymbol{\beta}$, with an additional correction term for the nonparametric component of the model. This estimator also shares some desirable properties with the fully parametric regression estimators. It is location and scale invariant, and it is calibrated for both the parametric and the nonparametric model components, in the sense that $\hat{\boldsymbol{x}}_{\text{reg}} = \bar{\boldsymbol{x}}_N$ and $\hat{\boldsymbol{z}}_{\text{reg}} = \bar{\boldsymbol{z}}_N$. The calibration for the variables in the parametric term can be checked directly by using expressions (6) and (7), while the calibration for the nonparametrically specified variable x_k follows from the fact that $\boldsymbol{s}_{Ak}^T \boldsymbol{X}_A = x_k$, where $\boldsymbol{X}_A = (x_k : k \in A)^T$ (we are ignoring the effect of the adjustment $\text{diag}(\delta N^{-2})$ in (5), because that adjustment can be made arbitrarily small). In addition, the estimator can be written as a weighted sum of the y_k, $k \in A$, so that a set of weights w_k can be obtained and applied to any survey variable of interest.

3. Properties and extensions

3.1 Design properties

In this section, we explore the design properties of the semiparametric estimator (8). In particular, we prove that $\hat{y}_{\text{reg}}$ is design $\sqrt{n}$-consistent, and we derive its asymptotic distribution, including an estimated variance. This will be done in the design-asymptotic context used in Isaki and Fuller (1982) and in Breidt and Opsomer (2000), in which both the population and the samples increase in size as $N \to \infty$. All proofs and the necessary assumptions are in the Appendix.

In the following theorem, we prove the design consistency of the semiparametric estimator. We also show that the convergence rate is $\sqrt{n}$, the usual rate for design estimators.

Theorem 3.1 *Under the assumptions* $A1 - A8$, *the estimator* $\hat{y}_{\text{reg}}$ *in* (8) *is design consistent with rate* $\sqrt{n}$, *in the sense that*

$$\hat{y}_{\text{reg}} = \bar{y}_N + O_p\left(\frac{1}{\sqrt{n}}\right).$$

The following theorem proves that a central limit theorem for $\hat{y}_{\text{reg}}$ exists whenever it exists for the expansion estimator $\bar{y}_\pi$.

Theorem 3.2 *Under the assumptions* $A1$ - $A8$, *if*

$$\frac{\bar{y}_\pi - \bar{y}_N}{\sqrt{\hat{V}(\bar{y}_\pi)}} \to N(0,\ 1),$$

with

$$\hat{V}(\bar{y}_\pi) = \frac{1}{N^2}\sum\sum_A \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

for a given sampling design, then we also have

$$\frac{\hat{y}_{\text{reg}} - \bar{y}_N}{\sqrt{\hat{V}(\hat{y}_{\text{reg}})}} \to N(0,\ 1),$$

with

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2}\sum\sum_A \frac{\pi_{kl} - \pi_k\,\pi_l}{\pi_{kl}}\frac{y_k - \hat{g}_k}{\pi_k}\frac{y_l - \hat{g}_l}{\pi_l}. \quad (10)$$

3.2 Semiparametric additive model

The results in Theorems 3.1 and 3.2 use the semiparametric model (1), which contains a single univariate nonparametric term $m(\cdot)$. In many practical applications, several auxiliary variables will be available that could be included in the nonparametric portion of a model, but the curse of dimensionality makes it often difficult to combine several variables into a single multi-dimensional nonparametric term. Instead, the variables that are to be included nonparametrically will be treated as univariate components. This results in the *semiparametric additive model*, which is written as

$$\mathrm{E}_\xi(y_k) = g(\boldsymbol{x}_k,\ \boldsymbol{z}_k) = m_1(x_{1k}) + \ldots + m_Q(x_{Qk}) + \boldsymbol{z}_k\boldsymbol{\beta}$$
$$\mathrm{Var}_\xi(y_k) = v(\boldsymbol{x}_k,\ \boldsymbol{z}_k)$$

where the $m_q(\cdot)$, $q = 1, \ldots, Q$ and $v(\cdot,\cdot)$ are unknown smooth functions.

When $Q = 2$, expressions similar to (6) and (7) can be developed, using the additive model decompositions of Opsomer and Ruppert (1997), and for $Q > 2$, recursive expressions can be derived using the approach of Opsomer (2000). The estimator would then be written as in equations (6) and (7), but with the smoother vectors $\boldsymbol{s}_{Ak}$ and smoother matrix $\boldsymbol{S}_A$ replaced by complicated higher-dimensional additive model smoothers (see Opsomer (2000) for details). Because of this, formally proving the properties of the model-assisted estimator for the case with arbitrary Q would be a challenging task beyond the scope of the current article.

In practice, the backfitting algorithm formulation provides a much more efficient and simple way to calculate the semiparametric estimator. Let $\boldsymbol{s}_{Aqk}$ represent the sample smoother vector, as defined in (5), for the variable x_q at the observation x_{qk} and $\boldsymbol{S}_{Aq}$ is the corresponding smoother matrix for the variable x_q. Also, $\hat{m}_{qk}$ denotes the sample-weighted backfitting estimator for $m_q(x_{qk})$ and $\hat{\boldsymbol{m}}_{Aq} = (\hat{m}_{qk},\ k \in A)$. The backfitting algorithm for a model including Q nonparametric terms consists of the following set of equations, iterated to converge:

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}_A^T\boldsymbol{\Pi}_A^{-1}\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^T\boldsymbol{\Pi}_A^{-1}\left(\boldsymbol{Y}_A - \sum_{q=1}^{Q}\hat{\boldsymbol{m}}_{Aq}\right)$$
$$\hat{\boldsymbol{m}}_{A1} = \boldsymbol{S}_{A1}\left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T\hat{\boldsymbol{B}} - \sum_{q\neq 1}\hat{\boldsymbol{m}}_{Aq}\right)$$
$$\vdots \quad \vdots$$
$$\hat{\boldsymbol{m}}_{AQ} = \boldsymbol{S}_{AQ}\left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T\hat{\boldsymbol{B}} - \sum_{q\neq Q}\hat{\boldsymbol{m}}_{Aq}\right).$$

These equations provide weighted fits at the sample locations $k \in A$ only. For the remaining locations $k \in U$ not in A, an additional smoothing step is required after obtaining the $\hat{\boldsymbol{m}}_{Aq}$, $q = 1, \ldots, Q$:

$$\hat{m}_{kq} = \boldsymbol{s}_{Aqk}^T\left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T\hat{\boldsymbol{B}} - \sum_{q'\neq q}\hat{\boldsymbol{m}}_{Aq'}\right).$$

The sample-based estimators for the mean function at all $k \in U$ are then defined as $\hat{g}_k = \hat{m}_{k1} + \ldots + \hat{m}_{kQ} + \boldsymbol{z}_k\hat{\boldsymbol{B}}$, which are used in expression (8) to construct the model-assisted estimator.

4. Application to Northeastern Lakes survey

In this section, we will show the applicability of the semiparametric regression estimator on a dataset of water chemistry samples. As will be illustrated, once a set of auxiliary variables and a model has been selected, computing survey estimators for the semiparametric model is as easy as for linear models, and hence can lead to improved precision for relatively little cost.

The National Surface Water Survey (NSWS) sponsored by the U.S. Environmental Protection Agency (EPA) between the years of 1984 and 1986 estimated 4.2 percent of the lakes in the northeastern region of the United States to be acidic (Stoddard, Kahl, Deviney, DeWalle, Driscoll, Herlihy, Kellogg, Murdoch, Webb and Webster 2003). Acid-sensitive Northeastern lakes were among the concerns addressed by the Clean Air Act Amendment (CAAA) of 1990, which placed restrictions on industrial sulfur and nitrogen emissions in an effort to reduce the acidity of these waters. A common measurement of acidity is acid neutralizing capacity (ANC), which is defined as a water's ability to buffer acid. An ANC value less than zero $\mu\text{eq}/L$ indicates that the water has lost all ability to buffer acid. Surface waters with ANC values below 200 $\mu\text{eq}/L$ are considered at risk of acidification, and values less than 50 $\mu\text{eq}/L$ are considered at high risk (National Acid Precipitation Assessment Program (1991), page 15).

Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency conducted a survey of lakes in the Northeastern states of the U.S. These data were collected in order to determine the effect that restrictions put in place by the CAAA had on the ecological condition of these waters. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited several times during the study period. Multiple measurements on the same lake were averaged in order to obtain one measurement per lake sampled. Lakes to be included in the survey were selected using a complex sampling design commonly employed by EMAP based on a hexagonal grid frame (see Larsen, Thornton, Urquhart and Paulsen (1993) for a description of the sampling design).

Let y_k represent the (possibly averaged) ANC value of the k^{th} sampled lake. A very simple estimate of the ANC mean of the lakes is represented by the expansion estimator $\bar{y}_\pi$. In this as in many surveys, a better choice is the Hájek estimator,

$$\hat{y}_H = \frac{1}{\hat{N}} \sum_{k \in A} \frac{y_k}{\pi_k}, \qquad (11)$$

which applies a ratio type adjustment for the estimation of the population size through $\hat{N} = \sum_{k \in A} 1/\pi_k$. However, auxiliary variables are available for each lake in this population, so that it should be possible to further improve upon the efficiency of the Hájek estimator. The following variables are available for each $k \in U$:

x_k = UTMX, x-geographical coordinate of the centroid of each lake in the UTM coordinate system,

$z_{j,k}$ = indicator variable for eco-region $j = 1, \ldots, 6$,

$z_{7,k}$ = UTMY, y-geographical coordinate,

$z_{8,k}$ = elevation.

There are seven different eco-regions included in the population, thus dummy variables $z_{j,k}$ are constructed for $j = 1, \ldots, 6$. A semiparametric regression estimator for the variable y will be constructed by treating the UTMX variable x as a nonparametric term and the remaining variables $z_1 - z_8$ as a parametric component. Model selection was used to determine that treating the other two continuous variables as nonparametric did not improve the model fit. For comparison purposes, we also computed a regression estimator that treats all terms as parametric. This estimator is therefore identical to the semiparametric estimator, except that the x-geographical coordinate is modeled linearly. We will denote this fully parametric regression estimator by $\hat{y}_{\text{par}}$.

In order to determine the estimated efficiency of survey estimators, we need to compute the variance estimates. However, second order inclusion probabilities were not available, thus we cannot evaluate $\hat{V}(\hat{y}_{\text{reg}})$ as in (10). In order to come up with appropriate variance estimates, we treat the complex sampling design as a stratified sample taken with replacement. The 14 strata we selected correspond to groups of spatial clusters of lakes that appeared in the original design, and that were used to ensure spatial distribution of the sampled lakes over the region of interest. Larsen *et al.* (1993) provide details on the construction of the spatial clusters.

Let H be the number of strata, n_h the number of observations within stratum h, and A_h the set of sampled elements that fall in stratum h. Define $p_k = n_h^{-1} \pi_k$. Using this notation and the assumption of a stratified sample with replacement, we rewrite the semiparametric estimator as

$$\hat{y}_{\text{reg}} = \frac{1}{N} \sum_{k \in U} \hat{g}(x_k, z_k) + \frac{1}{N} \sum_{h \in H} \frac{1}{n_h} \sum_{k \in A_h} \frac{y_k - \hat{g}(x_k, z_k)}{p_k} \qquad (12)$$

and the variance estimator as

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum_{h \in H} S_h^2,$$

where S_h^2 is the estimated within-stratum weighted residual variance for stratum h. Assuming the strata are sampled with replacement, Särndal *et al.* (1992, page 421 - 422) suggest S_h^2 can be calculated as

$$S_h^2 = \frac{1}{n_h(n_h - 1)} \sum_{k \in A_h} \left(\frac{y_k - \hat{g}(x_k, z_k)}{p_k} - \sum_{l \in A_h} \frac{y_l - \hat{g}(x_l, z_l)}{\pi_l} \right)^2. \qquad (13)$$

Similarly, we estimate $\hat{V}(\hat{y}_H)$ through

$$\hat{V}(\hat{y}_H) = \frac{1}{\hat{N}^2}\sum_{h\in H}\frac{1}{n_h(n_h-1)}\sum_{k\in A_h}\left(\frac{y_k-\hat{y}_H}{p_k}-\sum_{l\in A_h}\frac{y_l-\hat{y}_H}{\pi_l}\right)^2, \quad (14)$$

and the expression for $\hat{V}(\hat{y}_{\text{par}})$ is obtained completely analogously as for $\hat{V}(\hat{y}_{\text{reg}})$ except that $\hat{g}(x_k, z_k)$ is computed by linear regression.

This setup allows us to obtain the following estimates of mean ANC for the Northeastern lakes, together with variance estimates and approximate 95% confidence intervals (CI). A local linear fit has been employed for the nonparametric term with bandwidth set at one tenth of the range of UTMX.

$$\hat{y}_{\text{reg}} = 558.0\ \mu\text{eq/L} \quad \hat{V}(\hat{y}_{\text{reg}}) = 2534.6 \quad \text{CI} = (459.3; 656.6)$$

$$\hat{y}_{\text{par}} = 577.3\ \mu\text{eq/L} \quad \hat{V}(\hat{y}_{\text{par}}) = 3239.6 \quad \text{CI} = (465.8; 688.9)$$

$$\hat{y}_H = 555.9\ \mu\text{eq/L} \quad \hat{V}(\hat{y}_H) = 4313.3 \quad \text{CI} = (427.2; 684.7)$$

The confidence interval constructed using the Hájek estimator is about 31% wider than that constructed using the semiparametric estimator, while the interval for the fully parametric regression estimator is 13% wider. These results show evidence of an improvement in efficiency provided by accounting for the auxiliary information in both a parametric and nonparametric way in the mean estimation procedure, with the nonparametric estimator able to capture some additional efficiency beyond that of the parametric estimator.

As mentioned above, an important goal of this application is the assessment of how many lakes are at risk of acidification or are acidified already. That is, we are interested in estimating the proportion of Northeastern lakes with ANC values smaller than some specific threshold values. We can determine such proportions by estimating the finite population distribution function,

$$F_N(t) = \frac{1}{N}\sum_{k\in U} I_{\{y_k\le t\}}$$

at specific threshold values t, where $I_{\{y_k\le t\}}$ denotes the indicator function taking a value of 1 if $y_k \le t$ and 0 otherwise. Because all three estimators can be expressed as weighted sums of sample observations, the weights obtained for each can be applied directly to the $I_{\{y_k\le t\}}$ for the sample to estimate $F_N(t)$ for any desired t. Let us denote by $\hat{F}_H(t)$, $\hat{F}_{\text{reg}}(t)$ and $\hat{F}_{\text{par}}(t)$ the Hájek, semiparametric and parametric regression estimators of the distribution function, respectively. Estimates for their design variances are computed by plugging the indicator variables in equations (13) and (14).

Figure 1 shows estimates of the ANC cdf produced by $\hat{F}_H(t)$, $\hat{F}_{\text{par}}(t)$ and $\hat{F}_{\text{reg}}(t)$ evaluated on a grid of 1,000 equally spaced values for t. Included are their respective pointwise 95% confidence intervals calculated at each grid point. All three estimators are similar, but the confidence bands for the parametric and semiparametric regression estimators tend to be narrower. Averaged over all 1,000 grid points, the widths of the confidence bands are 0.093 for $\hat{F}_H(t)$, 0.084 for $\hat{F}_{\text{par}}(t)$ and 0.075 for $\hat{F}_{\text{reg}}(t)$, respectively.

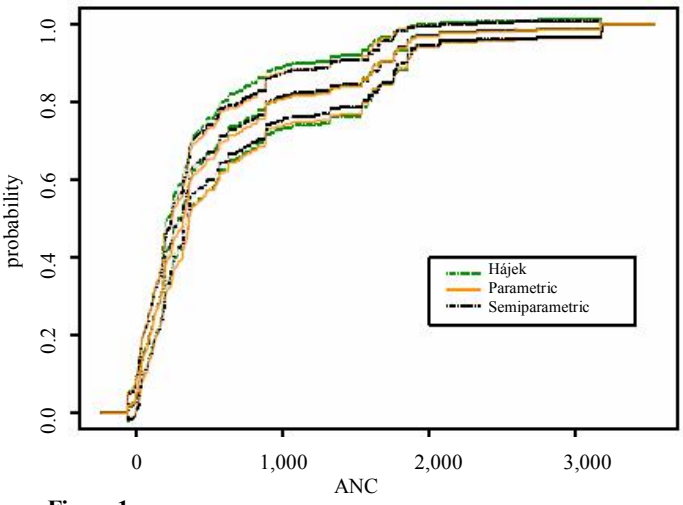

Figure 1
Estimates of the population cumulative distribution function for ANC and confidence bounds produced by Hájek, parametric and semiparametric regression estimators

Along with ANC, the EMAP survey of Northeastern lakes measured the concentration of multiple chemistry variables including sulfate, magnesium and chloride, so that the survey weights obtained for ANC can also be applied to these concentrations as well as their respective cdfs. As another illustration of the semiparametric estimation approach, it is possible to "invert" $\hat{F}_{\text{reg}}(t)$ to obtain quantile estimators $\hat{\theta}_{\text{reg}}(\alpha) = \min\{t : \hat{F}_{\text{reg}}(t) \ge \alpha\}$ of these additional chemistry variables. Table 1 displays semiparametric estimates of the first, second, and third quartiles of sulfate, magnesium, and chloride measured in $(\mu\text{eq}/L)$. Variance estimation for these quantiles could be handled using asymptotic results of Francisco and Fuller (1991), but will not be explored further here.

Table 1 Quartile estimates of chemistry variables

α	Sulfate	Magnesium	Chloride
0.25	73.3	63.8	27.4
0.50	104.3	127.0	162.2
0.75	201.4	221.9	462.2

5. Conclusion

In this article, we have described a model-assisted estimator that uses semiparametric regression to capture relationships between multiple population-level auxiliary variables and the survey variables. We have developed asymptotic theory that shows the resulting estimator is design consistent and asymptotically normal under mild conditions on the design and the population. This generalizes the results of Breidt and Opsomer (2000), who had proved similar results for a univariate nonparametric model-assisted estimator. The semiparametric estimator was applied to data from a survey of lakes in the Northeastern U.S., where it was shown to be more efficient than an estimator that does not take advantage of the auxiliary variables and than a fully parametric regression estimator.

In addition to its theoretical properties, the semiparametric model-assisted estimator has attractive practical properties as well. As noted earlier, it is fully calibrated for the auxiliary variables, whether used in the parametric or nonparametric model components, and it is location and scale invariant. The estimator can be expressed as a weighted sum of the sample observations, so that it conforms to the traditional survey estimation paradigm and a single set of weights can be applied to all the survey variables, hence preserving relationships between variables.

One issue which was not addressed in the current article is the selection of the smoothing parameter for the nonparametric component of the regression model. This is a challenging topic in the model-assisted context, further complicated by the just mentioned fact that a single set of survey regression weights is applied to all the survey variables: because the optimal bandwidth choice depends on the variable being smoothed, no single bandwidth (and hence set of weights) will be optimal for all variables in the survey. This topic is currently being explored by the authors.

Acknowledgments

The research for this article was supported by National Science Foundation grants DMS-0204531 and DMS-0204642, and by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

Appendix

Technical assumptions and derivations

We begin by stating the necessary assumptions, which extend those used in Breidt and Opsomer (2000) to the semiparametric model.

Assumptions:

- **A1** *Distribution of the errors under* ξ: *the errors* ε_k *are independent and have mean zero, variance* $v(x_k, z_k)$, *and compact support, uniformly for all* N.
- **A2** *Distribution of the covariates*: *the* x_k *and* z_k *are considered fixed with respect to the superpopulation model* ξ. *The* z_k *are assumed to have bounded support, and the* x_k *are independent and identically distributed* $F(x) = \int_{-\infty}^{x} f(t)\,dt$, *where* $f(\cdot)$ *is a density with compact support* $[a_x, b_x]$ *and* $f(x) > 0$ *for all* $x \in [a_x, b_x]$.
- **A3** *Nonparametric mean and variance functions*: *the mean function* $m(\cdot)$ *is continuous, and the variance function* $v(\cdot, \cdot)$ *is bounded and strictly greater than* 0.
- **A4** *Kernel* K: *the kernel* $K(\cdot)$ *has compact support* $[-1, 1]$, *is symmetric and continuous, and satisfies* $\int_{-1}^{1} K(u)\,du = 1$.
- **A5** *Sampling rate* nN^{-1} *and bandwidth* h_N: *as* $N \to \infty$, $nN^{-1} \to \pi \in (0, 1)$, $h_N \to 0$ *and* $Nh_N^2/(\log\log N) \to \infty$.
- **A6** *Inclusion probabilities* π_k *and* π_{kl}: *for all* N, $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k,\, l \in U_N} \pi_{kl} \geq \lambda^* > 0$ *and*
$$\limsup_{N\to\infty} n \max_{k,\, l \in U_{N:\, i \neq j}} |\pi_{kl} - \pi_k \pi_l| < \infty.$$
- **A7** *Additional assumptions involving higher-order inclusion probabilities*:
$$\lim_{N\to\infty} n^2 \max_{(k_1, k_2, k_3, k_4) \in D_{4,\, N}} |E_p(I_{k_1} - \pi_{k_1})(I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})(I_{k_4} - \pi_{k_4})| < \infty,$$
where $D_{t,\, N}$ *denotes the set of all distinct* t*-tuples* $(k_1, k_2, \ldots, k_t)$ *from* U_N,
$$\lim_{N\to\infty} \max_{(k_1, k_2, k_3, k_4) \in D_{4,\, N}} |E_p(I_{k_1} I_{k_2} - \pi_{k_1 k_2})(I_{k_3} I_{k_4} - \pi_{k_3 k_4})| = 0,$$
and
$$\limsup_{N\to\infty} n \max_{(k_1, k_2, k_3) \in D_{3,\, N}} |E_p(I_{k_1} - \pi_{k_1})^2 (I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})| < \infty.$$
- **A8** *The matrix* $N^{-1}\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Z}_U$ *is invertible for all* N *with model probability* 1.

Assumption A8 is required so that the population estimator $\boldsymbol{B}$ is well-defined. The invertibility of the matrix in A8 depends on the combined effect of the bandwidth h and the joint distribution of the x_k and z_k. While it would in principle be possible to write down sufficient conditions for this, we opted for this simpler and more explicit approach.

Before giving the proofs of Theorems 3.1 and 3.2, we state and prove a number of lemmas.

Lemma 1 *Under the assumptions A*1-*A*7,

(a) *for all* $k \in U$ *and* $d = 1, \ldots, D$,

$$\frac{1}{N}\sum_U E_p\,(\boldsymbol{s}_{Ak}^T \boldsymbol{Y}_A - \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U)^2 = O\left(\frac{1}{nh}\right)$$

and

$$\frac{1}{N}\sum_U E_p\,(\boldsymbol{s}_{Ak}^T \boldsymbol{Z}_{dA} - \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_{dU})^2 = O\left(\frac{1}{nh}\right);$$

(b) *the* $\boldsymbol{s}_{UK}^T \boldsymbol{Y}_U$ *and* $\boldsymbol{s}_{UK}^T \boldsymbol{Z}_U$ *are uniformly bounded over all* $k \in U$.

Proof of Lemma 1: Since both the y_k and z_{dk} are bounded by assumption, part (a) can be shown using an identical reasoning as in Lemma 4 of Breidt and Opsomer (2000). While that lemma did not include a rate of convergence, that rate is readily derived by noting that

$$\frac{1}{N}\sum_{i,\,k\in U_N} z_{ik}^2 = O\left(\frac{1}{nh}\right)$$

in the notation of Breidt and Opsomer (2000) and then proceeding as in that proof.

Part (b) was proven directly in Lemma 2 (iv) of Breidt and Opsomer (2000).

Lemma 2 *Under assumptions A*1-*A*8,

$$\hat{\boldsymbol{B}} = \boldsymbol{B} + O_p(1/\sqrt{nh}),$$

with the rate holding component-wise, and $\boldsymbol{B}$ *is bounded for all* N.

Proof of Lemma 2: Write $\tilde{y}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U$ and $\tilde{y}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Y}_A$ for the population and sample smoothed versions of y_k, and similarly, $\tilde{\boldsymbol{z}}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_U$ and $\tilde{\boldsymbol{z}}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Z}_A$. We rewrite expression (6) as a function of sample-weighted terms $\hat{\boldsymbol{t}}_l$, $l = 1, \ldots, 6$:

$$\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{t}_1 & \hat{\boldsymbol{t}}_2 \\ \hat{\boldsymbol{t}}_3 & \hat{\boldsymbol{t}}_4 \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_5 \\ \hat{\boldsymbol{t}}_6 \end{bmatrix},$$

where

$$\hat{t}_1 = \left(\frac{\hat{N}}{N}\right)^2$$

$$\hat{\boldsymbol{t}}_2 = \overline{\boldsymbol{z}}_\pi - \frac{1}{N}\sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k}\left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_3 = \overline{\boldsymbol{z}}_\pi^T \left(\frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_4 = \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \boldsymbol{z}_k}{\pi_k} - \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k} + \overline{\boldsymbol{z}}_\pi^T \frac{1}{N}\sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k}$$

$$\hat{t}_5 = \overline{y}_\pi - \frac{1}{N}\sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k}\left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_6 = \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T y_k}{\pi_k} - \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} + \overline{\boldsymbol{z}}_\pi^T \frac{1}{N}\sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k}.$$

The sample-weighted estimator $\hat{\boldsymbol{B}}$ will be expanded around

$$\boldsymbol{B} = \begin{bmatrix} 1 & \overline{\boldsymbol{z}}_N \\ \overline{\boldsymbol{z}}_N^T & \boldsymbol{t}_4 \end{bmatrix}^{-1} \begin{bmatrix} \overline{y}_N \\ \boldsymbol{t}_6 \end{bmatrix}, \tag{15}$$

where

$$\boldsymbol{t}_4 = \frac{1}{N}\sum_U \boldsymbol{z}_k^T \boldsymbol{z}_k - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_U]} + \overline{\boldsymbol{z}}_N^T \frac{1}{N}\sum_U \tilde{\boldsymbol{z}}_k^{[s_U]}$$

$$\boldsymbol{t}_6 = \frac{1}{N}\sum_U \boldsymbol{z}_k^T y_k - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} + \overline{\boldsymbol{z}}_N^T \frac{1}{N}\sum_U \tilde{y}_k^{[s_U]}$$

and the remaining $\boldsymbol{t}_l$ can be found in (15). The existence and continuity of the derivatives of $\hat{\boldsymbol{B}}$ with respect to the $\hat{\boldsymbol{t}}_l$ and evaluated at $\boldsymbol{t}_l$ follow from Lemma 1(b) and the existence of the inverse in (15), which is assumed by A8.

The result will follow from a 0th order Taylor expansion if we can show that $\hat{\boldsymbol{t}}_l - \boldsymbol{t}_l = O_p(1/\sqrt{nh})$ for all l (*e.g.*, Fuller (1996), Corollary 5.1.5). For $\hat{t}_1$ and $\hat{\boldsymbol{t}}_3$, this follows directly from A2 and A6. The remaining terms contain sums involving smoothed quantities $\tilde{\boldsymbol{z}}_k^{[s_A]}$ and $\tilde{y}_k^{[s_A]}$. We demonstrate the reasoning for one of those terms in $\hat{\boldsymbol{t}}_6$. We have

$$\frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} = \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]}\left(\frac{I_k}{\pi_k} - 1\right) + \frac{1}{N}\sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})\frac{I_k}{\pi_k},$$

and the first term is $O_p(1/\sqrt{n})$ by A6 and Lemma 1(b), using the same argument as in Lemma 4 of Breidt and Opsomer (2000). For the second term, use Schwarz's inequality

$$\left|\frac{1}{N}\sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})\frac{I_k}{\pi_k}\right| \le \sqrt{\frac{1}{N}\sum_U \boldsymbol{z}_k^{[2]T} \frac{I_k}{\pi_k^2}}\sqrt{\frac{1}{N}\sum_U (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})^2},$$

where $\boldsymbol{z}_k^{[2]}$ denotes that the squares are computed component-wise. The first term is bounded by A2 and A6, and the second term is $O_p(1/\sqrt{nh})$ by Lemma 1(a) and Markov's inequality. The desired result then follows by applying the same reasoning to the remaining terms in $\hat{\boldsymbol{t}}_2$, $\hat{\boldsymbol{t}}_4$, $\hat{t}_5$, $\hat{\boldsymbol{t}}_6$.

The boundedness of $\boldsymbol{B}$ follows directly from assumption A8, Lemma 1(b) and the boundedness of the z_k.

Lemma 3 *Under the assumptions A*1-*A*8, *we have*

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof of Lemma 3: Given expression (9), we need to show that

$$(\overline{\boldsymbol{z}}_N - \overline{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = o_p\left(\frac{1}{\sqrt{n}}\right) \tag{16}$$

$$\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{17}$$

Lemma 2 and assumptions A2, A5 and A6 show that $(\overline{\boldsymbol{z}}_N - \overline{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = O_p(1/nh)$. In order to prove (17), we can rewrite it as

$$\begin{aligned}\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) &= \frac{1}{N}\sum_U (\tilde{y}_k^{[s_U]} - \tilde{y}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right) \\ &\quad - \frac{1}{N}\sum_U (\tilde{\boldsymbol{z}}_k^{[s_U]} - \tilde{\boldsymbol{z}}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right)\boldsymbol{B} \\ &\quad - \frac{1}{N}\sum_U \tilde{\boldsymbol{z}}_k^{[s_A]}\left(1 - \frac{I_k}{\pi_k}\right)(\boldsymbol{B} - \hat{\boldsymbol{B}}).\end{aligned}$$

The first term on the right hand side has been proven to be $o_p(1/\sqrt{n})$ in Lemma 5 of Breidt and Opsomer (2000); this same Lemma and boundness of $\boldsymbol{B}$ provide the same rate for the second term. Assumptions A5-A6, Lemma 1(b) and Lemma 2 show that the third term is $O_p(1/n\sqrt{h})$ and the desired rate is achieved.

Lemma 4 *Under assumptions A*6 *and A*8,

$$\begin{aligned}E_p(\hat{y}_{\text{dif}}) &= \overline{y}_N \\ \text{Var}_p(\hat{y}_{\text{dif}}) &= \frac{1}{N^2}\sum\sum_{k,l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l} \\ &= O\left(\frac{1}{n}\right).\end{aligned}$$

Proof of Lemma 4: The properties of the difference estimator are readily computed. The rate of the design variance follows from the stated assumptions using the same reasoning as in Lemma 4 of Breidt and Opsomer (2000).

Lemma 5 *Under assumptions A*1-*A*8,

$$\hat{V}(\hat{y}_{\text{reg}}) = \text{Var}_p(\hat{y}_{\text{dif}}) + o_p\left(\frac{1}{n}\right).$$

Proof of Lemma 5: The reasoning for this proof will closely follow that of Theorem 3 of Breidt and Opsomer (2000). We write

$$\begin{aligned}\hat{V}(\hat{y}_{\text{reg}}) - \text{Var}_p(\hat{y}_{\text{dif}}) = &\ (\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) \\ &+ (\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})) \quad (18)\end{aligned}$$

with

$$\hat{V}(\hat{y}_{\text{dif}}) = \frac{1}{N^2}\sum\sum_A \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l}.$$

Since

$$\frac{1}{N}\sum_U (y_k - g_k)^4 < \infty.$$

by assumptions A1-A3 and from Lemmas 1(b) and 2, the approach used for the term A_N of Breidt and Opsomer (2000) can be used to show that

$$\text{E}_p|\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})| = o\left(\frac{1}{n}\right),$$

which provides the desired consistency by the Markov inequality.

For the first term in (18), note that

$$\begin{aligned}\hat{g}_k - g_k = &\ (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\,(\hat{\boldsymbol{B}} - \boldsymbol{B}) \\ &+ (\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_k^{[s_U]})(\hat{\boldsymbol{B}} - \boldsymbol{B}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\boldsymbol{B},\end{aligned}$$

so that

$$\begin{aligned}&(\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) = \\ &\frac{1}{N^2}\sum\sum_U \left\{\begin{array}{l} -2\dfrac{y_k - g_k}{\pi_k}\dfrac{\hat{g}_l - g_l}{\pi_l} \\ \quad + \dfrac{\hat{g}_k - g_k}{\pi_k}\dfrac{\hat{g}_l - g_l}{\pi_l}\end{array}\right\}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l\end{aligned}$$

can be decomposed into variance terms involving sample and population smooths and parameter estimators. Each of these terms can be shown to be $o_p(1/n)$. We demonstrate the approach on one of the terms:

$$\begin{aligned}&\left|\frac{1}{N^2}\sum\sum_U \frac{y_k - g_k}{\pi_k}\frac{\hat{z}_l - z_l}{\pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l(\hat{\boldsymbol{B}} - \boldsymbol{B})\right| \\ &\le \left(\frac{C_1}{N} + C_2 \max|\pi_{kl} - \pi_k\pi_l|\right)\frac{1}{N}\sum_U |\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]}||\hat{\boldsymbol{B}} - \boldsymbol{B}| \\ &= o_p\left(\frac{1}{n}\right)\end{aligned}$$

where $C_1, C_2 < \infty$ summarize the bounded terms (by assumptions A1-A3 and A6 and Lemma 1(b)), and the rate of convergence is the result of assumption A6 and Lemmas 1(a) and 2.

Proof of Theorem 3.1: In Lemma 3, we show that

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\hat{y}_{\text{dif}}$ is the difference estimator (3). The result immediately follows from assumption A5 and Lemma 4.

Proof of Theorem 3.2: Note that $\hat{y}_{\text{dif}}$ can be written as the sum of a population constant and an expansion estimator of the form $\bar{y}_\pi$ by defining a new variable $y_k - \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U + \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_U \boldsymbol{B} - \boldsymbol{z}_k \boldsymbol{B}$ for $k \in U$. As is the case for the original y_k, this new variable has bounded support by Lemma 1(b) and a variance of order $O(1/n)$ by Lemma 4. Hence, existence of the CLT for $\bar{y}_\pi$ implies existence of the CLT for $\hat{y}_{\text{dif}}$. Also, $\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p(1/\sqrt{n})$ by Lemma 3, so that $\sqrt{n}\,\hat{y}_{\text{reg}}$ and $\sqrt{n}\,\hat{y}_{\text{dif}}$ have the same asymptotic distribution. Applying Slutsky's Theorem and Lemma 5 complete the proof.

References

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

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

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

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*. Washington, D.C.: Chapman and Hall.

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

Larsen, D.P., Thornton, K.W., Urquhart, N.S. and Paulsen, S.G. (1993). Overview of survey design and lake selection. EMAP - Surface Waters 1991 Pilot Report. Technical Report EPA/620/R - 93/003, U.S. Environmental Protection Agency. (Eds. D.P Larsen and S.J. Christie).

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*. To appear.

Opsomer, J.-D., and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.

Opsomer, J.D., and Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, 8, 715-732.

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

Speckman, P.E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society*, Series B, 50, 413-436.

Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Webb, J.R. and Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Technical Report EPA/620/R-03/001, U.S. Environmental Protection Agency, Washington, DC.

U.S. National Acid Precipitation Assessment Program (1991, November). 1990 Integrated Assessment Report. Technical report, Washington, DC.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 45-50
Statistics Canada, Catalogue No. 12-001

Ex post weighting of price data to estimate depreciation rates

Marc Tanguay and Pierre Lavallée [1]

Abstract

To model economic depreciation, a database is used that contains information on assets discarded by companies. The acquisition and resale prices are known along with the length of use of these assets. However, the assets for which prices are known are only those that were involved in a transaction. While an asset depreciates on a continuous basis during its service life, the value of the asset is only known when there has been a transaction. This article proposes an *ex post* weighting to offset the effect of source of error in building econometric models.

Key Words: Price ratio; Survival data; Uniform distribution; Depreciation of vehicles.

1. Introduction

Various econometric models are used to estimate economic depreciation. To this end, we use a database containing information on assets discarded by companies. The acquisition and resale prices are known along with the length of use of these assets. From this information, we would like to infer results for the total population of assets used by companies. Regarding the use of the prices of used assets to estimate economic depreciation, we refer the reader to, Gellatly, Tanguay and Yan (2002) and Hulten and Wykoff (1981).

We question, however, the representativeness of the database used. Indeed, the assets for which prices are known are solely those subject to a transaction. We do not know the extent to which the losses of value observed on these assets are representative of the loss of value for all assets in production, regardless of whether they were the subject of a transaction. This situation can be a source of error in building econometric models because these models seek to measure depreciation of assets over their service lives, regardless of whether there was a transaction.

It is this second source of error that we propose to offset, at least in part, by applying *ex post* weighting when building econometric models. Section 2 of this article will describe the problem in greater detail, while in Section 3, we will describe the approach used to determine the weights. Finally, in Section 4, we present some numeric results.

2. Problem

We are seeking to describe the relationship between prices and asset age. There is a sample of n assets where we know, for each asset i, the price ratio r_i and the time t_i when this ratio was measured. Once prices are expressed in real dollars, this ratio is given as $r_i = P_i^t / P_i^0$ where P_i^0 is the initial value of the investment in asset i and P_i^t is its resale price at time t. This ratio is strictly decreasing in relation to the time axis t. At the start, we do not know the process that generates the loss in value and there are no specifics about the function that describes this loss except that it is strictly decreasing. However, it is possible to examine the distribution of the price ratios between 0 and 1. Here is an example constructed from data on manufacturing plants (note that 2/3 of the sample was excluded because it corresponds to discarded assets (the price is zero) and the estimation procedures take this component into account, each in its own way).

Since we want to use the data to infer statistics on the population of assets in production, we would like our data to have properties similar to those of a random sample drawn from that population. As we stated earlier, this is not the case because we only have the prices of assets i that were subject to a transaction at time t_i, $i = 1, ..., n$. In effect, while we would like to have price ratios for various periods in the existence of a given asset i, the ratio is only available when there has been a transaction, something that occurs in a non-uniform manner over an asset's service life.

Consequently, we can ask ourselves what form the above distribution might have if it had been drawn from a sample in which the price ratio had been measured, for the same asset i at different times t. Our argument is that it should converge toward a *uniform distribution*. We will therefore seek to obtain a weighting that will help us recreate a uniform distribution of price ratios. This weighting will help us offset the lack of uniformity in the distribution of observations, which may impact statistical analyses such as linear regression.

1. Marc Tanguay and Pierre Lavallée, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. E-mail: marc.tanguay@statcan.ca, pierre.lavallee@statcan.ca.

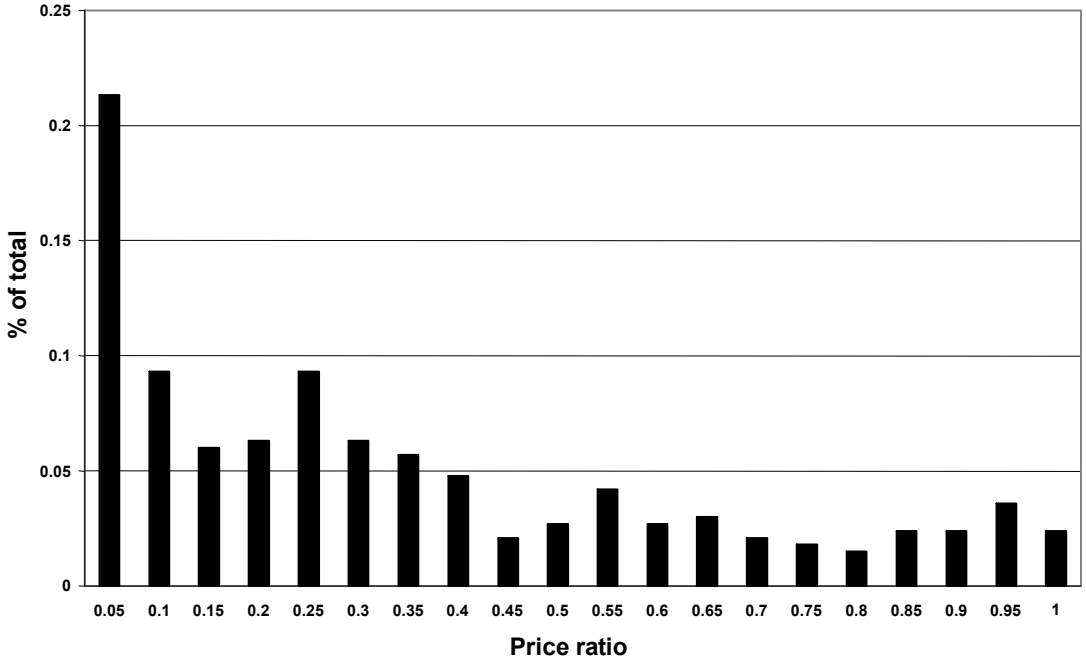

Figure 1 Distribution of observations by price ratio, manufacturing plants

3. Approach

Our starting point is that price ratios can be considered empirical realizations of an unknown form of survival function. In service life models, the survival function expresses the probability that an entity with a limited service life will survive beyond a certain point on the time axis. Accordingly, it provides the same information as a distribution function (or Cumulative Distribution Function). We will let r, be a random variable describing the service life of a unit of value incorporated in some asset. The value gradually erodes over time for as long as the asset is in service. The price ratio can therefore be interpreted as the surviving fraction that gradually becomes smaller and smaller. This fraction is written as $S(y)$ and gives

$$S(y) = 1 - F(y)$$

where $F(y) = P(r \leq y)$ is the distribution function, that is, the probability that a unit of value is lost before point y.

Fundamental transformation theorems of probability laws provide the means for defining the inverse function of $F(y)$ (Greene 1993 and Ross 2002). We let $z = F(y)$ and assume that the inverse function F^{-1} exists so that $y = F^{-1}(z)$. This shows that there is a direct match between the space of y, bounded at 0 but infinite to the right, and that of F which is bound between 0 and 1. The distribution function of z is $F(F^{-1}(z)) = z$. The law that generates this distribution is a uniform distribution between 0 and 1.

This result is generally at the core of data generation processes like Monte Carlo simulations because the uniform distribution is often used when a random sample is being generated, followed by the application of the inverse function (Davidson and MacKinnon 1993). This approach is not always practical and indeed is sometimes patently impossible, especially if the inverse function F^{-1} is not explicit. This result has also been used in generalized remainder approaches, notably to build specification tests (Lancaster 1985).

The result is that any random sample built using empirical realizations of survival proportion data must converge in distribution toward a uniform distribution.

In the case of price data, intuition suggests that between the time of investment and that of disposal, the full range of relative prices must be covered by an asset in production. Initially, value depreciates faster and therefore there are more observations with short periods of time. This is offset by the fact that the corresponding reference on the time scale is also shorter. For example, it takes less time to move from 100% of the initial value to 90%, than from 15% to 5% of the initial value.

It is easy to verify these findings numerically using simulated data and we will not spend time on this. Rather, we will examine how this result can be reintroduced in the database to produce, at least partially, properties similar to those of a random sample. W*e can do this by simply imposing ex post on the empirical price distribution a*

weight structure w_i *that ensures that the empirical distribution of the data, in the price space, is uniform.*

The empirical distribution of price ratios r is given by

$$\hat{F}_n(y) = \frac{\sum_{i=1}^{n} I_i(y)}{n} \quad (1)$$

where $I_i(y) = 1$ if the measured value r_i of asset *i* is less than or equal to *y* (specifically, $r_i \le y$), and 0 otherwise, and *n* is the total number of observations. Note that if the *n* units of the sample are independent and identically distributed (i.i.d.), when $n \to \infty$, $\hat{F}_n(y)$ converges in probability to $F(y)$, that is, $\hat{F}_n(y) \xrightarrow{P} F(y)$ (Bickel and Doksum 1977).

To obtain weight w_i for each asset *i*, we simply distribute the sample in a given number H of intervals (or classes) of a fixed size on the scale of price ratios, and we assign the same probability $\pi = 1/H$ to each of these intervals. Since the price ratios are bounded by 0 and 1, we then have the interval $h = 1$ given by $[0, H^{-1}]$, and for $h = 2, ..., H$, the intervals are given by $](h-1)H^{-1}, hH^{-1}]$. A weight w_h is then calculated in each interval *h* by the ratio $\pi / \hat{\pi}_h$ where $\hat{\pi}_h$ is the empirical probability specific to interval *h*, producing

$$\hat{\pi}_h = \frac{1}{n} \sum_{i=1}^{n} \delta_i(h) = \frac{n_h}{n} \quad (2)$$

where $\delta_i(h) = 1$ if $r_i \in h$, 0 otherwise. We then propose

$$w_i = w_h = \frac{\pi}{\hat{\pi}_h} = \frac{n}{Hn_h} \quad (3)$$

for $r_i \in h$. Using these weights, the *weighted empirical distribution* of the price ratios *r* is given by

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^{n} w_i I_i(y)}{\sum_{i=1}^{n} w_i}. \quad (4)$$

By writing $\sum_{i=1}^{n} w_i = \sum_{h=1}^{H} \sum_{i=1}^{n_h} n / Hn_h = n$, we finally get

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^{n} w_i I_i(y)}{n}. \quad (5)$$

Since $n_h = \sum_{i=1}^{n} \delta_i(h)$, we have

$$\begin{aligned} \hat{F}_{n,w}(y) &= \frac{\sum_{i=1}^{n} w_i I_i(y)}{n} \\ &= \frac{1}{H} \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n} \delta_i(h) I_i(y) \\ &= \frac{1}{H} \sum_{h=1}^{H} \frac{\sum_{i=1}^{n} \delta_i(h) I_i(y)}{\sum_{i=1}^{n} \delta_i(h)} \\ &= \frac{1}{H} \sum_{h=1}^{H} \hat{F}_n(y \mid h). \end{aligned} \quad (6)$$

When $n \to \infty$, we have $(1/n) \sum_{i=1}^{n} \delta_i(h) I_i(y) \xrightarrow{P} P(r \in h, r \le y)$ and $(1/n) \sum_{i=1}^{n} \delta_i(h) \xrightarrow{P} P(r \in h)$. Thus, when $n \to \infty$,

$$\begin{aligned} \hat{F}_n(y \mid h) &\xrightarrow{P} \frac{P(r \in h, r \le y)}{P(r \in h)} \\ &= P(r \le y \mid r \in h) = F(y \mid h) \end{aligned} \quad (7)$$

where $F(y \mid h)$ is the distribution of price ratios r within interval *h*.

For a sufficiently large *n*, *H* must be determined in such a way as to build the intervals *h* so that $\hat{F}_n(y \mid h)$ is distributed approximately uniformly, $h = 1, ..., H$. In other words, when $n \to \infty$, for a sufficiently large *H*, $F(y \mid h)$ should have a uniform distribution on interval *h*. Note that this argument was used by Dalenius and Hodges (1959) in a context of optimal stratification. In this case, the distribution $F(y \mid h)$ is given by

$$F(y \mid h) = \begin{cases} 0 \text{ for } y \le (h-1)H^{-1} \\ Hy - h + 1 \text{ for } (h-1)H^{-1} < y \le hH^{-1} \\ 1 \text{ for } y > hH^{-1}. \end{cases} \quad (8)$$

Since $F(y) = \sum_{h=1}^{H} F(y \mid h) / H$, we have $F(y) = y$, which corresponds to the uniform distribution. We conclude from this that for a sufficiently large *n*, the use of weighting (3) should ensure that the weighted empirical distribution $\hat{F}_{n,w}(y)$ given by (5) is distributed approximately uniformly.

Monte Carlo simulations have shown that estimates produced from a non-random sample could be improved by using this approach. Its main advantages can be attributed to:

- its simplicity;
- the fact that it can be introduced *ex ante*, or prior to introducing the econometric model as such. Consequently, it does not require strong working hypotheses.

If we go back to the histogram presented earlier and divide the sample in $H = 5$ intervals of a width of 0.2 and a value of $\pi = 1/5 = 0.2$, we then get the following histogram that was weighted *ex post*.

4. Application

We will now illustrate our approach using an example taken from the Kelly Blue Book, a source of information widely used to estimate depreciation of automobiles. Table 1 shows the prices of two models of cars at different ages between 1 and 18 years. For each car, we have a sample of $n = 18$ units. Prices are expressed in relative value in

relation to a new model. The ratios also have to be adjusted to take into account the survival probability at each of these ages. For each vehicle, the final ratio used r_i for year i is built from the product of the price ratio times the survival probability.

We are interested in the average depreciation rate $\bar{\tau}$ for each car. This can be estimated from a regression of the prices (or from a function of these prices) in relation to age (or a function of age). However, if we assume that the rate is constant and geometric, we obtain the relationship $r_i = 1 - \bar{\tau}^i$, where r_i is the relative price based on age i. In this case, a rate $\hat{\tau}_i$ can be estimated at each age i by $\hat{\tau}_i = 1 - r_i^{1/i}$. An estimate of the average rate of depreciation is then produced from the average for all ages, $\hat{\bar{\tau}} = \sum_{i=1}^{18} \hat{\tau}_i / 18$.

In the above example, we see that the depreciation rates $\hat{\tau}_i$ vary by age range and that they tend to increase with age. Moreover, the fact that we use a simple average of the ages in calculating $\hat{\bar{\tau}}$ again implicitly gives the same weight to each age. However, it is quite clear that this is not the distribution that we would get from a random sample of service vehicles. The figure below shows the distribution of price cells between ratios of 0 and 1.

The reweighting technique simply involves applying an equal weight to each of the relative price ranges. In this example, the $n = 18$ ages are distributed into $H = 7$ classes, resulting in 18/7 of the ages in each class (in reality, the structures of the cells was configured into 8 classes but the last is always empty). As mentioned in Section 3, the individual weights w_i for each age i are built using (3), that is, by dividing 18/7 by the number of observations found in each class, except for the empty cells where the weight remains zero. Table 2 shows the results and the impact of reweighting on the derived statistics.

This example clearly illustrates the problems of aggregation bias typical of regressions estimated from economic aggregates without taking account the real distribution of the units at the micro level. Thus, it is quite clear that the units at 17 and 18 years would not have the same regression weight as those at 1 year because the risk of loss at 1 year affects almost all vehicles to be put into circulation, while very few of them will be exposed to the risk of loss of value at more advanced ages. The result is that the unweighted estimate in this example produces an over-estimation of the depreciation rate in the order of 15%.

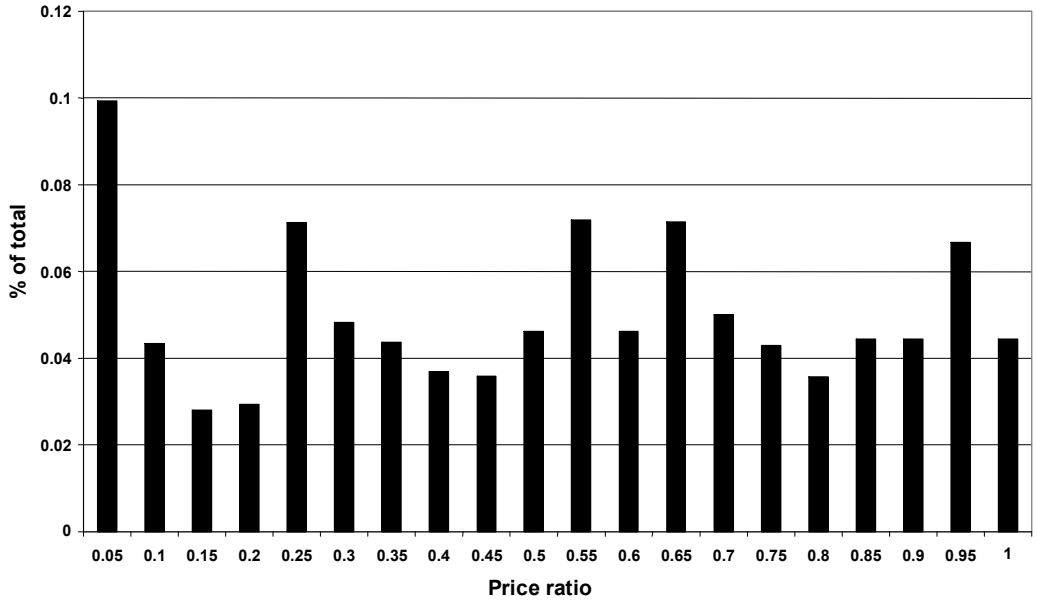

Figure 2 Weighted distribution of observations by price ratio, manufacturing plants *Ex post* weighting

Table 1 Relative prices of two models of cars based on the Kelly Blue Book and average depreciation rates before weighting

	Pr ($t > S$)*	Relative price				Average depreciation rates	
		Excluding disposals		Including disposals		Including disposals	
Year		*Buick*	*Chrysler*	*Buick*	*Chrysler*	*Buick*	*Chrysler*
1	0.9988	0.8633	0.8257	0.8622	0.8246	0.1367	0.1743
2	0.9901	0.7435	0.6801	0.7361	0.6734	0.1377	0.1753
3	0.9666	0.6410	0.5608	0.6195	0.5420	0.1378	0.1754
4	0.9220	0.5523	0.4621	0.5092	0.4261	0.1379	0.1755
5	0.8526	0.4740	0.3794	0.4042	0.3234	0.1387	0.1762
6	0.7582	0.4034	0.3087	0.3058	0.2341	0.1404	0.1779
7	0.6433	0.3391	0.2482	0.2181	0.1597	0.1432	0.1805
8	0.5164	0.2790	0.1953	0.1441	0.1009	0.1475	0.1846
9	0.3892	0.2227	0.1491	0.0867	0.0580	0.1537	0.1906
10	0.2731	0.1639	0.1050	0.0448	0.0287	0.1654	0.2018
11	0.1770	0.1261	0.0772	0.0223	0.0137	0.1716	0.2077
12	0.1051	0.0892	0.0523	0.0094	0.0055	0.1824	0.2180
13	0.0567	0.0614	0.0344	0.0035	0.0019	0.1932	0.2284
14	0.0276	0.0441	0.0236	0.0012	0.0007	0.1999	0.2347
15	0.0120	0.0320	0.0164	0.0004	0.0002	0.2050	0.2396
16	0.0046	0.0190	0.0093	0.0001	0.0000	0.2194	0.2534
17	0.0016	0.0088	0.0041	0.0000	0.0000	0.2432	0.2761
18	0.0005	0.0051	0.0023	0.0000	0.0000	0.2542	0.2867
					Average	0.1727	0.2087

* Survival probability based on estimates from the Micro-Economic Studies and Analysis Division of Statistics Canada.

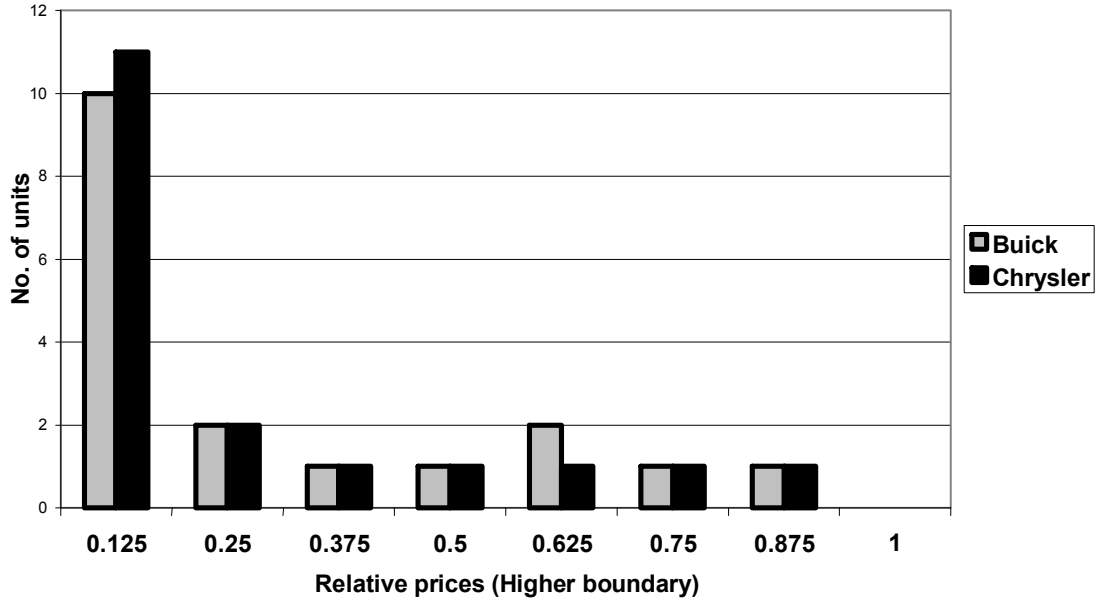

Figure 3 Distribution of cells used to estimate the average depreciation rate using data from the Kelly Blue Book before weighting (Total = 18)

Table 2 Relative prices of two models of cars based on the Kelly Blue Book and the average depreciation rate after weighting

	Relative prices		Average depreciation rates		*Ex post* weights	
	Including disposals		Including disposals			
Year	*Buick*	*Chrysler*	*Buick*	*Chrysler*	*Buick*	*Chrysler*
1	0.8622	0.8246	0.1367	0.1743	2.5714	2.5714
2	0.7361	0.6734	0.1377	0.1753	2.5714	2.5714
3	0.6195	0.5420	0.1378	0.1754	1.2857	2.5714
4	0.5092	0.4261	0.1379	0.1755	1.2857	2.5714
5	0.4042	0.3234	0.1387	0.1762	2.5714	2.5714
6	0.3058	0.2341	0.1404	0.1779	2.5714	1.2857
7	0.2181	0.1597	0.1432	0.1805	1.2857	1.2857
8	0.1441	0.1009	0.1475	0.1846	1.2857	0.2338
9	0.0867	0.0580	0.1537	0.1906	0.2571	0.2338
10	0.0448	0.0287	0.1654	0.2018	0.2571	0.2338
11	0.0223	0.0137	0.1716	0.2077	0.2571	0.2338
12	0.0094	0.0055	0.1824	0.2180	0.2571	0.2338
13	0.0035	0.0019	0.1932	0.2284	0.2571	0.2338
14	0.0012	0.0007	0.1999	0.2347	0.2571	0.2338
15	0.0004	0.0002	0.2050	0.2396	0.2571	0.2338
16	0.0001	0.0000	0.2194	0.2534	0.2571	0.2338
17	0.0000	0.0000	0.2432	0.2761	0.2571	0.2338
18	0.0000	0.0000	0.2542	0.2867	0.2571	0.2338
				Weighted average	0.1479	0.1836

Acknowledgements

The authors would like to express their sincere appreciation to the anonymous referee of *Survey Methodology*, whose thoughtful comments helped improve the quality of the article.

References

Bickel, P.J., and Doksum, K.A. (1977). *Mathematical Statistics*, Holden-Day, Oakland, CA.

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

Davidson, R., and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, N.Y.

Gellatly, G., Tanguay, M. and Yan, B. (2002). An alternative methodology for estimating economic depreciation: New results using a survival model. In *Productivity Growth in Canada - 2002*, Statistics Canada. #15-204-XPE.

Greene, W.H. (1993). *Econometric Analysis*. Second edition, Prentice Hall, Englewood Cliffs, N.J.

Hulten, C.R., and Wykoff, F.C. (1981). The measurement of economic depreciation. In *Depreciation, Inflation, and the Taxation of Income from Capital*, (Ed. C.R. Hulten). The Urban Institute Press, Washington, D.C, 81-125.

Lancaster, T. (1985). Generalized residuals and heterogeneous duration model: With applications to the weibull model. *Journal of Econometrics*, 28, 155-69.

Ross, S.M. (2002). *Introduction to Probability Models*, 8st Edition, Academic Press.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 51-60
Statistics Canada, Catalogue No. 12-001

Person-level and household-level regression estimation in household surveys

David G. Steel and Robert G. Clark [1]

Abstract

A common class of survey designs involves selecting all people within selected households. Generalized regression estimators can be calculated at either the person or household level. Implementing the estimator at the household level has the convenience of equal estimation weights for people within households. In this article the two approaches are compared theoretically and empirically for the case of simple random sampling of households and selection of all persons in each selected household. We find that the household level approach is theoretically more efficient in large samples and any empirical inefficiency in small samples is limited.

Key Words: Contextual effects; Generalized regression estimator; Intra-class correlation; Sampling variance; Model-assisted; Household surveys.

1. Introduction

Many household surveys involve selecting a sample of households and then selecting all people in the scope of the survey in the selected households. Data on one or more variables of interest are collected for the people in the sample. There may be some auxiliary variables whose population totals and sample values are known; for example these may consist of population counts by geographic and demographic classifications. The generalized regression (GREG) estimator is often used to combine auxiliary information and sample data to efficiently estimate the population totals of the variables of interest.

The GREG estimator makes use of a regression model relating the variable of interest to the auxiliary variables. The standard approach is to fit this model using data for each person in the sample (*e.g.*, Lemaître and Dufour 1987, first paragraph). This person-level GREG estimator is equal to a weighted sum of the sample values of the variable of interest, where the weights are in general different for each person.

It is sometimes convenient to have equal weights for people within a household, for surveys which collect information on both household and person level variables of interest. The same weights can then be used for both types of variables. This ensures that relationships between household variables and person variables are reflected in estimates of total. If a household level variable is equal to the sum of person level variables (for example if household income is the sum of personal incomes), then the estimated total of the household variable will equal the estimated total of the person variable. This is not generally the case where separate weighting procedures are used for person and household variables. Similarly, if there is an inequality relationship between a household level variable and the sum of the person level variables, this will also be reflected in the estimates of the two variables. For example, the estimated number of households using child care centres should not exceed the estimated number of children using centres.

The household-level GREG estimator achieves equal weights within households by fitting the regression model using household totals of the variable of interest and the auxiliary variables (*e.g.*, Nieuwenbroek 1993). Weights with this property are called integrated weights.

An alternative approach would be to use different estimation methods for household-level and person-level variables, and then make an adjustment to force agreement of estimates which should be equal. This approach is sometimes called benchmarking and has mainly been used to achieve consistency between estimates from annual and sub-annual business surveys (*e.g.*, Cholette 1984). A benchmarking approach to household and person-level variables from household surveys would require explicit identification of which person and household-level variables should have equal population totals. In this article we concentrate on integrated weighting and do not consider benchmarking approaches.

Luery (1986); Alexander (1987); Heldal (1992) and Lemaître and Dufour (1987) discussed a number of methods which give integrated weights for person-level and household-level estimates. However, none of these authors evaluated the impact on the sampling variance of calculating the generalized regression estimator at the household level rather than the person level. This is an important issue in practice because the cosmetic benefit of integrated weighting must be balanced against any effect on sampling efficiency.

1. David G. Steel and Robert G. Clark, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: David_Steel@uow.edu.au.

This article compares the design variance, which is the variance over repeated probability sampling from a fixed population, of the person-level and household-level generalized regression estimators. In Section 2, we prove that the large sample variance of the household-level estimator is less than or equal to that of the person-level estimator, by showing that the former is optimal in a large class of GREG estimators. We show that this is because the household-level estimator effectively models contextual effects whereas the person-level estimator does not. In Section 3 the two estimators are compared for a range of variables in a simulation study. Section 4 is a discussion. Three theorems are proved in an Appendix.

2. Theoretical comparison of person and household GREGs

2.1 The generalized regression estimator

In this subsection the generalized regression estimator is described for the general case of probability sampling from any population of units. Let U be a finite population of units and $s \subseteq U$ be the sample. The probabilities of selection are $\pi_i = \Pr[i \in s]$ for units $i \in U$. Let y_i be the variable of interest which is observed for units $i \in s$. Let $\boldsymbol{z_i}$ be the vector of auxiliary variables for unit i, which are observed for every unit in the population. The population totals of these variables are T_Y and $\boldsymbol{T_Z}$ respectively.

The generalized regression estimator of T_Y is based on a model relating the variable of interest to the auxiliary variables:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T \boldsymbol{z_i} \\ &\mathrm{var}_M[y_i] = v_i \sigma^2 \\ &y_i, y_j \text{ independent for } i \neq j \end{aligned}\right\} \quad (1)$$

where v_i are known variance parameters. Subscripts "M" refer to expectations under a model and subscripts "p" refer to design-based expectations, which are expectations over repeated probability sampling from a fixed population. For business surveys collecting continuous variables such as business income and expenses, v_i are often modelled as a function of business size. For household surveys, the variable of interest is often dichotomous, in which case v_i is usually set to 1 corresponding to a homoskedastic model.

Usually $\boldsymbol{z_i}$ have the property that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T \boldsymbol{z_i} = 1$ for all $i \in U$. For example, this is true if the regression model (1) contains an intercept parameter.

Definition 1. generalized regression estimator

The generalized regression estimator for model (1) is defined as

$$\hat{T}_r = \hat{T}_\pi + \hat{\boldsymbol{\beta}}^T (\boldsymbol{T_Z} - \hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}}) \quad (2)$$

where

$$\hat{T}_\pi = \sum_{i \in s} \pi_i^{-1} y_i$$
$$\hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}} = \sum_{i \in s} \pi_i^{-1} \boldsymbol{z_i}.$$

and $\hat{\boldsymbol{\beta}}$ is a solution of

$$\sum_{i \in s} c_i \pi_i^{-1} \left(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{z_i} \right) \boldsymbol{z_i} = \boldsymbol{0}$$

where c_i are regression weights. (Often c_i are set to $c_i = v_i^{-1}$.)

The coefficients $\hat{\boldsymbol{\beta}}$ are calculated from a weighted least squares regression of y_i on $\boldsymbol{z_i}$ for $i \in s$. The GREG estimator has low design variance if the model is approximately true but is design-consistent regardless of the truth of the model (*e.g.*, Särndal, Swensson and Wretman 1992, chapter 6).

For large samples the design variance of $\hat{T}_r$ is approximately equal to

$$\mathrm{var}_p[\hat{T}_r] \approx \mathrm{var}_p[\tilde{T}_r] \quad (3)$$

where

$$\tilde{T}_r = \hat{T}_\pi + \boldsymbol{B}^T (\boldsymbol{T_Z} - \hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}})$$

and $\boldsymbol{B}$ is a solution of

$$\sum_{i \in U} c_i (y_i - \boldsymbol{B}^T \boldsymbol{z_i}) \, \boldsymbol{z_i} = \boldsymbol{0}$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). The coefficients $\boldsymbol{B}$ are calculated from a weighted least squares regression of y_i on $\boldsymbol{z_i}$ for $i \in U$. The sample regression coefficients $\hat{\boldsymbol{\beta}}$ are design-consistent for $\boldsymbol{B}$.

2.2 Person and household level GREGs

We now consider the special case of household sampling, where the basic unit, i, is the person. Let $\boldsymbol{x_i}$ be the p-vector of auxiliary variables observed for all people $i \in U$. The elements of $\boldsymbol{x_i}$ may refer to characteristics of the person or of the household to which they belong. The population and sample of households will be denoted U_1 and s_1 respectively. The population of people in household $g \in U_1$ will be denoted U_g which is of size N_g. Let $y_{g1} = \sum_{i \in U_g} y_i$ and $\boldsymbol{x_{g1}} = \sum_{i \in U_g} \boldsymbol{x_i}$ be the household totals of y_i and $\boldsymbol{x_i}$. Let $\bar{\boldsymbol{x}}_{\boldsymbol{g}} = \boldsymbol{x_{g1}} / N_g$ be the household mean of $\boldsymbol{x_i}$.

We consider the common case where households are selected by probability sampling and all people are selected from selected households, so that $s = \bigcup_{g \in s_1} U_g$. Let

$\pi_{g1} = P[g \in s_1] > 0$ be the probability of selection for household g. It follows that $\pi_i = \pi_{g1}$ for $i \in U_g$.

The person-level GREG, $\hat{T}_P$, is the GREG under the following model:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T \boldsymbol{x_i} \\ &\text{var}_M[y_i] = v_i \sigma^2 \\ &y_i,\ y_j \text{ independent for } i \neq j. \end{aligned}\right\} \quad (4)$$

So the person-level GREG, $\hat{T}_P$, is given by substituting $\boldsymbol{x_i}$ for $\boldsymbol{z_i}$ in (2). Model (4) ignores any correlations between y_i and y_j for people i and j in the same household. These correlations were 0.3 or less in most of the variables considered by Clark and Steel (2002), although higher values occurred for variables related to ethnicity, such as Indigenous self-identification. Correlations of 1 could occur for environmental variables. Tam (1995) shows that the optimal model-assisted estimator for cluster sampling is robust to mis-specification of within-cluster correlations. One way of interpreting this result is that correlations within households are not relevant to estimating population totals, because all people are selected in selected households. So within-household correlations do not help to estimate for non-sample individuals, since the sampled and non-sampled people are in distinct households.

A number of methods have been suggested for GREG-type estimation with equal weights within households. Nieuwenbroek (1993) motivated an estimator by aggregating model (4) to household level:

$$\left.\begin{aligned} &E_M[y_{g1}] = \boldsymbol{\beta}^T \boldsymbol{x_{g1}} \\ &\text{var}_M[y_{g1}] = v_{g1} \sigma^2 \\ &y_{g1},\ y_{k1} \text{ independent for } g \neq k. \end{aligned}\right\} \quad (5)$$

where $v_{g1} = \sum_{i \in U_g} v_i$. The GREG estimator using sample data y_{g1} for $g \in s_1$ based on this model is $\hat{T}_H$:

$$\hat{T}_H = \hat{T}_\pi + \hat{\boldsymbol{\beta}}_{\boldsymbol{H}}^{\boldsymbol{T}} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}}) \quad (6)$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{H}}$ is a solution of

$$\sum_{g \in s_1} \pi_{g1}^{-1} a_g \, (y_{g1} - \hat{\boldsymbol{\beta}}_{\boldsymbol{H}}^{\boldsymbol{T}} \boldsymbol{x_{g1}}) \, \boldsymbol{x_{g1}} = \boldsymbol{0}. \quad (7)$$

The regression coefficient $\hat{\boldsymbol{\beta}}_{\boldsymbol{H}}$ is a household level weighted least squares regression of the sample values of y_{g1} on $\boldsymbol{x_{g1}}$ with weights $\pi_{g1}^{-1} a_g$. The values of a_g could be set to v_{g1}^{-1}. If $v_i = 1$ then $v_{g1} = N_g$ so $a_g = N_g^{-1}$. Alternatively, $a_g = 1$ could also be used.

Several other equivalent integrated weighting methods have been used. Lemaître and Dufour (1987) constructed a generalized regression estimator at person level, using $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ instead of $\boldsymbol{x_i}$ as the auxiliary variables. Nieuwenbroek (1993) commented that this is equivalent to (6) if $c_i = a_g N_g$ for $i \in U_g$. Alexander (1987) developed closely related weighting methods using a minimum distance criterion.

Both the person and household level GREG can be written in weighted form $\sum_{i \in s} w_i Y_i$. The weights for both estimators can be written as $w_i = \pi_i^{-1} g_i$ where

$$g_i = 1 + (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})^T \left(\sum_{i \in s} c_i \pi_i^{-1} \boldsymbol{x_i} \boldsymbol{x_i^T} \right)^{-1} c_i \, \boldsymbol{x_i}$$

for $\hat{T}_P$ and

$$g_i = 1 + (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})^T \left(\sum_{g \in s_1} a_g \, \pi_{g1}^{-1} \boldsymbol{x_{g1}} \boldsymbol{x_{g1}^T} \right)^{-1} a_g \, \pi_g^{-1} \boldsymbol{x_{g1}}$$

for $\hat{T}_H$, where person i belongs to household g. (Superscript "-" stands for generalized inverse of a matrix).

2.3 Theoretical results

In this section, we show that $\hat{T}_H$ has the lowest possible large sample variance in a class of estimators which also includes $\hat{T}_P$, for the sample design where households are selected by simple random sampling without replacement. We will then explain this result by showing that $\hat{T}_H$ is equivalent to a regression estimator calculated using person level data, where the model includes contextual effects.

For large samples, $\hat{T}_P$ and $\hat{T}_H$ can be approximated by

$$\tilde{T}_P = \hat{T}_\pi + \boldsymbol{B_P^T} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}});$$

and

$$\tilde{T}_H = \hat{T}_\pi + \boldsymbol{B_H^T} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})$$

respectively, where $\boldsymbol{B_P}$ and $\boldsymbol{B_H}$ are solutions of

$$\left.\begin{aligned} &\sum_{i \in U} c_i (y_i - \boldsymbol{B_P^T x_i}) \, x_i = \boldsymbol{0} \\ &\sum_{g \in U_1} a_g (y_{g1} - \boldsymbol{B_H^T x_{g1}}) \, \boldsymbol{x_{g1}} = \boldsymbol{0} \end{aligned}\right\} \quad (8)$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). Theorem 1 states the minimum variance estimator in a class including $\tilde{T}_P$ and $\tilde{T}_H$.

Theorem 1. Optimal estimator for simple cluster sampling

Suppose that m households are selected by simple random sampling without replacement from a population of M households, and all people are selected from selected households. Consider the estimator of T given by

$$\tilde{T} = \hat{T}_\pi + \boldsymbol{h}^T (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})$$

where $\boldsymbol{h}$ is a constant p-vector. It is assumed that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda^T x_i} = 1$ for all $i \in U$. The

variance of this estimator is minimised by h^* which are solutions of

$$\sum_{g \in s_1} (y_{g1} - \boldsymbol{h}^T \boldsymbol{x}_{g1}) \boldsymbol{x}_{g1} = \mathbf{0}.$$

Hence $\tilde{T}_H$ with $a_g = 1$ for all g is the optimal choice of $\tilde{T}$.

Theorem 1 has the perhaps surprising implication that $\hat{T}_H$ (with $a_g = 1$ for all g) has lower variance than $\hat{T}_P$ for large samples. This is in spite of the fact that $\hat{T}_H$ discards some of the information in the sample, because it uses the household sums of $\boldsymbol{x}_i$ and y_i. The Theorem suggests that $\hat{T}_H$ is the appropriate GREG estimator for the cluster sampling design assumed here, and that the information discarded by summing to household level is not relevant when this design is used. To explain why $\hat{T}_H$ can perform better than $\hat{T}_P$, we will make use of a "linear contextual model" which is a more general model for $E_M[Y_i]$ than (4). The model is:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\gamma}_1^T \bar{\boldsymbol{x}}_g + \boldsymbol{\gamma}_2^T \boldsymbol{x}_i \quad (i \in U_g) \\ &\mathrm{var}_M[y_i] = \sigma^2 \\ &y_i,\ y_j \text{ independent for } i \neq j. \end{aligned}\right\} \qquad (9)$$

Both $\bar{\boldsymbol{x}}_g$ and $\boldsymbol{x}_i$ are used as explanatory variables for y_i because the household mean of the person level auxiliary variables may capture some of the effect of household context (Lazarfeld and Menzel 1961). For example, if the elements of $\boldsymbol{x}_i$ are indicator variables summarising the age and sex of person i then $\bar{\boldsymbol{x}}_g$ are the proportions of people in the household falling into different age and sex categories. If the population of interest includes both adults and children, then $\bar{\boldsymbol{x}}_g$ includes the proportion of children in the household, which could be relevant to the labour force participation of adults in the household.

Theorem 2 shows that the improvement in the variance from using $\tilde{T}_H$ with $a_g = 1$ rather than using $\tilde{T}_P$ can be explained by the linear contextual model.

Theorem 2. Explaining the difference in the asymptotic variances

Suppose that households are selected by simple random sampling without replacement and all people are selected from selected households. Let $r_i = y_i - \boldsymbol{B}_P^T \boldsymbol{x}_i$, and let $\boldsymbol{B}_C$ be the result of regressing r_i on $\bar{\boldsymbol{x}}_g$ over $i \in U$ using weighted least squares regression weighted by N_g. Then

$$\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \boldsymbol{B}_C^T \left(\sum_{g \in U_1} \boldsymbol{x}_{g1} \boldsymbol{x}_{g1}^T\right) \boldsymbol{B}_C$$

where $\tilde{T}_H$ is calculated using $a_g = 1$ for all g.

The result shows that the reduction in variance from using $\tilde{T}_H$ (with $a_g = 1$) rather than $\tilde{T}_P$ is a quadratic form in $\boldsymbol{B}_C$. Hence the extent of the improvement depends on the extent to which $\bar{\boldsymbol{x}}_g$ helps to predict y_i after $\boldsymbol{x}_i$ has already been controlled for, *i.e.*, the extent to which a linear contextual effect helps to predict r_i over $i \in U$, using a weighted least squares regression weighted by N_g.

The proofs of Theorems 1 and 2 are very much dependent on the assumption of cluster sampling. The results would not be expected to apply if there was subsampling within households.

Theorems 1 and 2 only apply with $a_g = 1$ in the weighted least squares regression for $\hat{T}_H$. Other choices of a_g are often used, for example it would often be reasonable to assume that $v_{g1} = N_g$ in model (5), in which case it would be sensible to use $a_g = N_g^{-1}$. Theorem 3 shows that $\hat{T}_H$ is equivalent to a person-level GREG estimator fitted under the linear contextual model for other choices of a_g.

Theorem 3. The linear contextual GREG

For sample designs where all people are selected from selected households and $\pi_{g1} > 0$ for all $g \in U_1$, $\hat{T}_H$ with a given choice of a_g is the generalized regression estimator for model (9) where $c_i = a_g N_g$ for $i \in U_g$.

Theorem 3 means that $\hat{T}_H$ is the GREG under a more general model than $\hat{T}_P$. Nieuwenbroek (1993) showed that $\hat{T}_H$ is equal to a person-level GREG derived from regressing y_i on $\bar{\boldsymbol{x}}_g$. Theorem 3 states it is also equal to the person-level GREG from regressing y_i on both $\boldsymbol{x}_i$ *and* $\bar{\boldsymbol{x}}_g$, thereby automatically incorporating any household contextual effects. As a result, $\hat{T}_H$ would be expected to have lower variance than $\hat{T}_P$ for large samples. (In the case of $a_g = 1$, Theorem 1 stated that this is always the case). For small samples, however, a more general model may be counter-productive. Silva and Skinner (1997) showed for single-stage sampling that adding parameters to the model can increase the variance of the GREG estimator, although this effect is negligible for large samples. It is possible that the contextual effects have little or no predictive power for some variables. In this case, it would be expected that $\hat{T}_H$ would perform slightly worse than $\hat{T}_P$ for small samples, and about the same for large samples.

The contextual model, (9), includes all of the elements of $\boldsymbol{x}_i$ and all of the elements of $\bar{\boldsymbol{x}}_g$. An alternative would be to use only those elements of either $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_g$ which are significant, or which give improvements in the estimated variance of a GREG estimator. A GREG estimator based on

this type of model would probably have lower variance than the estimators considered in this paper, but would not give integrated weights unless the same elements of $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_g$ were used.

3. Empirical study

3.1 Methodology

A simulation study was undertaken to compare the person and household GREGs, $\hat{T}_P$ and $\hat{T}_H$, for a range of survey variables. We used two populations, consisting of 187,178 households randomly selected from the 2001 Australian Population Census and 210,132 households from the 1995 Australian National Health Survey. All adults and children in the households were included. The average household size was approximately 2.5.

We selected cluster samples from these populations, where households were selected by simple random sampling without replacement and all people from selected households were selected. We simulated samples of size m = 500, 1,000, 2,000, 5,000 and 10,000 households. In each case, 5,000 samples were selected. The auxiliary variables $\boldsymbol{x}_i$ consisted of indicator variables of sex by agegroup (12 categories). (This choice of $\boldsymbol{x}_i$ means that the GREG estimation is equivalent to post-stratification.) The person-level GREG with $c_i = 1(\hat{T}_P)$, the household-level GREG with $a_g = N_g^{-1}(\hat{T}_{H1})$, and the household-level GREG with $a_g = 1(\hat{T}_{H2})$ were all calculated. We also included the Hájek estimator

$$\hat{T}_1 = N\left(\sum_{i\in s}\pi_i^{-1}y_i\right)\Big/\left(\sum_{i\in s}\pi_i^{-1}\right)$$

which equals $N/n\sum_{g\in s_1}\sum_{i\in U_g}y_i$ for cluster sampling with simple random sampling of households, where n is the realized sample size of people.

The variables include labour force, health and other topics. All of the variables are dichotomous except for income (annual income in Australian dollars, based on range data reported from the Census). "Employment(F)" is the indicator variable which is 1 if a person is employed and female, and 0 otherwise. The first six variables are from the Census population and the remaining five variables are from the health population.

3.2 Results

Table 1 shows the relative root mean squared errors (RRMSEs) of $\hat{T}_1$, $\hat{T}_P$, $\hat{T}_{H1}$ and $\hat{T}_{H2}$, for a sample size of 1,000 households. The RRMSEs are expressed as a percentage of the true population total. The biases have not been tabulated because they were a negligible component of the MSE in all cases. The percentage improvements in MSE of $\hat{T}_{H1}$ and $\hat{T}_{H2}$ relative to $\hat{T}_P$ are also shown. The figures in brackets are the simulation standard errors of these percentage improvements.

For this sample size, $\hat{T}_{H1}$ and $\hat{T}_{H2}$ performed slightly worse than $\hat{T}_P$ for the health variables and slightly better for most other variables. The greatest gain was in estimating the number of sole parents; this variance was reduced by 10.8% and 16.3% by using the household-level GREGs. For all other variables, either the improvement was small or the household GREG was slightly worse than the person-level GREG. The inefficiency from using a household-level GREG rather than $\hat{T}_P$ was never more than 2.2%.

Table 2 shows the percentage improvement in MSE from using $\hat{T}_{H1}$ rather than $\hat{T}_P$ for different sample sizes. The simulation standard errors for each figure are shown in brackets. Table 3 shows the percentage improvements from using $\hat{T}_{H2}$ rather than $\hat{T}_P$. The asymptotic percentage improvements ($m = \infty$) are also shown, based on the large sample approximation to the variance of a GREG. For both household-level GREGs, the percentage improvements are generally increasing as the sample size increases. For m = 500, the household GREGs are generally worse than the person GREGs, although never more than 5% worse. For m = 10,000, an improvement is recorded for over half of the variables. The greatest improvements were for estimates of the number of sole parents (11.5%) and employed women (4.2%); all other improvements were small. $\hat{T}_{H1}$ and $\hat{T}_{H2}$ never had variances more than 0.2% higher than $\hat{T}_P$ for m = 10,000. Generally $\hat{T}_{H2}$ performs better than $\hat{T}_{H1}$ for larger sample sizes, as would be expected from Theorem 1, but the reverse is true for small sample sizes.

In practice estimates of subpopulation totals are often of as much interest as population totals. Table 4 shows the performance of the various estimators for age-sex domains (12 age categories) and region domains, for the sample size of 1,000 households. There were 49 regions in the census dataset. The health dataset did not contain a similar region variable, instead the socioeconomic quintile of the collection district (a geographical unit consisting of approximately 200 contiguous households) was used as the domain. The domain estimators were produced by calculating weights from each estimator and taking the weighted sum over the sample in the domain. This is equivalent to the domain ratio estimator described in Case 1, Section 2.1 of Hidiroglou and Patak (2004). We have used this method because it is the most commonly used in practice, as it enables all domains and population totals to be estimated with a single set of weights, although more efficient domain estimators exist (Hidiroglou and Patak 2004, cases 2-6).

In each case, the median RRMSE over the domains is shown. The table shows that there is not much difference

between the three GREG estimators. For age-sex domains, the household GREGs did slightly better than the person GREG for census variables and slightly worse for health variables. For region estimates, the household GREGs were slightly worse in all cases. Table 5 shows that the households GREGs performed very similarly to $\hat{T}_P$ for a sample size of 10,000 households. It is worth noting that Theorem 1 and 2 do not apply to the domain estimators we have used.

Table 1 Relative RMSEs for sample size of 1,000 households

Variable	RRMSE%				% improvement in MSE	
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	2.62	2.09	2.09	2.10	0.20 (0.26)	-0.28 (0.27)
employed F	3.78	3.05	3.01	3.02	2.63 (0.33)	2.09 (0.33)
income	2.56	2.20	2.19	2.19	1.04 (0.25)	0.75 (0.24)
low income	5.04	4.87	4.89	4.90	-0.62 (0.20)	-1.12 (0.22)
hrs worked	3.08	2.54	2.53	2.53	0.94 (0.28)	0.70 (0.28)
sole parent	12.50	12.73	12.02	11.65	10.84 (0.62)	16.31 (0.49)
arthritis	5.52	4.50	4.53	4.53	-1.38 (0.17)	-1.57 (0.18)
smoker	4.73	4.57	4.60	4.61	-1.64 (0.18)	-1.81 (0.20)
high BPR	6.80	5.30	5.35	5.36	-1.70 (0.17)	-2.06 (0.18)
fair/poor hlth	9.79	9.42	9.47	9.47	-1.16 (0.16)	-1.07 (0.18)
alcohol	4.81	4.66	4.70	4.71	-1.77 (0.16)	-2.15 (0.18)

Table 2 Improvement in MSE of household GREG $\hat{T}_{H1}$ compared to $\hat{T}_P$

Variable	% improvement in MSE					
	m = 500	1,000	2,000	5,000	10,000	∞
employed	-0.65 (0.31)	0.20 (0.26)	1.02 (0.24)	0.90 (0.21)	2.17 (0.21)	1.85
employed F	1.22 (0.37)	2.63 (0.33)	2.59 (0.33)	3.53 (0.31)	4.24 (0.31)	4.13
income	-1.53 (0.31)	1.04 (0.25)	0.48 (0.24)	0.61 (0.19)	1.43 (0.19)	1.07
low income	-2.45 (0.27)	-0.62 (0.20)	0.02 (0.18)	0.18 (0.15)	0.00 (0.00)	0.65
hrs worked	-0.26 (0.34)	0.94 (0.28)	1.72 (0.27)	1.61 (0.24)	2.64 (0.24)	2.12
sole parent	7.81 (0.69)	10.84 (0.62)	10.74 (0.61)	10.23 (0.57)	11.50 (0.58)	11.21
arthritis	-3.01 (0.24)	-1.38 (0.17)	-0.34 (0.12)	-0.08 (0.09)	-0.13 (0.07)	0.08
smoker	-3.91 (0.25)	-1.64 (0.18)	-1.02 (0.12)	-0.26 (0.08)	-0.06 (0.07)	0.16
high BPR	-2.93 (0.24)	-1.70 (0.17)	-0.86 (0.12)	-0.31 (0.08)	-0.04 (0.06)	0.08
fair/poor hlth	-3.67 (0.25)	-1.16 (0.16)	-0.71 (0.12)	-0.05 (0.08)	0.03 (0.06)	0.10
alcohol	-4.22 (0.23)	-1.77 (0.16)	-0.77 (0.12)	-0.31 (0.08)	-0.21 (0.07)	0.14

Table 3 Improvement in MSE of household GREG $\hat{T}_{H2}$ compared to $\hat{T}_P$

Variable	% improvement in MSE					
	m = 500	1,000	2,000	5,000	10,000	∞
employed	-1.85 (0.35)	-0.28 (0.27)	1.25 (0.25)	1.05 (0.21)	2.22 (0.21)	1.98
employed F	0.28 (0.39)	2.09 (0.33)	2.71 (0.33)	3.55 (0.29)	4.50 (0.30)	4.31
income	-2.64 (0.31)	0.75 (0.24)	0.71 (0.22)	0.90 (0.17)	1.30 (0.16)	1.37
low income	-3.15 (0.30)	-1.12 (0.22)	-0.15 (0.18)	0.06 (0.15)	0.00 (0.00)	0.94
hrs worked	-1.51 (0.35)	0.70 (0.28)	1.98 (0.25)	1.79 (0.21)	2.57 (0.22)	2.26
sole parent	14.70 (0.53)	16.31 (0.49)	16.39 (0.47)	15.41 (0.44)	16.44 (0.44)	16.35
arthritis	-3.31 (0.26)	-1.57 (0.18)	-0.05 (0.13)	-0.12 (0.09)	-0.10 (0.07)	0.16
smoker	-3.82 (0.28)	-1.81 (0.20)	-0.69 (0.14)	0.21 (0.11)	0.28 (0.10)	0.57
high BPR	-3.20 (0.26)	-2.06 (0.18)	-1.12 (0.13)	-0.40 (0.09)	-0.05 (0.07)	0.12
fair/poor hlth	-4.02 (0.28)	-1.07 (0.18)	-0.57 (0.13)	-0.09 (0.09)	0.00 (0.07)	0.15
alcohol	-5.00 (0.26)	-2.15 (0.18)	-0.82 (0.13)	-0.49 (0.09)	-0.29 (0.08)	0.18

Table 4 Median relative RMSEs for domain estimators for sample size m = 1,000

Variable	Age-Sex Domains				Region Domains			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	12.74	7.92	7.93	7.90	29.89	29.92	30.20	30.34
employed F	13.12	8.32	8.36	8.34	34.64	34.65	35.03	35.16
income	13.25	8.43	8.49	8.47	28.04	28.12	28.43	28.51
low income	21.17	18.77	18.96	18.94	42.71	42.85	43.24	43.33
hrs worked	14.56	10.69	10.76	10.72	31.24	31.23	31.52	31.63
sole parent	96.20	96.33	97.64	96.69	92.99	93.30	94.37	93.50
arthritis	24.94	20.94	21.12	21.11	13.31	12.94	13.02	13.04
smoker	32.10	29.25	29.39	29.37	12.32	12.27	12.35	12.38
high BPR	27.01	23.80	23.97	23.95	15.83	15.31	15.44	15.45
fair/poor hlth	39.64	37.73	38.05	38.08	22.38	22.30	22.51	22.55
alcohol	25.58	21.42	21.53	21.58	12.73	12.70	12.80	12.82

Table 5 Median relative RMSEs for domain estimators for sample size m = 10,000

Variable	Age-Sex Domains				Region Domains			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	3.77	2.35	2.32	2.31	8.85	8.85	8.87	8.88
employed F	3.86	2.43	2.43	2.42	10.30	10.26	10.25	10.25
income	3.91	2.53	2.51	2.51	8.24	8.23	8.23	8.24
low income	6.31	5.63	5.62	5.61	12.67	12.68	12.69	12.69
hrs worked	4.29	3.15	3.15	3.12	9.26	9.25	9.27	9.27
sole parent	28.40	28.26	28.29	28.23	27.11	27.14	27.16	27.11
arthritis	7.40	6.26	6.27	6.27	3.98	3.85	3.85	3.85
smoker	9.53	8.58	8.58	8.57	3.69	3.67	3.68	3.67
high BPR	8.07	7.02	7.01	7.01	4.66	4.48	4.49	4.49
fair/poor hlth	11.69	11.02	11.02	11.01	6.75	6.69	6.69	6.69
alcohol	7.74	6.43	6.43	6.43	3.87	3.85	3.85	3.85

4. Discussion

The standard person-level GREG estimator produces unequal weights within households. Household-level GREG estimators can be used to give integrated household and person weights, which is beneficial for surveys collecting information on both household-level and person-level variables. This article demonstrated that there is little or no loss associated with the practical benefit of integrated weighting arising from using a household-level GREG estimator. For large samples, the household-level GREG has lower design variance than the person-level GREG. For smaller samples there is at most a small increase in variance for some variables from using the household GREG, because this estimator is equivalent to using a regression model containing more parameters. Therefore, if integrated weights would improve the coherence of a household survey's outputs, the household-level GREG can be adopted with little or no detriment to the variance and bias of estimators.

Acknowledgements

This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics. The views expressed here do not necessarily reflect the views of either organisation. The authors thank Julian England, Frank Yu and Ray Chambers for their thoughtful comments.

Appendix

Proof of theorems

Proof of theorem 1

Let $\bar{Y}_1 = T_Y / M$ and $\bar{\boldsymbol{X}}_1 = \boldsymbol{T}_X / M$ be the population means of y_{g1} and $\boldsymbol{x}_{g1}$ respectively. The variance of $\tilde{T}$ is

$$\begin{aligned}\mathrm{var}_p[\tilde{T}] &= \mathrm{var}[\hat{T}_\pi + \boldsymbol{h}^T(\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})] \\ &= \mathrm{var}\left[\frac{M}{m}\sum_{g\in s_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\right] \\ &= \frac{M^2}{m}\left(1 - \frac{m}{M}\right)S_r^2\end{aligned}$$

where $S_r^2 = (M-1)^{-1}\sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}^2$. To minimise with respect to $\boldsymbol{h}$, we set the derivative of S_r^2 to zero:

$$
\begin{aligned}
0 &= (M-1)^{-1}\sum_{g\in U_1} \\
&\quad \{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1) \\
0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,\boldsymbol{x}_{g1} \\
&\quad -\sum_{g\in U_1}\{(y_{g1} - \bar{Y}_1) - \boldsymbol{h}^T(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1)\}\bar{\boldsymbol{X}}_1 \\
0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\boldsymbol{x}_{g1} \\
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\,\boldsymbol{T}_X. \qquad (10)
\end{aligned}
$$

We now show that (10) is satisfied by $\boldsymbol{h}^*$. By assumption, $\boldsymbol{h}^*$ satisfies

$$\boldsymbol{0} = \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{x}_{g1}. \qquad (11)$$

Hence the first sum in the right hand side of (10) is equal to zero for $\boldsymbol{h} = \boldsymbol{h}^*$. Premultiplying both sides of (11) by $\boldsymbol{\lambda}^T$ gives

$$
\begin{aligned}
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{\lambda}^T\boldsymbol{x}_{g1} \\
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*) \\
0 &= T_Y - \boldsymbol{T}_X^T\boldsymbol{h}^*.
\end{aligned}
$$

Dividing by M gives $\bar{Y}_1 - \bar{\boldsymbol{X}}_1^T\boldsymbol{h}^* = 0$. Hence the rest of the right hand side of (10) is equal to zero. So $\boldsymbol{h}^*$ satisfies (10).

Proof of theorem 2

Let "-" denote a generalized inverse of a matrix. Then $\boldsymbol{B}_C$ is equal to

$$
\begin{aligned}
\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g\bar{\boldsymbol{x}}_g^T\right\}^-\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g r_i \\
&= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}r_{g1}. \qquad (12)
\end{aligned}
$$

Now, $r_i = y_i - \boldsymbol{B}_P^T\boldsymbol{x}_i$ so $r_{g1} = y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}$. Hence (12) becomes

$$
\begin{aligned}
\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}) \\
&= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}y_{g1} \\
&\quad -\left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_P \\
&= \boldsymbol{B}_H - \boldsymbol{B}_P \qquad (13)
\end{aligned}
$$

since $\boldsymbol{B}_H = \{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\}\sum_{g\in U_1}\sum_{i\in U_g}\boldsymbol{x}_{g1}y_{g1}$. The difference in the variances is given by

$$
\begin{aligned}
&\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \\
&\qquad \left\{\sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2\right\}
\end{aligned}
$$

which becomes

$$
\begin{aligned}
&\{\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H]\}\Big/\left\{\frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\right\} \\
&= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} + \boldsymbol{B}_P^T\boldsymbol{x}_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1} + \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 + \sum_{g\in U_1}(\boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&\quad + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C^T \\
&\quad - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}\boldsymbol{B}_C^T\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_C + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C. \qquad (14)
\end{aligned}
$$

Now, $\boldsymbol{B}_C$ is an ordinary least squares regression of r_{g1} on $\boldsymbol{x}_{g1}$ so

$$\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} = \boldsymbol{0}.$$

Hence (14) becomes

$$
\begin{aligned}
&\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \\
&\qquad \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B}_C^T\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_C.
\end{aligned}
$$

Proof of Theorem 3

The GREG estimator is invariant under linear invertible transformations of the auxiliary variables. Hence model (9) can be re-parameterised to give

$$E_M[y_i] = \boldsymbol{\phi}_1^T \bar{\boldsymbol{x}}_g + \boldsymbol{\phi}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g) \tag{15}$$

or equivalently

$$E_M[y_i] = \boldsymbol{\phi}^T \boldsymbol{z}_i$$

where

$$\boldsymbol{z}_i = \begin{pmatrix} \bar{\boldsymbol{x}}_g \\ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_g \end{pmatrix}$$

and

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix}.$$

The parameters in model (15) are related to those in model (9) by $\boldsymbol{\phi}_1 = \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2$ and $\boldsymbol{\phi}_2 = \boldsymbol{\gamma}_2$.

From Definition 1, noting that

$$s = \bigcup_{g \in s_1} U_g$$

for the assumed design, the generalized regression estimator under model (15) is

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{i \in U} \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i - \sum_{i \in s} \pi_i^{-1} \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i \\ &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \{\hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g + \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ &\quad - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \{\hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g + \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\}. \end{aligned} \tag{16}$$

However, $\sum_{i \in U_g} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g) = \boldsymbol{0}$ for each g. Hence (16) becomes

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T \sum_{g \in U_1} \sum_{i \in U_g} \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_1^T \sum_{g \in s_1} \pi_{g1}^{-1} \sum_{i \in U_g} \bar{\boldsymbol{x}}_g \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T \sum_{g \in U_1} \bar{\boldsymbol{x}}_{g1} - \hat{\boldsymbol{\phi}}_1^T \sum_{g \in s_1} \pi_{g1}^{-1} \bar{\boldsymbol{x}}_{g1} \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi}). \end{aligned} \tag{17}$$

Notice that (17) does not include the estimator of $\boldsymbol{\phi}_2$. The least squares estimators

$$\hat{\boldsymbol{\phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_1 \\ \hat{\boldsymbol{\phi}}_2 \end{pmatrix}$$

are the solution of:

$$\sum_{i \in s} \pi_i^{-1} c_i (y_i - \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i)\, \boldsymbol{z}_i = \boldsymbol{0}$$

which is equivalent to:

$$\sum_{i \in s} \pi_i^{-1} c_i \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_1^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \begin{pmatrix} \bar{\boldsymbol{x}}_g \\ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_g \end{pmatrix} = \boldsymbol{0}.$$

By assumption, $c_i = a_g N_g$ so the first p elements of this equation are:

$$\begin{aligned} \boldsymbol{0} &= \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} a_g N_g \bar{\boldsymbol{x}}_g \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g N_g \bar{\boldsymbol{x}}_g \sum_{i \in U_g} \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g \boldsymbol{x}_{g1} \{y_{g1} - \hat{\boldsymbol{\phi}}_1^T \boldsymbol{x}_{g1} - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_{g1} - \boldsymbol{x}_{g1})\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g \boldsymbol{x}_{g1} (y_{g1} - \hat{\boldsymbol{\phi}}_1^T \boldsymbol{x}_{g1}). \end{aligned}$$

Hence $\hat{\boldsymbol{\phi}}_1$ is a solution to (7). So the GREG estimator for model (9) is equal to $\hat{T}_H$ provided that $c_i = a_g N_g$.

References

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Cholette, P. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10, 35-49.

Clark, R.G., and Steel, D.G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, 70 (2), 289-314.

Heldal, J. (1992). A method for calibration of weights in sample surveys. In *Workshop on uses of auxiliary information in surveys*. University of Orebro, Sweden.

Hidiroglou, M., and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67-78.

Lazarfeld, P.F., and Menzel, H. (1961). On the relation between individual and collective properties. In *Complex Organizations*: *A Sociological Reader*. Holt, Reinhart and Winston. 422-440.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

Luery, D.M. (1986). Weighting sample survey data under linear constraints on the weights. In *Proceedings of the Social Statistics Section*, American Statistical Association, (Alexandria, VA), 325-330.

Nieuwenbroek, N. (1993). *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Netherlands Central Bureau of Statistics.

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

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

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 61-68
Statistics Canada, Catalogue No. 12-001

Mean - Adjusted bootstrap for two - Phase sampling

Hiroshi Saigo [1]

Abstract

Two-phase sampling is a useful design when the auxiliary variables are unavailable in advance. Variance estimation under this design, however, is complicated particularly when sampling fractions are high. This article addresses a simple bootstrap method for two-phase simple random sampling without replacement at each phase with high sampling fractions. It works for the estimation of distribution functions and quantiles since no rescaling is performed. The method can be extended to stratified two-phase sampling by independently repeating the proposed procedure in different strata. Variance estimation of some conventional estimators, such as the ratio and regression estimators, is studied for illustration. A simulation study is conducted to compare the proposed method with existing variance estimators for estimating distribution functions and quantiles.

Key Words: Double Sampling; Resampling; Variance estimation.

1. Introduction

Two-phase sampling or double sampling is a powerful tool for efficient estimation in surveys. Usually, a large-scale first phase sample is taken where auxiliary variables, correlated with the characteristics of interest and relatively easily obtained, are observed. Then, a small-scale sub-sample is chosen from the first phase sample to measure the characteristics of interest that are harder to obtain. At the estimation stage, the auxiliary variables at the first phase are employed to obtain an efficient estimator.

A closed-form sample variance formula for an estimator can be complicated or even unavailable under two-phase sampling. Consequently, resampling methods, such as the jackknife and bootstrap, are appealing for two-phase sampling. Rao and Sitter (1995) and Sitter (1997) studied the delete-1 jackknife approach to the ratio and regression estimators under two-phase sampling and found the method provides design-consistent variance estimation with desirable conditional properties given the auxiliary variables.

A weakness of the delete-1 jackknife is that it cannot handle quantile estimation. Moreover, it is not trivial how one can incorporate the finite population correction into the jackknife variance estimation under two-phase sampling (see Lee and Kim 2002 and Berger and Rao 2006). The bootstrap, on the other hand, eliminates these problems if properly formulated.

Several bootstrap methods for two-phase sampling have been proposed and studied. Schreuder, Li and Scott (1987), Biemer and Atkinson (1993) and Sitter (1997) considered similar bootstrap methods which provide consistent variance estimation when sampling fractions are negligible. Rao and Sitter (1997) proposed a rescaling bootstrap for high sampling fractions.

A disadvantage of the rescaling approach is that it cannot handle the estimation of distribution functions and quantiles. In this paper, we propose a mean-adjusted bootstrap for two-phase sampling that accommodates the estimation of distribution functions and quantiles. The method is simple and includes the existing ones for negligible sampling fractions as a special case. Recently, Kim, Navarro, and Fuller (2006) studied replication variance estimation without rescaling for two-phase sampling in a more generalized framework than that of this paper. Our method, however, is different in that it internally incorporates the finite population correction.

This paper is organized as follows. Section 2 presents the mean-adjusted bootstrap for two-phase sampling. Section 3 illustrates how the proposed method works for some conventional estimators. A simulation for estimating distribution functions and quantiles is conducted in Section 4. Section 5 discusses further applications of the mean-adjusted bootstrap. Concluding remarks are given in Section 6.

2. Mean - Adjusted bootstrap

For notational simplicity, we assume there is only one stratum. To extend our method to stratified sampling, repeat the same procedure independently in different strata to obtain a bootstrap sample (see Rao and Sitter 1997, pages 759-762).

Let P be the set of unit labels in a population of size N. Suppose a simple random sample without replacement (SRSWOR) of size n_{A+B} from P is taken and denote the sampled labels by $A+B$. The auxiliary variable (vector) x_i is observed for $i \in A+B$. Then take a second phase SRSWOR of size $n_A < n_{A+B}$ from $A+B$ and denote the sampled labels by A. The characteristic (vector) y_i is

1. Hiroshi Saigo, Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku Tokyo 169-8050, Japan.

measured for $i \in A$. Let $B = (A + B) - A$, $n_B = n_{A+B} - n_A$, $\mathbf{y}_A = \{y_i : i \in A\}$, $\mathbf{x}_A = \{x_i : i \in A\}$, and $\mathbf{x}_B = \{x_j : j \in B\}$. An approximately design-unbiased estimator of parameter θ is assumed to be written as $\hat{\theta} = t(\mathbf{y}_A, \mathbf{x}_A, \mathbf{x}_B)$.

Under the proposed method, a bootstrap sample is constructed as follows.

1. Regard A as an SRSWOR of size n_A from P. Choose n_A units from A by a bootstrap method suitable for an SRSWOR of size n_A from P. Denote the sampled labels by A^*.
2. Regard B as an SRSWOR of size n_B from $P - A$ conditional on A having been selected. Choose n_B units from B by a bootstrap method suitable for an SRSWOR of size n_B from $P - A$. Denote the sampled labels by B^*.
3. For $j \in B^*$, define the mean-adjustment as $\tilde{x}_j$, where
$$\tilde{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_{A^*})/(1 - f_A), \qquad (1)$$
with $\bar{x}_A = n_A^{-1}\sum_{i\in A} x_i$, $\bar{x}_{A^*} = n_A^{-1}\sum_{i\in A^*} x_i$, and $f_A = n_A / N$.
4. Let $\mathbf{y}_{A^*} = \{y_i : i \in A^*\}$, $\mathbf{x}_{A^*} = \{x_i : i \in A^*\}$, and $\tilde{\mathbf{x}}_{B^*} = \{\tilde{x}_j : j \in B^*\}$. The bootstrap analogue of $\hat{\theta}$ is then given by $\hat{\theta}^* = t(\mathbf{y}_{A^*}, \mathbf{x}_{A^*}, \tilde{\mathbf{x}}_{B^*})$.

For bootstrap methods for a finite population, see Shao and Tu (1995, Chapter 6). The Bernoulli Bootstrap (BBE) proposed by Funaoka, Saigo, Sitter and Toida (2006) is appropriate for our method because of a reason specified later. To obtain a bootstrap sample A^* in the BBE, we conduct random replacement for each i in A: keep (x_i, y_i) in the bootstrap sample with probability $p = \{1 - (1 - n_A^{-1})^{-1}(1 - f_A)\}^{1/2}$ or replace it with one randomly selected from A. For the case where $p \notin [0, 1]$, see Funaoka *et al.* (2006).

To estimate the variance of $\hat{\theta}$, repeat steps 1-4 a large number of times K and use
$$v_{\text{boot}}(\hat{\theta}) = K^{-1}\sum_{k=1}^{K}(\hat{\theta}^*_{(k)} - \hat{\theta}^*_{(\cdot)})^2, \qquad (2)$$
where $\hat{\theta}^*_{(k)}$ is the value of $\hat{\theta}^*$ in the k^{th} bootstrap sample and $\hat{\theta}^*_{(\cdot)} = K^{-1}\sum_k \hat{\theta}^*_{(k)}$.

When f_A is negligible, the mean adjustment (1) is unnecessary. The above method then reduces for large n_A to that by Schreuder *et al.* (1987) and Sitter (1997).

The proposed bootstrap method is motivated by the following two observations. First, let sampling schemes I and II be $[P \to A + B,\ A + B \to A]$ and $[P \to A,\ P - A \to B]$, respectively, where $\to$ means "the right hand side is an SRSWOR from the left hand side." Then, I and II implement the identical design. In fact, the design probability assigned to a particular sample $\{\mathbf{i} = (i_1, i_2, \ldots, i_{n_A}) \in A,\ \mathbf{j} = (j_1, j_2, \ldots, j_{n_B}) \in B\}$ in I is $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_{A+B}} \times {}_{n_{A+B}} C_{n_A}]^{-1} = n_A!\, n_B!\,(N - n_{A+B})!/N!$ while it is $\Pr\{\mathbf{i} \in A,\ \mathbf{j} \in B\} = [{}_N C_{n_A} \times {}_{N-n_A} C_{n_B}]^{-1} = n_A!\, n_B!\,(N - n_{A+B})!/N!$ in II. Obviously, the sampling distribution of an estimator under repeated sampling depends on the sampling design. So, it is a matter of convenience to assume II is carried out even when I is employed.

Second, to motivate the mean adjustment (1), observe that the mean of x of the set $P - A$, or the conditional expectation of $\bar{x}_B$ under repeated sampling given A, is $\bar{X}_{P-A} = (\bar{X} - f_A\bar{x}_A)/(1 - f_A)$. The bootstrap analogue of $\bar{X}_{P-A}$ is given by $\bar{X}_{P-A^*} = (\bar{X} - f_{A^*}\bar{x}_{A^*})/(1 - f_{A^*})$. So, equation (1) amounts to $\tilde{x}_j = x_j - \bar{X}_{P-A} + \bar{X}_{P-A^*}$, a mean adjustment similar to that proposed by Rao and Shao (1992) in the context of hot deck imputation under the uniform response mechanism. This mean adjustment ensures appropriate correlations between x in A^* and x in B^* required for consistent variance estimation with high sampling fractions (see Rao and Sitter 1997, page 760). Note that the condition $n_A = n_{A^*}$ or $f_A = f_{A^*}$ is essential for cancelling out $\bar{X}$ in the mean adjustment. Therefore, the mean-adjusted bootstrap requires a bootstrap method for SRSWOR which retains the original sample size, such as the BBE.

It is shown in Appendix A that the proposed bootstrap method provides design-consistent variance estimation for the class of estimators studied by Rao and Sitter (1997). Since no rescaling is performed, the method also works for estimation of distribution functions. Under some regularity conditions for the population distribution function, it provides design-consistent variance estimates for quantiles.

3. Illustrations

3.1 Ratio estimator

To illustrate, let us first consider the ratio estimator $\bar{y}_r = r_A\bar{x}_{A+B}$, where $r_A = \bar{y}_A/\bar{x}_A$, $w_A = n_A/n_{A+B}$, and $\bar{x}_{A+B} = w_A\bar{x}_A + (1 - w_A)\bar{x}_B$. Let $\bar{y}^*_r = (\bar{y}_{A^*}/\bar{x}_{A^*})\{w_A\bar{x}_{A^*} + (1 - w_A)\bar{x}_{B^*}\}$, the bootstrap analogue of $\bar{y}_r$. Using the results in Appendix A with $h(\bar{y}_A, \bar{x}_A, \bar{x}_B) = (\bar{y}_A/\bar{x}_A)\{w_A\bar{x}_A + (1 - w_A)\bar{x}_B\}$, we may approximate variance of $\bar{y}^*_r$ under the proposed bootstrap method $V_*(\bar{y}^*_r)$ by

$$V_*(\bar{y}^*_r) \doteq (\bar{x}_{A+B}/\bar{x}_A)^2 \frac{(1 - f_A)}{n_A}\hat{S}^2_{dA} + 2(\bar{x}_{A+B}/\bar{x}_A)\frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2\left[\frac{(w_A - f_A)}{(1 - f_A)}\hat{S}^2_{xA} + \frac{(1 - w_A)}{(1 - f_A)}\hat{S}^2_{xB}\right], \qquad (3)$$

where $\hat{S}_{dA}^2 = (n_A - 1)^{-1}\sum_{i\in A}(y_i - r_A x_i)^2$, $\hat{S}_{dxA} = (n_A - 1)^{-1}$ $\sum_{i\in A}(y_i - r_A x_i)(x_i - \bar{x}_A)$, $\hat{S}_{xA}^2 = (n_A - 1)^{-1}\sum_{i\in A}(x_i - \bar{x}_A)^2$, and $\hat{S}_{xB}^2 = (n_B - 1)^{-1}\sum_{i\in B}(x_i - \bar{x}_B)^2$. The right hand side of (3) can be described as a "bootstrap-linearization" variance estimator. We denote it by $v_{\mathrm{BL}}(\bar{y}_r)$. Note that $v_{\mathrm{BL}}(\bar{y}_r)$ is almost identical to the jackknife-linearization variance estimator by Rao and Sitter (1995),

$$v_{\mathrm{JL}}(\bar{y}_r) = (\bar{x}_{A+B}/\bar{x}_A)^2 \frac{(1-f_A)}{n_A}\hat{S}_{dA}^2 + 2(\bar{x}_{A+B}/\bar{x}_A)\frac{(1-f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1-f_{A+B})}{n_{A+B}} r_A^2 \hat{S}_{xA+B}^2, \quad (4)$$

where $\hat{S}_{xA+B}^2 = (n_{A+B} - 1)^{-1}\sum_{A+B}(x_i - \bar{x}_{A+B})^2$, which agrees with equation 4.8 of Demnati and Rao (2004), page 25. Since they are close to $v_{\mathrm{JL}}(\bar{y}_r)$, $V_*(\bar{y}_{lr})$, its Monte Carlo approximation $v_{\mathrm{boot}}(\bar{y}_r^*)$ and $v_{\mathrm{BL}}(\bar{y}_{lr})$ should perform well not only unconditionally but conditionally on $(\bar{x}_{A+B}/\bar{x}_A)$ as well. It is interesting to note that Taylor linearization in deriving $v_{\mathrm{BL}}(\bar{y}_r)$ is performed around the sample means, not the population means (see the comment made by Demnati and Rao 2004, page 21).

3.2 Regression estimator

We next consider the regression estimator. The estimator of the population mean is $\bar{y}_{lr} = \bar{y}_A + b_A(\bar{x}_{A+B} - \bar{x}_A) = \bar{y}_A + (1 - w_A)\, b_A(\bar{x}_B - \bar{x}_A)$, where $b_A = \hat{S}_{xyA}/\hat{S}_{xA}^2$ with $\hat{S}_{xyA} = (n_A - 1)^{-1}\sum_{i\in A}(x_i - \bar{x}_A)(y_i - \bar{y}_A)$. Let $\bar{y}_{lr}^* = \bar{y}_{A^*} + (1 - w_A)\, b_{A^*}(\bar{\bar{x}}_{B^*} - \bar{x}_{A^*})$. Using the results in Appendix A (see also Appendix B), we have

$$V_*(\bar{y}_{lr}^*) \doteq \frac{(1-f_A)}{n_A} m_{02} + \frac{(1-f_{A+B})}{n_{A+B}} b_A^2\left[\frac{(w_A - f_A)}{(1-f_A)}\hat{S}_{xA}^2 + \frac{(1-w_A)}{(1-f_A)}\hat{S}_{xB}^2\right] + z_A^2\frac{(1-f_A)}{n_A} m_{22} + 2z_A\frac{(1-f_A)}{n_A} m_{12} + 2z_A\frac{(1-f_{A+B})}{n_{A+B}} b_A m_{21} + 4z_A^2\frac{(1-f_A)}{n_A} a_A b_A \bar{x}_A \hat{S}_{xA}^2, \quad (5)$$

where $z_A = n_A(\bar{x}_{A+B} - \bar{x}_A)/\{(n_A - 1)\hat{S}_{xA}^2\}$, $m_{pq} = (n_A - 1)^{-1}\sum_{i\in A}(x_i - \bar{x}_A)^p e_i^q$, $e_i = y_i - \bar{y}_A - b_A(x_i - \bar{x}_A)$, and $a_A = \bar{y}_A - b_A\bar{x}_A$. We call the right hand side of (5) a bootstrap-linearization variance estimator of $\bar{y}_{lr}$ and denote it by $v_{\mathrm{BL}}(\bar{y}_{lr})$. The jackknife-linearization variance estimator for $\bar{y}_{lr}$ (Sitter 1997, page 781) is

$$v_{\mathrm{JL}}(\bar{y}_{lr}) = \frac{(1-f_A)}{n_A} m_{02} + \frac{(1-f_{A+B})}{n_{A+B}} b_A^2 \hat{S}_{xA+B}^2 + \frac{z_A^2}{n_A^2}\sum_{i\in A}\frac{(x_i - \bar{x}_A)^2 e_i^2}{(1-c_i)^2} + \frac{2z_A}{n_A^2}\sum_{i\in A}\frac{(x_i - \bar{x}_A)e_i^2}{(1-c_i)} + \frac{2z_A b_A}{n_A(n_{A+B}-1)}\sum_{i\in A}\frac{(x_i - \bar{x}_A)(x_i - \bar{x}_{A+B})e_i}{(1-c_i)}, \quad (6)$$

where $c_i = n_A^{-1} + (x_i - \bar{x}_A)^2/\{(n_A - 1)\hat{S}_{xA}^2\}$, the leverage values. From (5) and (6), $v_{\mathrm{boot}}(\bar{y}_{lr})$, $v_{\mathrm{BL}}(\bar{y}_{lr})$ and $v_{\mathrm{JL}}(\bar{y}_{lr})$ perform in a similar fashion conditionally provided that $f_{A+B} \doteq 0$, n_A is large enough for all c_i to be nearly zero and the last term on the right hand side of (5) is negligible.

3.3 Estimation of distribution functions

As an example, let us take the model-calibrated pseudo-empirical maximum likelihood estimator (ME) under two-phase sampling proposed by Wu and Luan (2003) defined by

$$\hat{F}_{\mathrm{ME}}(t) = \sum_{i\in A}\hat{p}_i I(y_i \le t), \quad (7)$$

where $\hat{p}_i$ maximizes the pseudo-likelihood function $\hat{l}(p) = \sum_A (N/n_A)\log p_i$ subject to (a) $\sum_A p_i = 1 (0 < p_i < 1)$; and (b) $\sum_A p_i g_i = n_{A+B}^{-1}\sum_{A+B} g_i$ where $g_i = g(x_i, t) = P(y \le t\,|\,x_i)$ under a certain working model. For example, we may assume $\log(g_i/(1-g_i)) = x_i'\theta$ with variance function $V(g) = g(1-g)$. Chen, Sitter and Wu (2002) showed a simple algorithm for computing $\hat{p}_i$. It can be shown (see Wu and Luan 2003) that under the two-phase sampling considered in this paper,

$$\hat{F}_{\mathrm{ME}}(t) = n_A^{-1}\sum_{i\in A} I(y_i \le t) + \left\{n_{A+B}^{-1}\sum_{i\in A+B} g_i - n_A^{-1}\sum_{i\in A} g_i\right\}\beta + o_p(n_A^{-1/2}),$$

where $\beta = \sum_P (g_i - \bar{g})I(y \le t)/\sum_P (g_i - \bar{g})^2$ with $\bar{g} = N^{-1}\sum_P g_i$. Note that this equation is not used in estimation, but it shows that the variance of $\hat{F}_{\mathrm{ME}}(t)$ can be estimated by the mean-adjusted bootstrap since $\hat{F}_{\mathrm{ME}}(t)$ is approximated by a regression-type estimator.

3.4 Quantile estimation

Quantile estimation can be obtained by directly inverting $\hat{F}(t)$ by $\hat{F}^{-1}(\alpha) = \inf\{t : \hat{F}(t) \ge \alpha\}$ for some $\alpha \in (0, 1)$. For example, if (7) is used, then a quantile estimate is given by $y_{(k)}$, where $y_{(k)}$ is the k^{th} order statistic of y such that $\sum_{i=1}^{k-1}\hat{p}_{(i)} < \alpha$ and $\sum_{i=1}^{k}\hat{p}_{(i)} \ge \alpha$ (Chen and Wu 2002). Under some conditions specified in Chen and Wu (2002), a

Bahadur-type representation for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ can be established. Thus the mean-adjusted bootstrap variance estimator for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ is design-consistent. Note that no closed form variance estimator for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ is available, but a consistent variance estimator based on Woodruff's interval estimation (Woodruff 1952) can be applied.

4. Simulation

4.1 Population and sampling

A simulation study was conducted to examine the mean-adjusted bootstrap variance estimator for the estimators in Section 3. We report here the results for estimating distribution functions and quantiles. The results for the ratio and regression estimators are available from the author upon request.

First, the auxiliary variable x for a finite population P of size $N = 2{,}000$ were generated as Gamma(1, 1). The characteristic variable y was then generated by $y_i = x_i + \sqrt{x_i}\,v_i$, where $v_i \sim N(0,\ 0.5^2)$. An SRSWOR $A + B$ of size $n_{A+B} = 800$ was taken from the population and then an SRSWOR A of size $n_A = 200$ was selected from $A + B$. The population was fixed throughout all simulation runs since we focus on design-based repeated-sampling properties.

4.2 Estimation of distribution functions

For the estimation of distribution functions, we took $\hat{F}_{\mathrm{ME}}(t)$ as an example. Other estimators, *e.g.*, Chambers and Dunstan (1986) and Rao, Kovar and Mantel (1990), can be handled similarly when an estimator is approximately design-unbiased. The working model for g in $\hat{F}_{\mathrm{ME}}(t)$ was assumed to be logit with binomial variance. The bootstrap variance estimator $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ was calculated with $K = 200$. The BBE was used in constructing a bootstrap sample. The total simulation runs were $M = 5{,}000$ while the true MSE of $\hat{F}_{\mathrm{ME}}(t)$ at a given t was estimated by 50,000 runs.

We compared $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ with three variance estimators: Wu and Luan's (2003) analytical estimator, the standard delete-1 jackknife and an *ad hoc* fpc-adjusted delete-1 jackknife. Wu and Luan's (2003) estimator is

$$v_a(\hat{F}_{\mathrm{ME}}(t)) = (n_{A+B}^{-1} - N^{-1})\,\hat{S}_I^2 + (n_A^{-1} - n_{A+B}^{-1})\,\hat{S}_D^2,$$

where the two $\hat{S}^2$ components are estimated respectively by

$$\hat{S}^2 = s^2 + \left[\frac{1}{n_{A+B}(n_{A+B}-1)}\sum_{j>i:i,\,j\in A+B} u_{ij} - \frac{1}{n_A(n_A-1)}\sum_{j>i:i,\,j\in A} u_{ij}\right]\hat{\beta}_F,$$

where $s^2 = \{n_A(n_A-1)\}^{-1}\sum_{i<j:i,\,j\in A} v_{ij}$, and $\hat{\beta}_F = \sum_{i<j:i,\,j\in A} u_{ij}v_{ij} / \sum_{i<j:i,\,j\in A} u_{ij}^2$ with u_{ij} and v_{ij} specified as follows: For $\hat{S}_I^2$, $v_{ij} = (I_i - I_j)^2$ and $u_{ij} = (\hat{g}_i - \hat{g}_j)^2$ with $I_i = I(y_i \le t)$ and $\hat{g}_i = \hat{g}(x_i,\ t)$ estimated in A; For $\hat{S}_D^2$, $v_{ij} = (\hat{D}_i - \hat{D}_j)^2$ and $u_{ij} = \hat{g}_i(1-\hat{g}_i) + \hat{g}_j(1-\hat{g}_j)$ with $\hat{D}_i = I_i - \hat{g}_i\hat{\beta}$, $\hat{\beta} = \sum_{i\in A} I_i(\hat{g}_i - \bar{\hat{g}}_A)/\sum_{i\in A}(\hat{g}_i - \bar{\hat{g}}_A)^2$ and $\bar{\hat{g}}_A = n_A^{-1}\sum_{i\in A}\hat{g}_i$.

The standard delete-1 jackknife formula is given by

$$v_J(\hat{\theta}) = \frac{(n_{A+B}-1)}{n_{A+B}}\sum_{j\in A+B}(\hat{\theta}_{(-j)} - \hat{\theta}_{(\cdot)})^2,$$

where $\hat{\theta} = \hat{F}_{\mathrm{ME}}(t)$, $\hat{\theta}_{(-j)}$ is the j^{th} jackknife pseudo-estimate and $\hat{\theta}_{(\cdot)} = n_{A+B}^{-1}\sum_{j\in A+B}\hat{\theta}_{(-j)}$. Note that for $j \in A$, both y_j and x_j are deleted from the sample while for $j \in B$, only x_j is deleted (see Rao and Sitter 1995 and Sitter 1997). The *ad hoc* fpc-adjusted formula is $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t)) = (1 - f_{A+B})v_J(\hat{F}_{\mathrm{ME}}(t))$.

Table 1 shows the relative bias (%Bias) and the coefficient of variation (CV) of the four variance estimators for $\hat{F}_{\mathrm{ME}}(t_\alpha)(\alpha = 0.10,\ 0.25,\ 0.50,\ 0.75,\ 0.90)$, where $F(t_\alpha) = \alpha$. Here, %Bias and CV were calculated as $\%\mathrm{Bias} = 100 \times (M^{-1}\sum_{m=1}^{M} v^{(m)} - \mathrm{MSE})/\mathrm{MSE}$ and $\mathrm{CV} = [M^{-1}\sum_{m=1}^{M}(v^{(m)} - \mathrm{MSE})^2]^{1/2}/\mathrm{MSE}$, respectively, where $v^{(m)}$ is a variance estimate in the m^{th} simulation run. Table 1 demonstrates that $v_J(\hat{F}_{\mathrm{ME}}(t))$ is biased upward since the sampling fractions are not negligible, that $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$ is biased downward since the *ad hoc* adjustment factor $(1 - f_{A+B})$ is too small, and that both $v_a(\hat{F}_{\mathrm{ME}}(t))$ and $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ are approximately unbiased although the latter is slightly more unstable, as is typical for a resampling method.

Table 1 Variance estimation for the pseudo-empirical MLE $\hat{F}_{\mathrm{ME}}(t_\alpha)$

		α				
Estimator		**0.10**	**0.25**	**0.50**	**0.75**	**0.90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	0.27	-0.22	0.64	0.83	2.73
	CV	0.19	0.14	0.14	0.15	0.24
$v_a(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	-2.29	-2.03	-0.47	-1.95	-3.26
	CV	0.17	0.11	0.09	0.11	0.19
$v_J(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	14.24	17.29	22.98	23.80	24.97
	CV	0.24	0.21	0.25	0.27	0.36
$v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	-31.45	-29.63	-26.21	-25.72	-25.02
	CV	0.33	0.30	0.27	0.27	0.30

Paralleling Royall and Cumberland (1981a, 1981b), we ordered the $M = 5{,}000$ simulated samples on the values of $\bar{x}_{A+B} - \bar{x}_A$, classified them into 20 consecutive groups of $G = 250$ in each of which the simulated conditional $\mathrm{MSE}(\mathrm{MSE}_c)$ and conditional mean of $v(E_c(v))$ were computed. Figure 1 shows MSE_c and $E_c(v)$ plotted against the group averages of $\bar{x}_{A+B} - \bar{x}_A$ for $t_{0.10}$ and $t_{0.90}$. It is seen that both $v_a(\hat{F}_{\mathrm{ME}}(t))$ and $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ behave

similarly conditioned on $\bar{x}_{A+B} - \bar{x}_A$. The jackknife variance estimators, $v_J(\hat{F}_{\mathrm{ME}}(t))$ and $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$, though biased, track a trend in MSE_c.

4.3 Quantile estimation

By directly inverting $\hat{F}_{\mathrm{ME}}(t)$, we estimated the α quantile. To obtain $\hat{p}_i$ for $\hat{F}_{\mathrm{ME}}(t)$, we fixed t at $\hat{t}_\alpha$, where $\hat{t}_\alpha = \inf\{t: n_A^{-1}\sum_A I(y_i \le t) \ge \alpha\}$, an estimator using only $\{y_i : i \in A\}$. For variance estimation, $K = 1{,}000$ bootstrap samples were created. For comparison, we also computed the Woodruff variance estimator (Woodruff 1952 and Shao and Tu 1995, page 238),

$$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha)) = \left[\frac{\hat{F}_{\mathrm{ME}}^{-1}(\alpha+\zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}}) - \hat{F}_{\mathrm{ME}}^{-1}(\alpha-\zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}})}{2\zeta_{1-\kappa/2}}\right]^2,$$

where $\hat{\sigma}^2_{\hat{F}} = v(\hat{F}_{\mathrm{ME}}(t))$ with $t = \hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ and $\zeta_{1-\kappa/2}$ is the $(1-\kappa/2)$ quantile of $N(0, 1)$. We let $\kappa = 0.05$ although the best choice of κ is unknown. The performance measures, %Bias and CV, were calculated through $M = 5{,}000$ runs while the true MSE was estimated through 50,000 simulation runs.

Table 2 summarizes the results for quantile estimation. It demonstrates that the mean-adjusted bootstrap has an upward bias in estimating $V(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ while the bias in the Woodruff variance estimator is negligible.

Table 2 Variance estimation for quantiles

		α				
Estimator		**0.10**	**0.25**	**0.50**	**0.75**	**0.90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Bias	6.27	14.32	10.05	10.02	10.28
	CV	0.53	0.53	0.51	0.52	0.61
$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Bias	1.64	3.75	2.92	0.70	-3.67
	CV	0.50	0.45	0.45	0.46	0.52

Figure 2 shows conditional properties of $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ and $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ for $\alpha = 0.10, 0.90$. We see that both $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ and $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ track MSE_c similarly although the former uniformly possesses an upward bias.

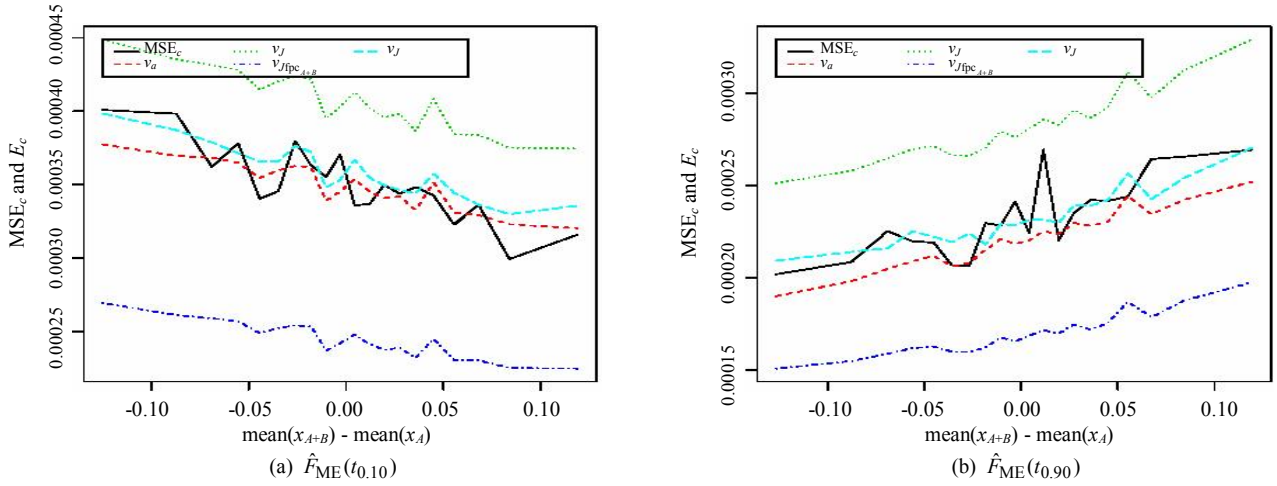

Figure 1 MSE_c and $E_c(v)$ for $\hat{F}_{\mathrm{ME}}(t_\alpha)$

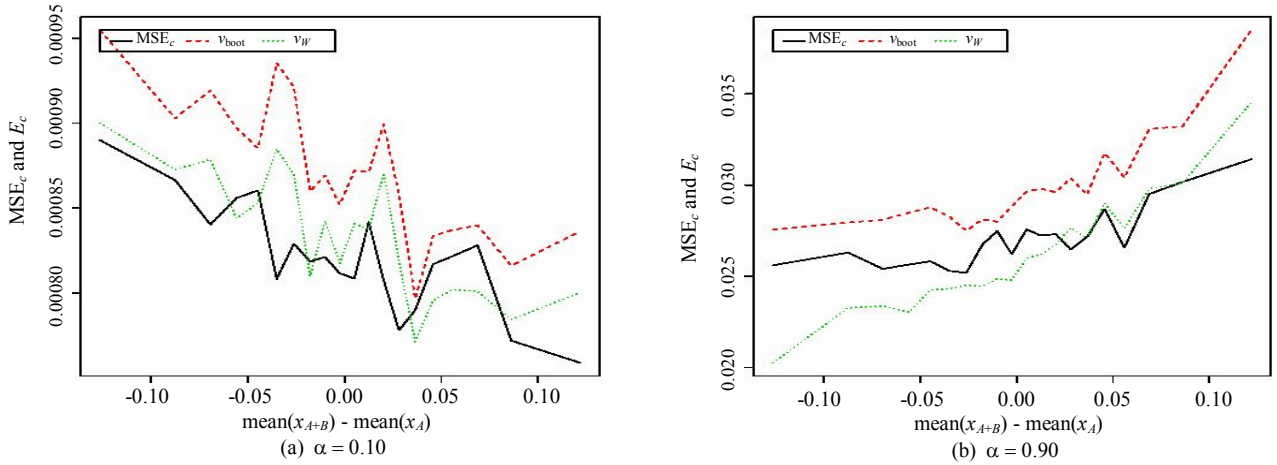

Figure 2 MSE_c and $E_c(v)$ for quantile estimation

5. Further remarks

5.1 Stratified two-phase sampling

Suppose a population is to be stratified into H strata but no information for stratification is available. A possible solution for this situation is to first obtain an SRSWOR of size n' from the population, observe auxiliary variables including the ones for stratification, stratify the sample into H strata, and in each stratum take an SRSWOR of size n_h from n'_h units belonging to stratum h in the sample. See, for example, Cochran (1977, section 12.2) for details.

Let N_h be the size of stratum h in the population. Conditioned on $n'_h > 0$, the first-phase sampling in stratum h described above is equivalent to simple random sampling without replacement of size n'_h in stratum h independent across strata. Thus, given n'_h $(h = 1, ..., H)$, the mean-adjusted bootstrap can be applied independently in different strata to obtain a bootstrap sample. When N_h is unknown, as is usually the case for stratified two-phase sampling, an unbiased estimator $\hat{N}_h = N(n'_h/n')$ can be used in the mean-adjusted bootstrap. In this case, the sampling fraction n'/N is used commonly throughout all the strata.

Note, however, that the present discussion is legitimate for estimates conditioned on the first phase sample sizes. Variance due to the variable n'_h may be large. For unconditional variance estimation, see Kim *et al.* (2006).

5.2 Non-response

The above comment applies to imputed survey data under the uniform response mechanism. Let us suppose that a population is stratified into S_h $(h = 1, ..., H)$ where simple random sampling without replacement is undertaken independently. A sample is divided into imputation classes C_l $(l = 1, ..., L)$ in each of which the response rate is assume to be uniform and imputation is performed. An imputation class may cut across strata. We also assume which imputation class a sampled unit belongs to is correctly identified before imputation. Let us denote the numbers of sampled units and respondents in $S_h \cap C_l$ by n_{hl} and r_{hl}, respectively. Then, it is seen that given n_{hl} and r_{hl}, the corresponding design in $S_h \cap C_l$ is the same as the one discussed in this paper if we regard the n_{hl} units and r_{hl} respondents as $A + B$ and A, respectively. Therefore, the mean-adjusted bootstrap can be conducted independently in different $S_h \cap C_l$ $(h = 1, ..., H; l = 1, ..., L)$. The size of $S_h \cap C_l$, denoted by N_{hl}, can be estimated by $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Note that this is a bootstrap method conditioned on the number of respondents.

6. Conclusion

In this paper, we have proposed the mean-adjusted bootstrap for two-phase sampling. The method requires a simple mean adjustment and can handle the estimation of distribution functions and quantiles because it requires no rescaling. The Taylor series expansion shows that the method has desirable conditional properties for the ratio and regression estimators. A simulation study demonstrates that it also has similar conditional properties in estimating distribution functions and quantiles. An extension to stratified two-phase sampling is straightforward. Conditioned on the first phase sample sizes, the method can handle stratified two-phase sampling and imputation under the uniform response mechanism. We are currently invesigating an extension of the proposed method to more generalized multi-phase sampling designs.

Acknowledgements

This research was supported by a grant from the Japan Society of the Promotion of Science. The author would like to thank Professor Randy R. Sitter, the Editor, the Associate Editor and the two referees for their helpful comments and suggestions.

Appendix A

In this appendix, we show that the proposed bootstrap method provides consistent variance estimates for a class of estimators considered by Rao and Sitter (1997). We use the same setting as in Rao and Sitter (1997) with slightly different notation. For simplicity, we assume there exists only one stratum, but an extension to stratified two-phase sampling is straightforward.

Consider a class of estimators, $\theta = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$, of a population parameter $\theta = h(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, where $\bar{\boldsymbol{Y}}$ and $\bar{\boldsymbol{X}}$ are the population means of vectors $\boldsymbol{y}$ and $\boldsymbol{x}$, *i.e.*, $\bar{\boldsymbol{Y}} = N^{-1}\sum_{i\in P}\boldsymbol{y}_i$ and $\bar{\boldsymbol{X}} = N^{-1}\sum_{i\in P}\boldsymbol{x}_i$. Here, $\boldsymbol{x}$ is observed in the first phase sample $A + B$ whereas $\boldsymbol{y}$ is measured only in the second phase sample A. The sample means $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A)$ and $\bar{\boldsymbol{x}}_B$ are calculated in A and B, respectively, *i.e.*, $\bar{\boldsymbol{y}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{y}_i$, $\bar{\boldsymbol{x}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{x}_i$, and $\bar{\boldsymbol{x}}_B = n_B^{-1}\sum_{i\in B}\boldsymbol{x}_i$.

By a Taylor expansion, we have

$$\hat{\theta} = \theta + \nabla h'(\Delta\bar{\boldsymbol{y}}'_A, \Delta\bar{\boldsymbol{x}}'_A, \Delta\bar{\boldsymbol{x}}'_B)' + o_p(n_A^{-1/2}),$$

where ∇h is the gradient vector of h evaluated at $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, $\Delta\bar{\boldsymbol{y}}_A = \bar{\boldsymbol{y}}_A - \bar{\boldsymbol{Y}}$, $\Delta\bar{\boldsymbol{x}}_A = \bar{\boldsymbol{x}}_A - \bar{\boldsymbol{X}}$, $\Delta\bar{\boldsymbol{x}}_B = \bar{\boldsymbol{x}}_B - \bar{\boldsymbol{X}}$, and $'$ means a transposed matrix (see equation 33.7 of Rao and Sitter 1997, page 757 and the required conditions therein). Then, the variance of $\hat{\theta} = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$ is approximated by

$$V(\hat{\theta}) \doteq \nabla h' \sum_{(\bar{\boldsymbol{y}}'_A, \bar{\boldsymbol{x}}'_A, \bar{\boldsymbol{x}}'_B)'} \nabla h,$$

where $\Sigma_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'}$ is the variance-covariance matrix of $(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'$ under repeated two-phase sampling. Because A and B are SRSWOR's of size n_A and n_B from the population P, respectively, we see that $\Sigma_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A')'} = (1-f_A) S^2_{(\boldsymbol{y}', \boldsymbol{x}')'}/n_A$ and $\Sigma_{\bar{\boldsymbol{x}}_B} = (1-f_B) S^2_{\boldsymbol{x}}/n_B$, where $S^2_{\boldsymbol{u}} = (N-1)^{-1}\sum_{i\in P}(\boldsymbol{u}_i - \bar{\boldsymbol{U}})(\boldsymbol{u}_i - \bar{\boldsymbol{U}})'$ is the population variance of $\boldsymbol{u} = (\boldsymbol{y}', \boldsymbol{x}')'$ or $\boldsymbol{x}$ and $f_B = n_B/N$. For $\mathrm{Cov}(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_B)$, let E_A and $E_{B|A}$ be the expectation for selecting an SRSWOR A from P and choosing an SRSWOR B from $P-A$ given A, respectively. Note that $E_{B|A}(\boldsymbol{x}_B) = (\bar{\boldsymbol{X}} - f_A \bar{\boldsymbol{x}}_A)/(1-f_A)$. So, we have

$$\begin{aligned}\mathrm{Cov}(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_B) &= E(\bar{\boldsymbol{y}}_A \bar{\boldsymbol{x}}_B') - E(\bar{\boldsymbol{y}}_A) E(\bar{\boldsymbol{x}}_B') \\ &= E_A(\bar{\boldsymbol{y}}_A E_{B|A}(\bar{\boldsymbol{x}}_B)) - \bar{\boldsymbol{Y}}\bar{\boldsymbol{X}}' \\ &= -S_{\boldsymbol{yx}}/N,\end{aligned}$$

where $S_{\boldsymbol{yx}} = (N-1)^{-1}\sum_{i\in P}(\boldsymbol{y} - \bar{\boldsymbol{Y}})(\boldsymbol{x} - \bar{\boldsymbol{X}})'$. Similarly, $\mathrm{Cov}(\bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B) = -S^2_{\boldsymbol{x}}/N$.

Now consider a Taylor expansion of $\hat{\theta}^* = h(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ with $\bar{\tilde{\boldsymbol{x}}}_{B^*} = \bar{\boldsymbol{x}}_{B^*} + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)$, the bootstrap analogue of $\hat{\theta} = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$. Let E_* and V_* be the expectation and variance under the proposed bootstrap procedure, respectively. First, observe that $E_*(\bar{\boldsymbol{y}}_{A^*}) = \bar{\boldsymbol{y}}_A$, $E_*(\bar{\boldsymbol{x}}_{A^*}) = \bar{\boldsymbol{x}}_A$ and

$$\begin{aligned}E_*(\bar{\tilde{\boldsymbol{x}}}_{B^*}) &= E_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) \\ &= E_{*A^*}(\bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)) \\ &= \bar{\boldsymbol{x}}_B,\end{aligned}$$

where E_{*A^*} and $E_{*B^*|A^*}$ are respectively the expectation with respect to sampling A^* and the conditional expectation with respect to sampling B^* given A^* under the proposed bootstrap method. Then, $\hat{\theta}^* = h(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ is approximated by

$$\hat{\theta}^* = \hat{\theta} + \nabla h^{*'}(\Delta\bar{\boldsymbol{y}}_{A^*}', \Delta\bar{\boldsymbol{x}}_{A^*}', \Delta\bar{\tilde{\boldsymbol{x}}}_{B^*}')' + o_p(n_A^{-1/2}),$$

where ∇h^* is the gradient of h evaluated at $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$, $\Delta\bar{\boldsymbol{y}}_{A^*} = \bar{\boldsymbol{y}}_{A^*} - \bar{\boldsymbol{y}}_A$, $\Delta\bar{\boldsymbol{x}}_{A^*} = \bar{\boldsymbol{x}}_{A^*} - \bar{\boldsymbol{x}}_A$, and $\Delta\bar{\tilde{\boldsymbol{x}}}_{B^*} = \bar{\tilde{\boldsymbol{x}}}_{B^*} - \bar{\boldsymbol{x}}_B$ (see equation 33.A.1 of Rao and Sitter 1997, page 767 and the required conditions therein). Therefore, $V_*(\hat{\theta}^*)$ is approximated by

$$V_*(\hat{\theta}^*) \doteq \nabla h^{*'} \Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}', \bar{\tilde{\boldsymbol{x}}}_{B^*}')'} \nabla h^*,$$

where $\Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}', \bar{\tilde{\boldsymbol{x}}}_{B^*}')'}$ is the variance-covariance matrix of $(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})'$ under the proposed bootstrap sampling.

Consistent variance estimation under the proposed method is proved by showing ∇h^* and $\Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}', \bar{\tilde{\boldsymbol{x}}}_{B^*}')'}$ are consistent for ∇h and $\Sigma_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'}$, respectively. Consistency of ∇h^* for ∇h follows from consistency of $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)'$ for $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$ and continuity of h.

Consistency of $\Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}', \bar{\tilde{\boldsymbol{x}}}_{B^*}')'}$ can be shown as follows. First, since we use a bootstrap method suitable for simple random sampling without replacement in subsampling A^*, we have $\Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}')'} = (1-f_A)\hat{S}^2_{(\boldsymbol{y}', \boldsymbol{x}')'A}/n_A$, where $\hat{S}^2_{\boldsymbol{u}A} = (n_A-1)^{-1}\sum_{i\in A}(\boldsymbol{u}_i - \bar{\boldsymbol{u}}_A)(\boldsymbol{u}_i - \bar{\boldsymbol{u}}_A)'$ with $\boldsymbol{u} = (\boldsymbol{y}', \boldsymbol{x}')'$. Second, because

1. $\Sigma_{\bar{\tilde{\boldsymbol{x}}}_{B^*}} = E_{*A^*}(V_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) + V_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}))$, where V_{*A^*} and $V_{*B^*|A^*}$ are respectively the variance with respect to sampling A^* and the conditional variance with respect to sampling B^* given A^*,
2. $V_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}) = (1-f_{B|A})\hat{S}^2_{\boldsymbol{x}B}/n_B$, where $\hat{S}^2_{\boldsymbol{x}B} = (n_B-1)^{-1}\sum_{i\in B}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_B)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_B)'$ and $f_{B|A} = n_B/(N-n_A)$, and
3. $E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}) = \bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)$, we have $\Sigma^*_{\bar{\tilde{\boldsymbol{x}}}_{B^*}} = (1-f_{B|A})\hat{S}^2_{\boldsymbol{x}B}/n_B + f_A\hat{S}^2_{\boldsymbol{x}A}/(N-n_A)$. Since both $\hat{S}^2_{\boldsymbol{x}A}$ and $\hat{S}^2_{\boldsymbol{x}B}$ are consistent for $S^2_{\boldsymbol{x}}$, $\Sigma^*_{\bar{\tilde{\boldsymbol{x}}}_{B^*}}$ is consistent for $\Sigma_{\bar{\boldsymbol{x}}_B} = (1-f_B)S^2_{\boldsymbol{x}}/n_B$. Finally, we compute $\mathrm{Cov}_*(\bar{\boldsymbol{y}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ and $\mathrm{Cov}_*(\bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$. For the former, we have

$$\begin{aligned}\mathrm{Cov}_*(\bar{\boldsymbol{y}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*}) &= E_*(\bar{\boldsymbol{y}}_{A^*}\bar{\tilde{\boldsymbol{x}}}_{B^*}') - E_*(\bar{\boldsymbol{y}}_{A^*})E_*(\bar{\tilde{\boldsymbol{x}}}_{B^*})' \\ &= E_{*A^*}(\bar{\boldsymbol{y}}_{A^*}E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) - \bar{\boldsymbol{y}}_A\bar{\boldsymbol{x}}_B' \\ &= E_{*A^*}(\bar{\boldsymbol{y}}_{A^*}\{\bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)\}) - \bar{\boldsymbol{y}}_A\bar{\boldsymbol{x}}_B' \\ &= -\hat{S}_{\boldsymbol{yx}A}/N,\end{aligned}$$

where $\hat{S}_{\boldsymbol{yx}A} = (n_A-1)^{-1}\sum_{i\in A}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_A)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A)'$. Similarly, $\mathrm{Cov}_*(\bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*}) = -\hat{S}^2_{\boldsymbol{x}A}/N$. This completes the proof of consistency of $\Sigma^*_{(\bar{\boldsymbol{y}}_{A^*}', \bar{\boldsymbol{x}}_{A^*}', \bar{\tilde{\boldsymbol{x}}}_{B^*}')'}$ for $\Sigma_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'}$.

Appendix B

In this appendix, we derive $v_{\mathrm{BL}}(\bar{y}_{lr})$. Under the mean-adjusted bootstrap,

$$\begin{aligned}\bar{y}^*_{lr} &= \bar{y}_{A^*} \\ &+ (1-w_A)b_{A^*}\left\{-\frac{(\bar{x}_{A^*} - \bar{x}_A)}{(1-f_A)} + (\bar{x}_{B^*} - \bar{x}_B) + (\bar{x}_B - \bar{x}_A)\right\}.\end{aligned}$$

Define

$$\begin{aligned}\hat{\xi}^*_{pq} &= n_A^{-1}\sum_{i\in A^*} x_i^p\, y_i^q, \\ \hat{\xi}^* &= [\hat{\xi}^*_{10}, \hat{\xi}^*_{01}, \hat{\xi}^*_{11}, \hat{\xi}^*_{20}, \bar{x}_{B^*}]'\end{aligned}$$

and

$$\xi = \left[\bar{x}_A, \bar{y}_A, n_A^{-1}\sum_{i\in A} x_i y_i, n_A^{-1}\sum_{i\in A} x_i^2, \bar{x}_B\right]' = E_*(\hat{\xi}^*).$$

Note that $b_{A^*} = (\hat{\xi}^*_{11} - \hat{\xi}^*_{10}\hat{\xi}^*_{01})/(\hat{\xi}^*_{20} - \hat{\xi}^{*2}_{10})$. Let $\bar{y}^*_{lr} = h(\hat{\xi}^*)$. This expression is slightly different from that in Appendix A, but we may exploit independent subsampling of A^* and B^*. Then, by Taylor linearization of $\bar{y}^*_{lr} = h(\hat{\xi}^*)$ around ξ,

we obtain $\bar{y}^*_{lr} \doteq \bar{y}_{lr} + \nabla h^{*'}(\hat{\xi}^* - \xi)$ and $V_*(\bar{y}^*_{lr}) \doteq \nabla h^{*'} \Sigma^*_{\hat{\xi}^*} \nabla h^*$, where

$$\nabla h^* = [-b_A(1-w_A)/(1-f_A) - z_A(\bar{y}_A - 2b_A\bar{x}_A), 1 - z_A\bar{x}_A, z_A, -z_A b_A, b_A(1-w_A)]'$$

and $\Sigma^*_{\hat{\xi}^*} = [v_{ij}]$ with

$$v_{11} = c_A \hat{S}^2_{xA},$$

$$v_{21} = c_A \hat{S}_{xyA},$$

$$v_{22} = c_A \hat{S}^2_{yA},$$

$$v_{31} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(x_i - \bar{x}_A),$$

$$v_{32} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(y_i - \bar{y}_A),$$

$$v_{33} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})^2,$$

$$v_{41} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i - \bar{x}_A),$$

$$v_{42} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(y_i - \bar{y}_A),$$

$$v_{43} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i y_i - \xi_{11}),$$

$$v_{44} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})^2,$$

$$v_{51} = v_{52} = v_{53} = v_{54} = 0,$$

$$v_{55} = \{n_B^{-1} - (N - n_A)^{-1}\} \hat{S}^2_{xB},$$

$v_{ij} = v_{ji}$, and $c_A = (1 - f_A)/n_A$. Rewriting the moments from the origin as the central moments, noting that $y_i - \bar{y}_A = b_A(x_i - \bar{x}_A) + e_i$ and using properties of e_i as the least-squares residuals, we obtain the right hand side of (5) after some algebra.

References

Berger, Y.G., and Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 68, 531-547.

Biemer, P.P., and Atkinson, D. (1993). Estimation of measurement bias using a model prediction approach. *Survey Methodology*, 19, 127-136.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

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

Chen, J., and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

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

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

Funaoka, F., Saigo, H., Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.

Kim, J.-K., Navarro, A. and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Lee, H., and Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.

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

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K., and Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In *Survey Measurement and Process Quality*: *Wiley Series in Probability and Statistics*. (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York. 753-768.

Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Schreuder, H.T., Li, H.G. and Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.

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

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

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

Wu, C., and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 69-79
Statistics Canada, Catalogue No. 12-001

On standard errors of model-based small-area estimators

Nicholas Tibor Longford [1]

Abstract

We derive an estimator of the mean squared error (MSE) of the empirical Bayes and composite estimator of the local-area mean in the standard small-area setting. The MSE estimator is a composition of the established estimator based on the conditional expectation of the random deviation associated with the area and a naïve estimator of the design-based MSE. Its performance is assessed by simulations. Variants of this MSE estimator are explored and some extensions outlined.

Key Words: Composite estimation; Empirical Bayes estimation; Shrinkage; Small-area estimation.

1. Introduction

Design-based methods have over the years been proven to be inefficient for small-area estimation because, unlike empirical Bayes and related methods, they cannot make effective use of auxiliary information. However, the assumptions associated with the models that are applied remain a weakness of model-based methods because inferences based on them have the ubiquitous caveat of 'If the model is valid …'. In the application of empirical Bayes models to small-area estimation, the local areas (districts) are associated with random effects. In the design-based perspective, this assumption is not valid because in a hypothetical replication of the survey the same districts would be realised (except for some districts that happen not to be represented in the sample drawn), and the target quantities associated with them would also be the same. That is, the districts should be associated with fixed effects. The lack of validity in this aspect of empirical Bayes models has no adverse impact on estimation of small-area quantities (means, totals, proportions, and the like). Associating small areas with random effects is key to borrowing strength from or exploiting the similarity of the areas, as well as to doing so across variables, time points, surveys and other data sources, but it distorts the assessment of the precision of the estimators. Some composite estimators and estimators of their mean squared errors have the same deficiency.

In the next section we diagnose this problem in detail, and in Section 3 propose a solution, which is then illustrated and assessed in Section 4 by simulations using a set of examples. These range from the simplest and most congenial (agreeing with most of the assumptions made) to more complex and realistic but least congenial, so as to explore the robustness of the method. Its fuller potential is discussed in the concluding section.

2. Fixed and random

By sampling variance of a general estimator $\hat{\theta}$ based on a given data-generating (sampling) process χ we understand the variation of the values of $\hat{\theta}(\mathbf{X})$ in replications of the processes that generate datasets $\mathbf{X}$ and apply $\hat{\theta}$ to them. In the design-based perspective, the replication of a survey of a country with its division to D districts yields the same district-level population quantities $\theta_d, d = 1, ..., D$; these D quantities are *fixed*. In contrast, each replication in the model-based perspective, using empirical Bayes models, starts by generating a fresh set of D values θ_d, independently of the previous replications.

We regard the design-based perspective as appropriate, because, in principle, each quantity θ_d could be established with precision and a hypothetical replication of the survey would draw a sample from the same population, with the same division of the country into its districts and the same values of the recorded variables for each member of the population. Most established design-based methods are valid when the survey is based on a perfect sampling frame, which contains no duplicates and is exclusive for the studied population, and the sampling design is implemented with perfection, without any departures from the protocol. That is, the estimators they yield are (approximately) unbiased, the expressions for their sampling variances are correct, or nearly so, and these variances are estimated with small or no bias.

In contrast, model-based methods carry a much heavier burden of assumptions that often cannot be verified. Various model diagnostic procedures are available, but they are all subject to uncertainty. Interpreting failure to find a contradiction as evidence of absence of any contradiction is a commonly committed logical inconsistency. It can be overcome only by quoting properties of estimators when the assumptions are not valid, but such methods are difficult to develop because of a wide range of model violations that

1. Nicholas Tibor Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: NTL@SNTL.co.uk.

one would have to take into account. Yet, despite these drawbacks, model-based methods have proven their worth in small-area estimation and are nowadays rightly regarded as indispensable (Ghosh and Rao 1994; Rao 2003; and Longford 2005).

The EURAREA project (EURAREA Consortium 2004) carried out a large-scale simulation study involving sampling from artificially generated populations that resemble the human populations of several European countries and application of several classes of estimators. It confirmed the superiority of model-based estimators, with several qualifications, but reported rather disappointing results regarding estimators of their standard errors. We trace this problem to an averaging applied in deriving the standard errors of shrinkage estimators.

Suppose a population is divided into D districts, each of them of population size that can for all practical purposes be regarded as infinite, and independent simple random sampling schemes are applied in the districts. We assume that within each district d the outcome variable Y has the normal distribution with mean μ_d and the same variance σ^2_{W}, $N(\mu_d, \sigma^2_{\mathrm{W}})$. For the within-district population means μ_d, we assume the superpopulation model $\mu_d \sim N(\mu^*, \sigma^2_{\mathrm{B}^*})$, but we want to make inferences about a fixed set of (realised) means $\{\mu_d\}$. In Section 5, we discuss the more general regression setting defined by the within-district models

$$(Y|d) \sim N(\mathbf{X}_d\boldsymbol{\beta} + \delta_d, \sigma^2_{\mathrm{W}}),$$

in which $\mathbf{X}_d$ are the within-district regression matrices, $\boldsymbol{\beta}$ the set of corresponding regression parameters common to the districts, and δ_d is the deviation of the within-district regression from the typical regression defined by $\delta_d = 0$. In the superpopulation, δ_d are a random sample from $N(0, \sigma^2_{\mathrm{B}^*})$, but we want to make inferences about the fixed (realised) set $\{\delta_d\}$. Thus, we use model-based estimators, but assess their properties by design-based criteria.

Denote by μ the (national) mean of the quantities μ_d and by σ^2_{B} the district-level variance, $\sigma^2_{\mathrm{B}} = D^{-1}\sum_d (\mu_d - \mu)^2$. Note that they differ from their respective superpopulation counterparts μ^* and $\sigma^2_{\mathrm{B}^*}$. We assume first that $\sigma^2_{\mathrm{B}}, \sigma^2_{\mathrm{W}}$ and μ are known. Let $\hat{\mu}_d$ and $\hat{\mu}$ be the sample means of the variable of interest in district d and in the whole domain (country). They are based on samples of respective sizes n_d and $n = n_1 + \cdots + n_D$. When no covariates are used the empirical Bayes (shrinkage) estimator of μ_d is

$$\tilde{\mu}_d = \left(1 - \frac{1}{1 + n_d\omega}\right)\hat{\mu}_d + \frac{1}{1 + n_d\omega}\hat{\mu}, \qquad (1)$$

where $\omega = \sigma^2_{\mathrm{B}}/\sigma^2_{\mathrm{W}}$ is the variance ratio. The model-based conditional variance of μ_d, given the data, $\mu, \sigma^2_{\mathrm{W}}$ and σ^2_{B}, equal to $\sigma^2_{\mathrm{B}}/(1 + n_d\omega)$, is often regarded as the sampling variance of $\tilde{\mu}_d$; the origins of this practice can be traced to the application of the EM algorithm. A more careful derivation acknowledges that in the design-based perspective $\tilde{\mu}_d$ is biased for μ_d,

$$\mathrm{E}(\tilde{\mu}_d|\mu_d) - \mu_d = -\frac{\mu_d - \mu}{1 + n_d\omega},$$

and its mean squared error is

$$\begin{aligned}\mathrm{MSE}(\tilde{\mu}_d; \mu_d) &= \left(1 - \frac{1}{1 + n_d\omega}\right)^2 \mathrm{var}(\hat{\mu}_d) + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2} \\ &= \sigma^2_{\mathrm{W}}\frac{n_d\omega^2}{(1 + n_d\omega)^2} + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2}, \qquad (2)\end{aligned}$$

assuming, for simplicity, that $\hat{\mu} \equiv \mu$. To emphasise that MSE depends on the target, we include both the estimator and the target in its argument. In particular, $\mathrm{MSE}(\hat{\mu}; \mu) \neq \mathrm{MSE}(\hat{\mu}; \mu_d)$, unless $\mu_d = \mu$. An inconvenient feature of the identity in (2) is that it involves μ_d, the target of estimation. If we replace $(\mu_d - \mu)^2$ with its expectation over the districts, σ^2_{B}, we obtain the more familiar identity

$$\overline{\mathrm{MSE}}(\hat{\mu}_d; \mu_d) = \frac{\sigma^2_{\mathrm{B}}}{1 + n_d\omega}, \qquad (3)$$

the EM-related conditional model-based variance of μ_d. The bar over MSE indicates expectation (averaging) of $(\mu_d - \mu)^2$, the numerator in the last term of (2), over the districts, with the sample sizes n_d intact. Throughout, we condition on the within-district sample sizes $n_d, d = 1, ..., D$, even though in the sampling design each of them may be variable. $\overline{\mathrm{MSE}}$ can be interpreted as model expectation, although the expectation or average of the squared deviations $(\mu_d - \mu)^2$ could be considered and estimated for a given set of districts without any reference to a model. The conditional variance in (3) is appropriate for districts with μ_d in the 'typical' distance, σ_{B}, from the national mean μ. When $|\mu_d - \mu| \neq \sigma_{\mathrm{B}}$, an unbiased estimator of the conditional variance $\sigma^2_{\mathrm{B}}/(1 + n_d\omega)$ is biased for $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$. As the bias is related to the population quantity $\mu_d - \mu$, it is not reduced by increasing the sample size n_d.

3. Composite estimation of MSE

To estimate $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$, we reuse the idea of shrinkage and combine the alternative estimators, $\sigma^2_{\mathrm{B}}/(1 + n_d\omega)$ and a naïve estimator of the MSE in (2). This composite estimator can be motivated as follows. If $n_d = 0$, and therefore $\tilde{\mu}_d = \hat{\mu}$, we have no direct information about μ_d, so we cannot improve on $\sigma^2_{\mathrm{B}}/(1 + n_d\omega)$ as an estimator of $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$. When n_d is large, μ_d is estimated with precision sufficient for using $(\tilde{\mu}_d - \hat{\mu})^2$, possibly with an adjustment for bias, as an estimator of $(\mu_d - \mu)^2$. For

intermediate sample sizes, we search for a composition (compromise) of these two alternatives that are suitable in the extreme settings, when $n_d = 0$ and as $n_d \to +\infty$. We therefore derive expressions for their MSEs and then for the MSE of their combination.

We regard the constant $\sigma_B^2/(1+n_d\omega)$ as an estimator, and refer to it as the *averaged* estimator of MSE. Although it has no variance, it is biased, with mean squared error

$$\begin{aligned}
&\mathrm{MSE}\left\{\frac{\sigma_B^2}{1+n_d\omega}; \mathrm{MSE}(\tilde{\mu}_d;\mu_d)\right\} \\
&\quad = \left\{\frac{\sigma_B^2}{1+n_d\omega} - \frac{\sigma_W^2 n_d\omega^2}{(1+n_d\omega)^2} - \frac{(\mu_d-\mu)^2}{(1+n_d\omega)^2}\right\}^2 \\
&\quad = \left\{\frac{\sigma_B^2-(\mu_d-\mu)^2}{(1+n_d\omega)^2}\right\}^2. \qquad (4)
\end{aligned}$$

The squared deviation $(\mu_d-\mu)^2$, involved in (2), is estimated naïvely by $(\hat{\mu}_d-\hat{\mu})^2$ with bias equal to $\sigma_W^2(n_d^{-1}-n^{-1}) \doteq \sigma_W^2/n_d$ and, assuming that $\hat{\mu}_d$ is normally distributed,

$$\begin{aligned}
&\mathrm{MSE}\{(\hat{\mu}_d-\hat{\mu})^2;(\mu_d-\mu)^2\} \\
&\quad = \mathrm{var}\{(\hat{\mu}_d-\hat{\mu})^2|\mu_d\} \\
&\qquad + [\mathrm{E}\{(\hat{\mu}_d-\hat{\mu})^2-(\mu_d-\mu)^2|\mu_d\}]^2 \\
&\quad \doteq \frac{2\sigma_W^4}{n_d^2} + 4(\mu_d-\mu)^2\frac{\sigma_W^2}{n_d} + \frac{\sigma_W^4}{n_d^2} \\
&\quad = \frac{\sigma_W^2}{n_d}\left\{\frac{3\sigma_W^2}{n_d} + 4(\mu_d-\mu)^2\right\}, \qquad (5)
\end{aligned}$$

derived from the properties of the non-central χ^2 distribution and an approximation by letting $n \to +\infty$. As an alternative, $\tilde{\mu}_d$ may be used instead of $\hat{\mu}_d$; elementary operations yield the approximations

$$\mathrm{E}\{(\tilde{\mu}_d-\hat{\mu})^2|\mu_d\} \doteq (1-b_d)^2\left\{\frac{\sigma_W^2}{n_d} + (\mu_d-\mu)^2\right\}$$

$$\mathrm{var}\{(\tilde{\mu}_d-\hat{\mu})^2|\mu_d\} \doteq \frac{(1-b_d)^4}{n_d^2}\sigma_W^2\{2\sigma_W^2 + 4n_d(\mu_d-\mu)^2\},$$

where $b_d = 1/(1+n_d\omega)$, and so

$$\begin{aligned}
&\mathrm{MSE}\{(\tilde{\mu}_d-\hat{\mu})^2;(\mu_d-\mu)^2\} \\
&\quad = \mathrm{var}\{(\tilde{\mu}_d-\hat{\mu})^2|\mu_d\} + [\mathrm{E}\{(\tilde{\mu}_d-\hat{\mu})^2-(\mu_d-\mu)^2|\mu_d\}]^2 \\
&\quad \doteq (1-b_d)^4\frac{3\sigma_W^4}{n_d^2} \\
&\qquad + 2(1-b_d)^2(2-6b_d+3b_d^2)\frac{\sigma_W^2(\mu_d-\mu)^2}{n_d} \\
&\qquad + b_d^2(2-b_d)^2(\mu_d-\mu)^4. \qquad (6)
\end{aligned}$$

This approximation is valid only for $b_d = 1/(1+n_d\omega)$, so further approximation is involved when we substitute a possibly suboptimal choice or an estimate of b_d based on an estimate of ω. In general, the coefficient b_d that minimises the MSE in (6) differs from $1/(1+n_d\omega)$ because the shrinkage with $b_d = 1/(1+n_d\omega)$ is optimal only for targets that are linear transformations of μ_d (Shen and Louis 1998). We do not pursue this avenue because the solution, being a complicated function of the parameters, is likely to be sensitive to the error in estimation of the parameters. The estimator $(\hat{\mu}_d-\hat{\mu})^2$ could be corrected for its bias in estimating $(\mu_d-\mu)^2$, although this may result in a negative estimate, especially when n_d is small.

Finally, we combine the two (biased) estimators of $\mathrm{MSE}(\tilde{\mu}_d;\mu_d)$, the averaged estimator $\sigma_B^2/(1+n_d\omega)$ and the naïve estimator derived from the identity in (2), using $(\hat{\mu}_d-\hat{\mu})^2$ as an estimator of $(\mu_d-\mu)^2$. The MSEs of these two estimators depend on $(\mu_d-\mu)^2$, so we replace the relevant terms by their expectations across the districts *d*. We replace $(\mu_d-\mu)^2$ with σ_B^2, and $(\mu_d-\mu)^4$ with $3\sigma_B^4$ or, in general, with $\kappa\sigma_B^4$, where κ is the kurtosis of the (district-level) distribution of μ_d. Although it may at first appear that we have not gained anything, because we still have to remove the dependence of MSE on $(\mu_d-\mu)^2$ by using σ_B^2 instead, now we make this step at a later stage. In the simulations in Section 4, we show that this reduces the undesirable impact of averaging.

Thus, we search for the coefficient c_d that minimises the expected MSE of the composite estimator of the MSE,

$$\begin{aligned}
&\widetilde{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) \\
&\quad = (1-c_d)\widehat{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) + c_d\,\overline{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) \\
&\quad = (1-c_d)\left\{(1-b_d)^2\frac{\sigma_W^2}{n_d} + b_d^2(\hat{\mu}_d-\hat{\mu})^2\right\} + c_d b_d\sigma_B^2. \qquad (7)
\end{aligned}$$

To evaluate the MSE of this MSE estimator, as a function of c_d, we use the expressions

$$\begin{aligned}
&\overline{\mathrm{MSE}}\{b_d\sigma_B^2;\mathrm{MSE}(\tilde{\mu}_d;\mu_d)\} = 2b_d^4\sigma_B^4, \\
&\overline{\mathrm{MSE}}\{(\hat{\mu}_d-\hat{\mu})^2;(\mu_d-\mu)^2\} \doteq \frac{\sigma_W^4}{n_d^2}(3+4n_d\omega), \\
&\overline{\mathrm{MSE}}\{(\tilde{\mu}_d-\hat{\mu})^2;(\mu_d-\mu)^2\} \\
&\qquad \doteq \frac{\sigma_W^4}{n_d^2}\{3(1-b_d)^4 + 3b_d^2(2-b_d)^2 n_d^2\omega^2 \\
&\qquad\qquad + 2(1-b_d)^2(2-6b_d+3b_d^2)n_d\omega\},
\end{aligned}$$

derived by averaging of the respective equations (4), (5) and (6); $(\mu_d-\mu)^2$ is replaced by σ_B^2 and $(\mu_d-\mu)^4$ by $3\sigma_B^4$.

Assuming that the district-level targets μ_d are normally distributed, the MSE of the composite estimator in (7) is

$$
\begin{aligned}
&\mathrm{E}\{(1-c_d)(1-b_d)^2\frac{\sigma_{\mathrm{W}}^2}{n_d}+(1-c_d)b_d^2(\hat{\mu}_d-\hat{\mu})^2\\
&\quad+c_d b_d\sigma_{\mathrm{B}}^2-b_d^2\sigma_{\mathrm{W}}^2 n_d\omega^2-b_d^2(\mu_d-\mu_d)^2\}^2\\
&=b_d^4\,\mathrm{E}\{(1-c_d)\sigma_{\mathrm{B}}^2 n_d\omega+(1-c_d)(\hat{\mu}_d-\hat{\mu})^2\\
&\qquad+c_d\sigma_{\mathrm{B}}^2(1+n_d\omega)-\sigma_{\mathrm{B}}^2 n_d\omega-(\mu_d-\mu)^2\}^2\\
&=b_d^4\,\mathrm{E}\{(1-c_d)(\hat{\mu}_d-\hat{\mu})^2+c_d\sigma_{\mathrm{B}}^2-(\mu_d-\mu)^2\}^2\\
&\doteq b_d^4\left[(1-c_d)^2\left\{\frac{2\sigma_{\mathrm{W}}^4}{n_d^2}+\frac{4\sigma_{\mathrm{W}}^2}{n_d}(\mu_d-\mu)^2\right\}\right]\\
&\quad+b_d^4\left[(1-c_d)\frac{\sigma_{\mathrm{W}}^2}{n_d}+c_d\{\sigma_{\mathrm{B}}^2-(\mu_d-\mu)^2\}\right]^2,
\end{aligned}
$$

using the identities $(1-b_d)^2=b_d^2 n_d^2\omega^2$ and $\sigma_{\mathrm{B}}^2=\sigma_{\mathrm{W}}^2\omega$ to extract the factor b_d^4. By taking the expectation over the districts, keeping the sample sizes intact, we obtain

$$
\begin{aligned}
&\overline{\mathrm{MSE}}\{\widetilde{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d)\}\\
&\qquad\doteq\frac{b_d^4}{n_d^2}\{(1-c_d)^2(3+4n_d\omega)\sigma_{\mathrm{W}}^4+2c_d^2 n_d^2\sigma_{\mathrm{B}}^4\}.
\end{aligned}
$$

The minimum of this quadratic function of c_d is attained for

$$
c_d^*=\frac{3+4n_d\omega}{3+4n_d\omega+2n_d^2\omega^2}.
$$

This choice of a coefficient c_d agrees with our expectations. For $n_d=0, c_d^*=1$ and we rely solely on the averaged MSE estimator, equal to σ_{B}^2. Further, c_d^* is a decreasing function of n_d, converging to zero as n_d diverges to $+\infty$; for large n_d we rely on the naïve estimator of MSE. It is also a decreasing function of ω; for $\omega=0$, that is, $\sigma_{\mathrm{B}}^2=0, c_d^*=1$ for every district d, confirming that $\mu_d\equiv\mu$ and μ_d would be estimated precisely *if* μ were known. With increasing ω, $\sigma_{\mathrm{B}}^2/(1+n_d\omega)$ becomes less and less useful because the squared deviations $(\mu_d-\mu)^2$ are widely spread (around σ_{B}^2).

If we adjust $(\hat{\mu}_d-\hat{\mu})^2$ for its bias in estimating $(\mu_d-\mu)^2$, the expected MSE of the shrinkage estimator is minimised for

$$
c_d^\dagger=\frac{1+2n_d\omega}{(1+n_d\omega)^2}.
$$

It is easy to check that

$$
c_d^*-c_d^\dagger=\frac{n_d^2\omega^2}{(1+n_d\omega)^2}\,\frac{1}{3+4n_d\omega+2n_d^2\omega^2},
$$

so the bias-adjusted estimator derived from (2) is assigned greater weight (equal to $1-c_d^\dagger$) than the naïve estimator would be. But the difference is small for all values of $n_d\omega$.

The composite MSE estimator based on $(\tilde{\mu}_d-\hat{\mu})^2$ is derived similarly, but the resulting expression is much more complex. The optimal shrinkage coefficient is

$$
\begin{aligned}
c_d^{*\prime}&=3(1-b_d)^4+2(1-b_d)^2 f(b_d)n_d\omega-b_d(2-b_d)f(b_d)n_d^2\omega^2\\
&\quad\times[3(1-b_d)^4+2(1-b_d)^2 f(b_d)n_d\omega-\\
&\qquad\{2-4b_d(2-b_d)+3b_d^2 f(b_d)\}n_d^2\omega^2],
\end{aligned}
$$

where $f(b_d)=2-6b_d+3b_d^2$. The dependence on b_d is particularly problematic, because in practice b_d is estimated and the properties of the MSE estimator based on estimated $c_d^{*\prime}$ are bound to be affected by the uncertainty about b_d. In the derivations, we used the identity $b_d=1/(1+n_d\omega)$, so this expression could not be used when the values of b_d are set *a priori*.

4. Simulations

Properties of the composite estimator of MSE cannot be derived analytically, and so we resort to simulations. We consider the artificial setting of a national survey with a stratified sampling design, with strata coinciding with the country's 100 districts for which estimates of the means of a variable Y are sought. Simple random sampling is applied within each stratum, assumed to be of practically infinite population size. We have generated the values of the means μ_d from the normal distribution $N(\mu=20,\sigma_{\mathrm{B}}^2=8)$, and the sample sizes n_d from scaled conditional beta distributions, given the means μ_d, so as to inject a modicum of dependence of the means on the sample sizes. With this adjustment, the assumption underlying the averaged MSE estimator is false, but this could not be detected by a diagnostic procedure or a hypothesis test, not even with μ_d known. The sample size of one district was altered to be much greater than the rest, to represent the capital of the fictitious country. The within-stratum distributions of Y are $N(\mu_d,\sigma_{\mathrm{W}}^2=100)$. The district-level means and sample sizes are fixed in the replications. For orientation, they are plotted in Figure 1. The districts are assigned order numbers from 1 to 100 in the ascending order of their sample sizes. The smallest sample size is $n_1=15$ and the overall sample size is 3,698.

In the simulations, comprising 1,000 replications, we generate the direct estimates $\hat{\mu}_d$ as independent random draws from $N(\mu_d,\sigma_{\mathrm{W}}^2/n_d)$ and the within-district corrected sums of squares as independent draws from the appropriately scaled χ^2 distributions with n_d-1 degrees of freedom. Then we evaluate the shrinkage estimator $\tilde{\mu}_d$ for each district d, followed by evaluation of the averaged, naïve and the two composite MSE estimators using the coefficients c_d^* and $c_d^\dagger$ or their naïve estimates.

In the first set of replications, we assume that μ, σ_W^2 and σ_B^2 are known, so that the simulation reproduces the theoretically derived results and enables us to assess the quality of the composite MSE estimators without the interference of uncertainty about the shrinkage coefficient $b_d = 1/(1+n_d\omega)$. The results are summarised graphically in Figure 2. The empirical biases (their absolute values) of the four MSE estimators are plotted in the left-hand panel. Circles and black dots are used for the averaged and naïve estimators, respectively, and the biases of the composite estimators are connected by solid lines. The absolute values of the empirical biases are plotted, to highlight their strong association with the sample size for the naïve estimator. For 60 districts (60%), the composite estimator of MSE has a positive bias. For the naive estimator, this count or percentage is higher (78), and for the averaged estimator lower (52). Throughout, the main contributor to the bias of the averaged MSE estimator is the deviation of the squared distance $(\mu_d - \mu)^2$ from the district-level variance σ_B^2. The two composite estimators, based on $(\hat{\mu}_d - \hat{\mu})^2$ and on its bias-adjusted version, differ so little that their biases cannot be distinguished in the plot. The diagram shows that the averaged estimator of MSE entails substantial bias for a few districts, including several with large sample sizes. The biases of the naïve and composite estimators are without such extremes.

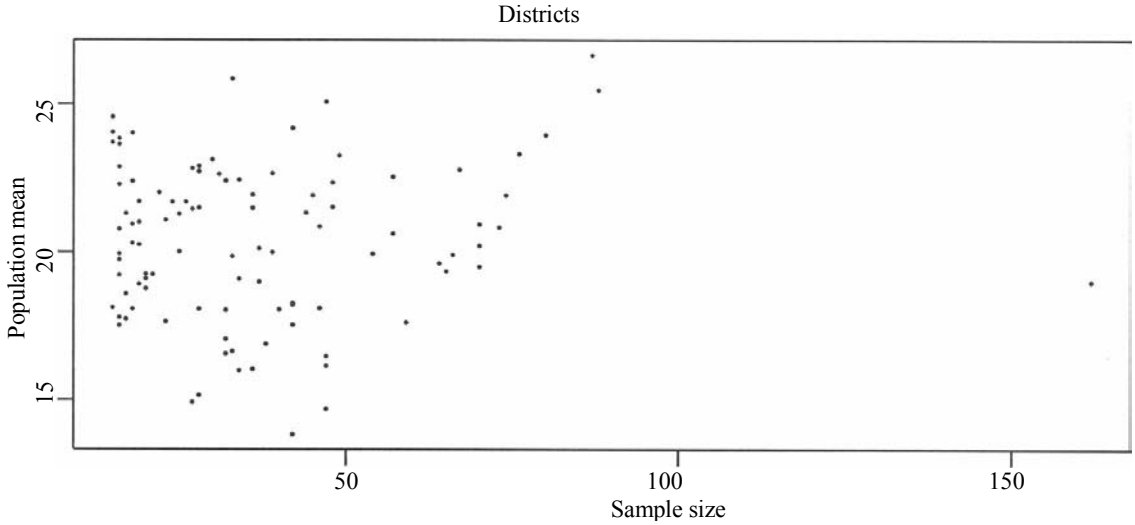

Figure 1 The district-level sample sizes and population means of *Y*. Artificially generated values

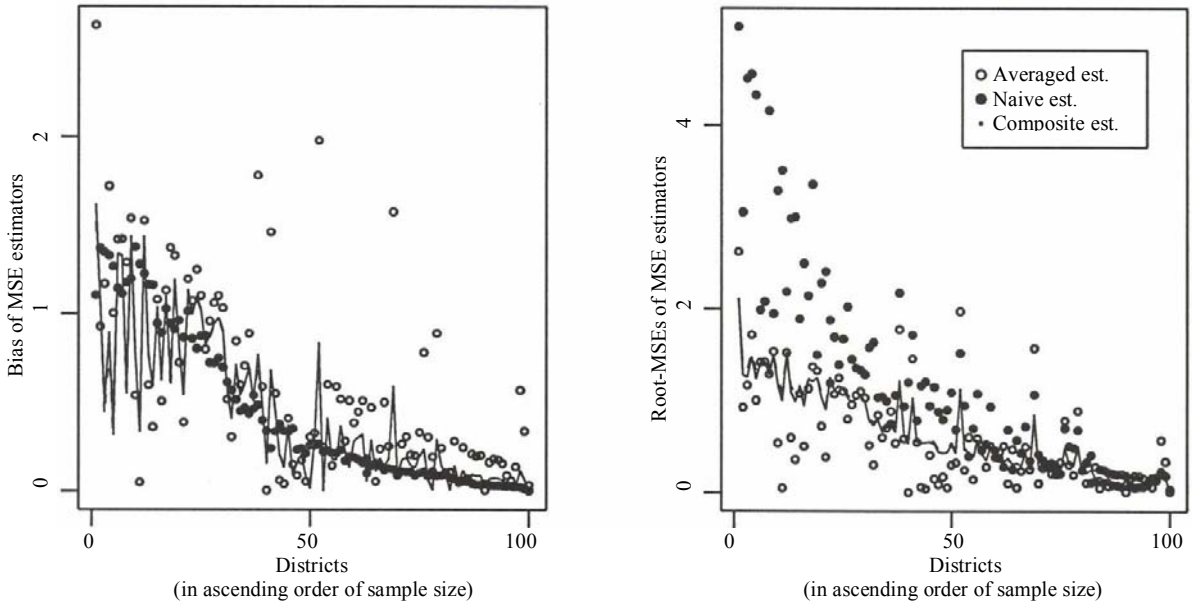

Figure 2 The bias and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. Based on simulations with an artificial setting. The bias and root-MSE of the composite estimators are connected by solid lines

In the right-hand panel, the root-MSEs of the MSE estimators are plotted, using the same symbols and layout. The diagram shows that the naïve estimator is inefficient, especially for districts with the smallest sample sizes, whereas the averaged estimator is very efficient for some but inefficient for some other districts, without any apparent relation to their sample sizes. In fact, apart from sample size, high efficiency is associated with proximity of $(\mu_d - \mu)^2$ to σ_B^2 and low efficiency with the smallest and largest values of $(\mu_d - \mu)^2$. For example, the empirical root-MSE of the averaged MSE estimator for district 1, with $n_1 = 15$, is 2.63, whereas its counterpart for district 11 $(n_{11} = 16)$ is 0.049. Their population means are $\mu_1 = 24.55$, exceeding $\mu + \sigma_B$ by 1.72, and $\mu_{11} = 22.87$, differing from $\mu + \sigma_B$ by only 0.04. The root-MSEs of the naïve estimator are 5.08 and 3.51, and those of the composite estimator are 2.10 and 1.00 for the respective districts 1 and 11. The composite MSE estimator performs much more evenly, moderating the deficiencies of the averaged and naïve estimators.

All three estimators are conservative (have positive biases) for districts with relatively small MSE of $\tilde{\mu}_d$. The averaged estimator has negative biases when the MSEs are relatively large. The composite estimator also has negative biases for such districts, but they tend to be smaller in absolute value. For districts with the smallest sample sizes, the composite estimator is not very effective because the naïve estimator is very inefficient. For a few of these districts, the composition is counterproductive, as a result of averaging, but such districts cannot be identified from a single realisation of the survey.

Next we study a less congenial setting, in which the normality assumptions of μ_d across the districts and of the elementary observations y_{id} within the districts are still satisfied, but the global parameters, μ, σ_W^2 and σ_B^2, are not known and are estimated. We use the same means μ_d and sample sizes n_d as in Figure 1. The results of the simulations are summarised in Figure 3. In the left-hand panel, the empirical means of the MSE estimators are plotted, using the same symbols as in Figure 2, together with the empirical MSEs (crosses '+') of the shrinkage estimators $\tilde{\mu}_d$. The empirical means of the averaged estimators have a regular pattern because the estimates in each replication depend only on the sample size n_d and the estimated variance ratio $\hat{\omega}$. For biases, the naive estimators have a regular pattern, similar to their pattern in Figure 2. The naïve estimators have positive biases that decline with the sample size. The averaged estimators are far too conservative; their means do not veer from the smooth trend. The composite MSE estimators deviate from this trend in the appropriate direction, but not to full degree. Their average bias is positive, equal to 0.22, or 10% (2.42 *vs.* 2.20), and they overestimate the target MSE for 70 out of the 100 districts.

The right-hand panel displays the root-MSEs of the MSE estimators. The naïve estimator is inefficient, whereas the averaged estimator is very efficient for some and rather inefficient for other districts. The composite MSE estimator is more efficient than either naïve or averaged estimator for 36 districts; it is more efficient than the averaged estimator in exactly half of the districts, but it does not have its glaring weaknesses. As in the congenial setting (Figure 2), the differences due to bias adjustment of $(\hat{\mu}_d - \hat{\mu})^2$ in composite MSE estimation (using coefficients c_d^* or $c_d^{*\prime}$) are negligible.

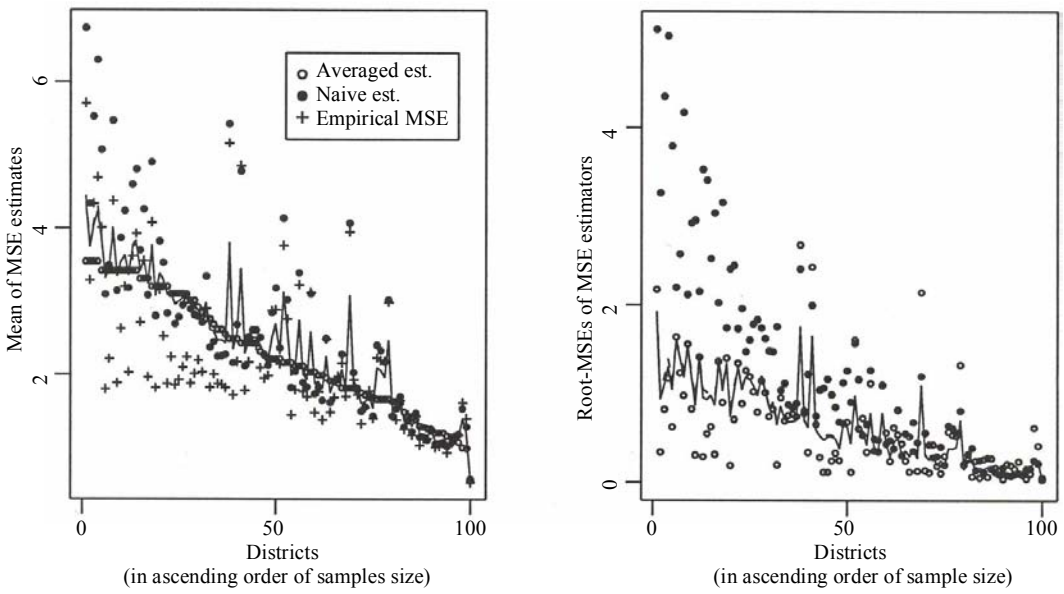

Figure 3 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. The global parameters μ, σ_W^2 and σ_B^2 are estimated

Next we compare the MSE estimators for the district-level means of $Y^2/100$, denoted by ν_d. The assumptions of normality both within and across districts are no longer appropriate. We apply the methods that rely on the normality assumptions, to assess the robustness of the composite estimators, but also to contrast the deficiencies of the averaging with the consequences of using 'incorrect' models. We chose the square transformation because the within-district expectations are known, equal to $(\mu_d^2 + \sigma_W^2)/100$, and could be estimated by

$$\tilde{\nu}_d^* = \frac{\tilde{\mu}_d^2 - \widehat{\mathrm{MSE}}(\tilde{\mu}_d) + \hat{\sigma}_W^2}{100}. \qquad (8)$$

We denote by $\tilde{\nu}_d$ the empirical Bayes estimators applied to $y_{id}^2/100$.

The results of the simulations based on the values of $y_{id}^2/100$ are presented in Figure 4, using the same layout and symbols as in Figure 3. The same conclusions about the biases and root-MSEs are arrived at as before, except that the naïve estimator is even more inefficient and the performance of the averaged estimator even more erratic - it is both very efficient and inefficient for more districts than in the more congenial setting of Figure 3. The naïve estimator is conservative, but for some districts with small n_d far too much so, and its MSEs for these districts are very large.

We contrast these conclusions with a comparison of estimating the district-level means of $Y^2/100$ by $\tilde{\nu}_d^*$, transforming the estimates $\tilde{\mu}_d$ according to (8). The estimator $\tilde{\nu}_d^*$ is more efficient than $\tilde{\nu}_d$ for most districts (90, in fact), and when less efficient, the relative difference of their MSEs is less than 4%. For a few districts, the difference in efficiency is perceptible, exceeding 20% for ten districts. However, the differences in the MSEs are small in comparison with the biases in estimating these MSEs, as shown in Figure 5. The biases and MSEs of $\tilde{\nu}_d$ are marked by black dots connected to their counterparts for $\tilde{\nu}_d^*$.

Part of the lack of efficiency of $\tilde{\nu}_d$ is due to its bias; the bias of $\tilde{\nu}_d$ exceeds the bias of $\tilde{\nu}_d^*$ for all but two districts, but the difference is non-trivial only when both estimators are positively biased. Thus, little efficiency is gained by arranging the analysis so that the distributional assumptions are satisfied. The gains are modest in comparison with the increase in the difficulty of estimating the efficiency, as expressed by $\mathrm{MSE}(\tilde{\nu}_d^*; \nu_d)$. Although the sampling variation of $\hat{\sigma}_W^2$ is trivial in large-scale surveys, the contribution of $\widehat{\mathrm{MSE}}(\tilde{\mu}_d; \mu_d)$ to $\mathrm{MSE}(\tilde{\nu}_d^*; \nu_d)$ cannot be ignored.

Figure 6 compares the composite MSE estimator with the naïve estimator of MSE of $\tilde{\mu}_d$ based on the empirical Bayes estimator of μ_d; it is derived by substituting $\tilde{\mu}_d$ for μ_d in (2). For brevity, we refer to it as the EB-naïve estimator. As anticipated in Section 3, it tends to underestimate its target. It is more efficient than the composite estimator of MSE for about half the districts (52 out of 100), but its performance is more uneven than that of the composite MSE estimator. In principle, the EB-naïve estimator could be improved by combining it with the averaged estimator; however, only minor improvement is made even in the congenial setting (known μ, σ_W^2 and σ_B^2), and the composition is detrimental for several districts in the less congenial settings. Details are omitted.

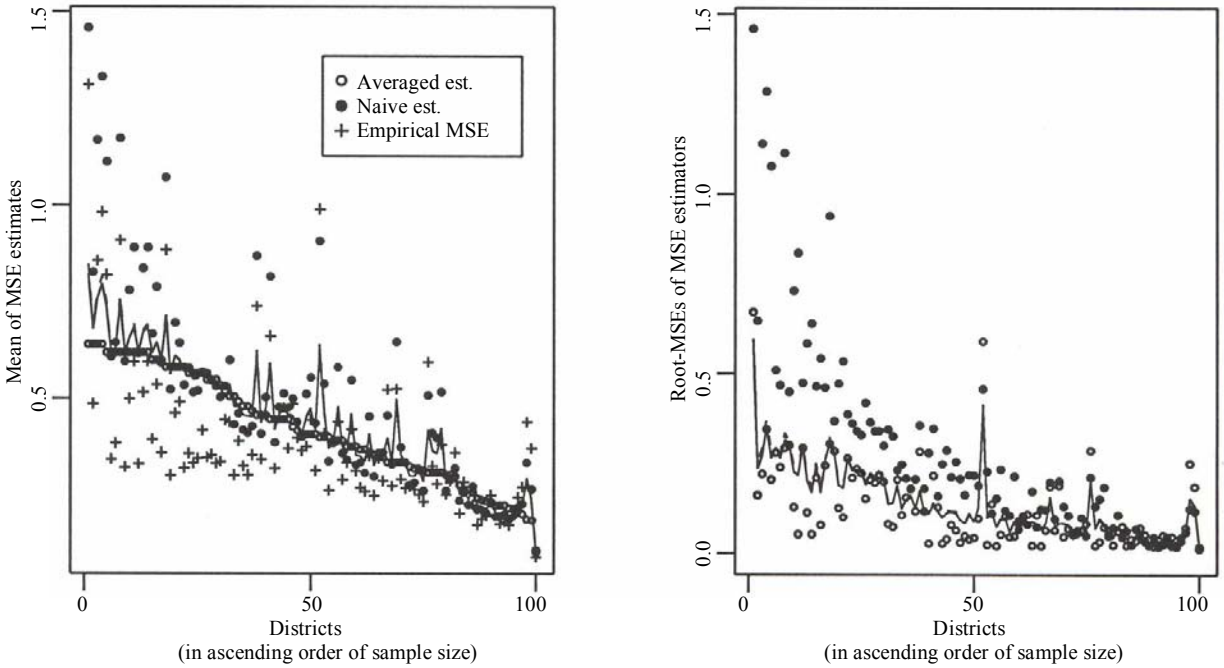

Figure 4 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators; estimation of the means of $Y^2/100$

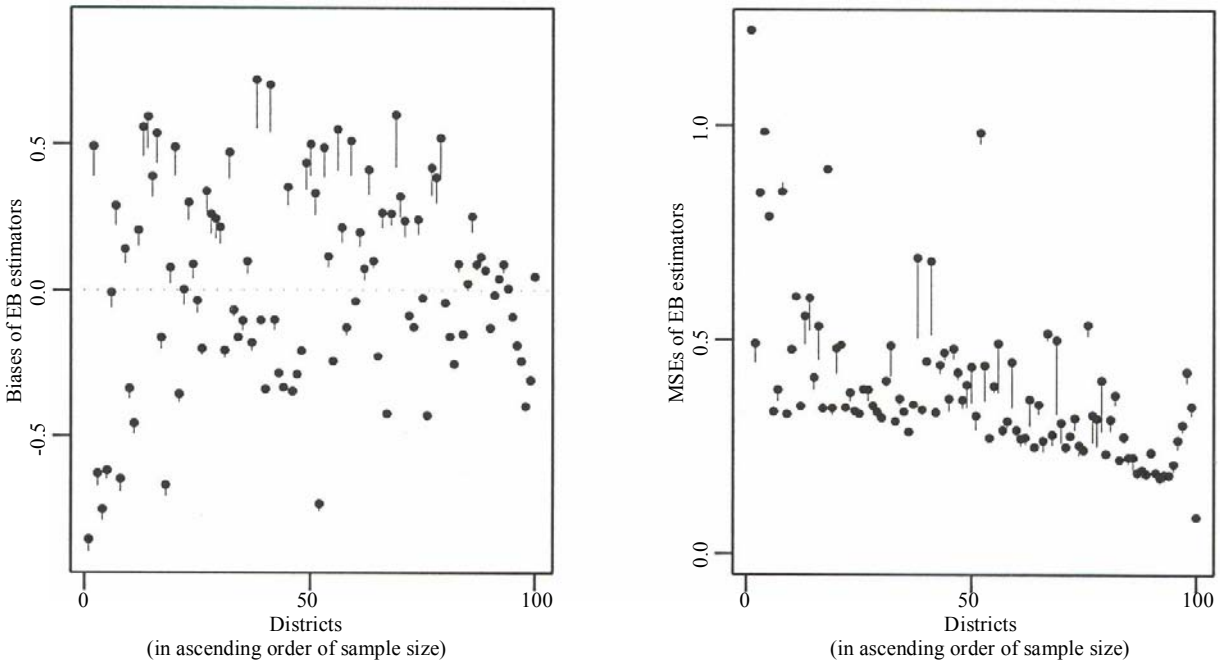

Figure 5 The biases and MSEs of estimators of ν_d. The vertical segments connect the quantities associated with $\tilde{\nu}_d^*$ and $\tilde{\nu}_d$. The quantities associated with $\tilde{\nu}_d$ are marked by black dots

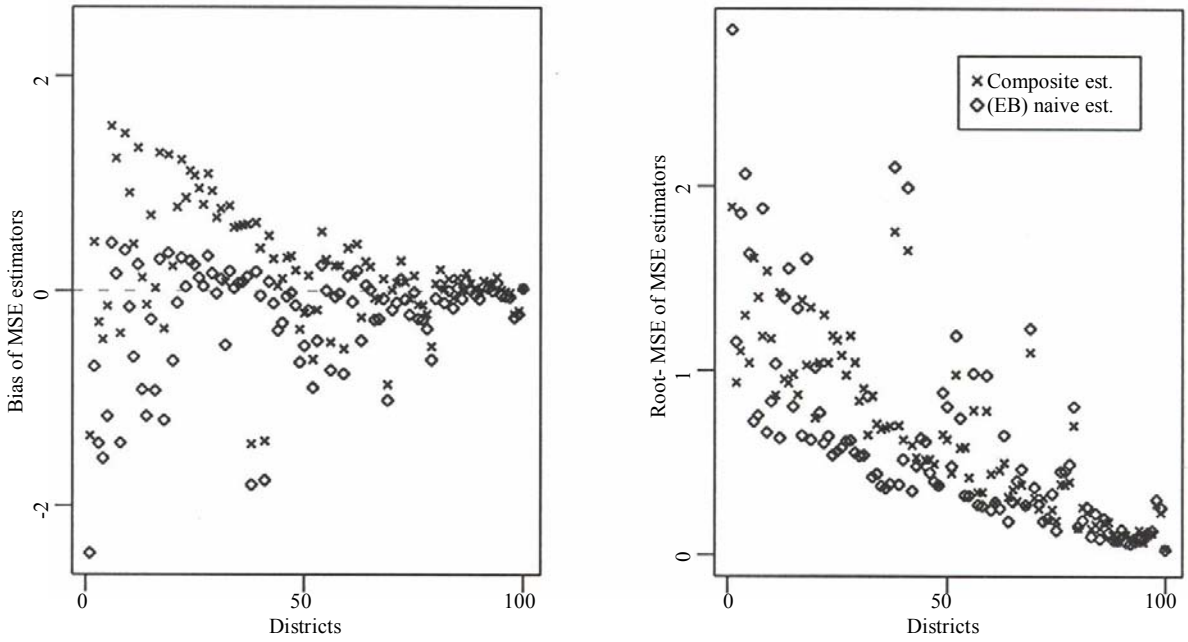

Figure 6 The bias and root-MSE of the composite and empirical-Bayes naïve estimators of the MSE of $\tilde{\mu}_d$

As a final simulation, we consider a binary outcome variable that indicates whether $Y < 5$, so that the district-level percentages are in the range 1.5-18.8 and the dependence of the percentage on the variance within districts is substantial. The mean of the district-level percentages is 6.85; the substantial skew of these percentages (skewness coefficient equal to 1.01 and kurtosis to 3.78) provides a stern test of the method.

In the simulation, the district-level percentages are estimated by the univariate version of the shrinkage method described in Longford (1999 and 2005, Chapter 8). The results are summarised in Figure 7. The MSE is over-estimated by all three estimators for most districts, except for a minority for which the empirical MSE is several times higher than for the rest. The naïve estimator has a substantial bias for most districts. The averaged estimator is less

regimented than for normally distributed outcomes because the shrinkage coefficient depends also on the estimated proportion, which is truncated from below at 2% to avoid zero estimated variance $\hat{p}_d(1-\hat{p}_d)/n_d$. The graph of the composite MSE estimates has the spikes for the appropriate districts, but the spikes are far too short to reduce the bias substantially.

The MSEs of the averaged estimator are satisfactory for most, but are very large for several districts. For the latter districts, the naïve MSE estimator is even less efficient. The composite MSE estimator is less efficient than the averaged estimator for many districts, but the difference is rather small, compensated by the gains in efficiency for districts for which the averaged estimator is less efficient. The EB-naïve MSE estimator resembles in many features the naïve MSE estimator; it is not plotted in the diagram.

In conclusion, this simulation shows that when one of the MSE estimators, in this case the naïve estimator, is very inefficient, it nevertheless contributes, even if very modestly, to the efficiency of the composite MSE estimator. The composite estimator draws on the best that the constituent estimators, averaged and naïve, have to offer, even in uncongenial settings. A remaining challenge is to combine the naïve and averaged estimators to satisfy a particular criterion which trades off the precision for districts that are estimated with high precision for higher precision in estimating in the districts with low precision. For example, we may be less concerned about estimation of the MSEs for districts with abundant representation in the sample and much more about the sparsely represented districts. Also, some districts (*e.g.*, those in a particular region) may be of specific interest, unrelated to their representation. Of course, the first step in this is the definition of one or a class of criteria that reflect the inferential priorities, and this is bound to be specific to each survey and client. See Longford (2006) for some proposals.

4.1 Refinements and extensions

Several elements of realism can be incorporated in the derivation of the composite MSE estimator. First, uncertainty about μ can be reflected by acknowledging that $\hat{\mu}_d$ and $\hat{\mu}$ are correlated. Thus, $\mathrm{var}(\hat{\mu}_d - \hat{\mu}) = \sigma^2_W\,(1/n_d - 1/n)$ and the approximation in (5) becomes equality when both instances of σ^2_W/n_d are replaced by $\sigma^2_W(1/n_d - 1/n)$. This brings about only a slight change when $n_d \ll n$, the case for most districts. If the country has a dominant district, with sample size that is a large fraction of the overall sample size, then this adjustment might be relevant, but it has a negligible impact on MSE estimation because even direct estimation of the mean for the district is nearly efficient.

A similar refinement can be applied to the empirical Bayes estimator of μ_d. It amounts to replacing n_d with $1/(n_d^{-1} - n^{-1}) = n_d n/(n - n_d)$ in the coefficient $b_d = 1/(1 + n_d\omega)$. The change is not trivial only for a dominant district, but for such a district shrinkage yields only minute improvement over direct estimation with or without this adjustment.

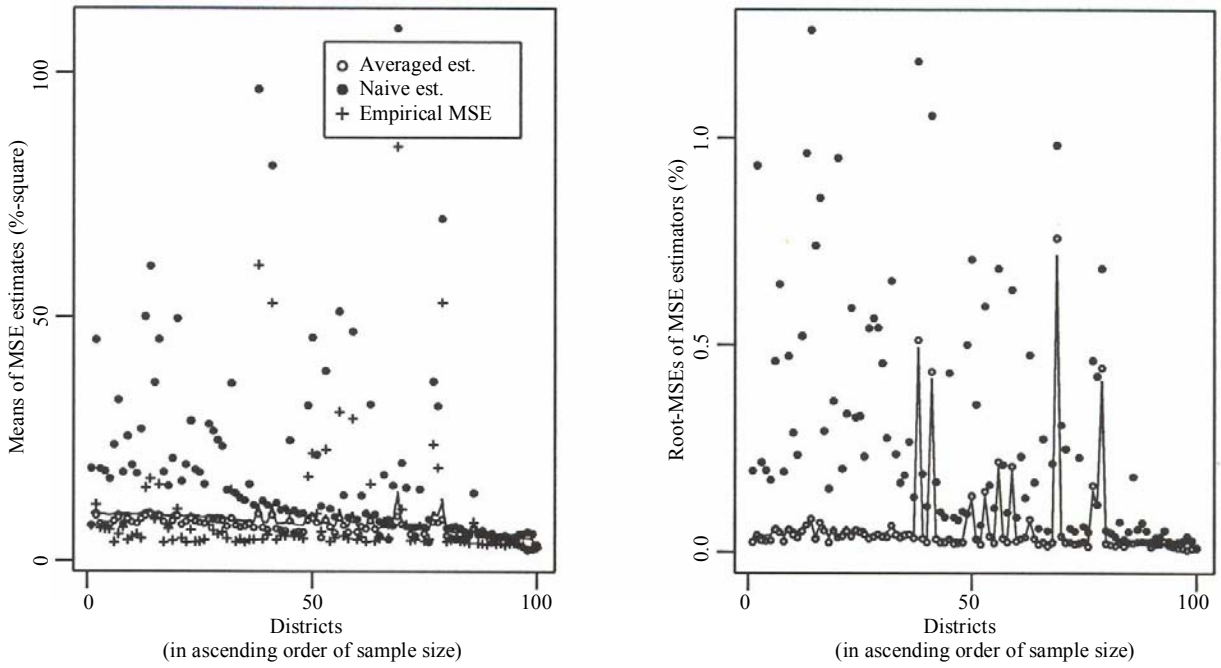

Figure 7 The mean and root-MSE of the composite naïve and averaged estimators of the MSEs of district-level percentages

Accommodating sampling designs that differ from stratified random sampling, and which associate subjects with sampling weights, generates in composite estimation no problems additional to direct estimation with such designs and weights, because we require only the sampling variances of $\hat{\mu}_d, \hat{\mu}$ and functions of these. Similarly, exploiting auxiliary information by applying (empirical Bayes) regression

$$y_{jd} = \mathbf{x}_{jd}\boldsymbol{\beta} + \delta_d + \varepsilon_{jd},$$

with independent random samples $\delta_d \sim N(0, \sigma_B^2)$ and $\varepsilon_{jd} \sim N(0, \sigma_W^2)$, amounts to replacing $\hat{\mu}$ in (1) with the prediction $\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}$, where $\hat{\mathbf{x}}_d$ is the vector of means of the regressors for district d and $\hat{\boldsymbol{\beta}}$ is the vector of regression parameter estimates. To see this, we express the empirical Bayes fit for district d as

$$\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}} + \frac{n_d\omega}{1+n_d\omega}(\hat{\mu}_d - \hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}) = \frac{n_d\omega}{1+n_d\omega}\hat{\mu}_d + \frac{1}{1+n_d\omega}\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}.$$

Pfeffermann *et al.* (1998) discuss issues related to fitting empirical Bayes models to observations with sampling weights. Composite estimation uses direct estimators $\hat{\boldsymbol{\mu}}_d$ and $\hat{\boldsymbol{\mu}}$ for the vectors of all the variables involved and their estimated sampling variance matrices; their evaluation is a standard task in sampling theory. An outstanding problem with empirical Bayes estimators arises when $\hat{\mathbf{x}}_d$ is based on very few observations because the uncertainty about μ_d is then inflated, even when the model fit is very good; if the vector of means $\mathbf{x}_d$ were known (available from sources external to the survey), μ_d would be estimated much more efficiently using $\mathbf{x}_d\hat{\boldsymbol{\beta}}$. Composite estimation bypasses this problem by searching for the combination of district-level means of auxiliary variables, whether known or estimated from the survey or from other sources, aiming directly to minimise the MSE of the combination (Longford 1999).

The approach developed in Section 3 can be adapted to distributions other than normal straightforwardly, so long as the kurtoses required for evaluating the district-level variance of $(\mu_d - \mu)^2$ and the sampling variance of $(\hat{\mu}_d - \mu)^2$ are known. In practice, kurtosis depends on the mean μ_d, creating difficulties that can be overcome only by approximations or averaging. Estimating proportions p_d with dichotomous data is a case in point. We have

$$\begin{aligned}\mathrm{var}\{(\hat{p}_d - p)^2\} &= \frac{v_d}{n_d^3}(1 - 3p_d + 3p_d^2) \\ &+ \frac{4v_d}{n_d^2}(1-2p_d)(p_d - p) + \frac{6v_d}{n_d}(p_d - p)^2 - \frac{v_d^2}{n_d^2},\end{aligned}$$

where $v_d = p_d(1-p_d)/n_d$ and p is the national proportion. The complex dependence on the poorly estimated p_d presents an analytical challenge that does not have a universal solution.

Throughout, we assumed that the value of the variance ratio ω is known. In practice, ω is estimated. It is difficult to take account of the uncertainty about ω analytically, but its impact on estimation of μ_d and $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$ can be assessed by sensitivity analysis which repeats the simulations described in Section 4 for a range of plausible values of ω. As one set of simulations takes about one minute of CPU time, this is a manageable computational task. One difficulty in such an assessment is that with an altered assumed value of ω the estimator $\tilde{\mu}_d$ is changed, and so the target of the composite MSE estimator is also changed. An alternative informal approach considers the consequences of under- and over-stating the value of ω. In estimating μ_d it is advisable to err on the side of greater ω, giving more weight to the direct estimator $\hat{\mu}_d$ (Longford 2005, Chapter 8). For estimating the MSE of $\tilde{\mu}_d$, we may prefer to err on the side of the more stable averaged estimator. That corresponds to increasing the value of the coefficient c_d^* and, as c_d^* is a decreasing function of ω, to reducing the value of ω used for setting c_d^*. Of course, this should be done in moderation, not to discard the contribution of the naïve estimator of MSE altogether.

5. Conclusion

The approach developed in this paper applies the general idea of shrinkage to estimation of MSE of small-area estimators and reduces the impact of averaging, regarded as undesirable when viewed from the design-based perspective, in which the country's districts have fixed population quantities μ_d. We have focussed on improvement in estimation of the MSE for each district separately. In practice, improvement of estimation for some districts is more important than for others. Many surveys are designed for inferences other than small-area estimation, or take small areas into account in planning only peripherally, and so they may yield more than satisfactory estimators for some districts, typically the most populous ones, and less satisfactory for others, often the sparsely populated districts. In such a setting, relatively higher inferential priority should be ascribed to the latter districts. Shrinkage estimators of small-area means and proportions have this property, and the simulations documented in Section 4 indicate that composite estimation of MSE has a similar property, at least in relation to the averaged estimator.

For a given size of the bias in estimating an MSE, we prefer the positive bias, because we regard understating the precision as statistically 'dishonest', whereas overstating it merely fails to present the estimate in the light it deserves - we undersell the results of our analytical effort. With this perspective, the optimal coefficient c_d in (7) should not be

sought by minimising the MSE of the combination, but by a criterion that regards underestimation of MSE as an error more severe than its overestimation by the same amount. Finding a suitable criterion for this, for which optimisation is tractable, is an open problem. The composite MSE estimator derived in Section 3 tends to overestimate the MSE, but this is not by our design.

We have experimented with ML and REML estimation; in the setting used for the simulations, the differences between the two approaches are minute. The advantage of unbiased estimation of the variance σ_B^2 is lost when $\hat{\sigma}_B^2$ is subjected to a non-linear transformation, and efficiency is maintained by transformations only asymptotically. However, small-area estimation is a quintessentially small-sample problem.

The approach presented in this paper illustrates the universality of the general idea of combining alternative estimators. The composite estimator exploits the strengths and reduces the drawbacks of the constituent estimators. Applying it is not detrimental when one of the estimators is far inferior to the other. As a form of averaging is involved even in the composite MSE estimator, it contributes to its robustness by ameliorating departures from the assumptions made in the theoretical development, such as heteroscedasticity and asymmetric (non-normal) within-district distributions.

Incorporating inferential priorities, in effect, redistributing the precision in estimating the MSEs for the small areas, is an open problem. A similar problem, designing surveys for small-area estimation so as to ensure sufficient precision in the model-based perspective (with averaging) is addressed by Longford (2006).

Acknowledgement

Partial support for the work on this manuscript by Grants SEC2003-04476 and SAB2004-0190 from the Spanish Ministry of Education and Science is acknowledged. Insightful and constructive comments of two referees and an Associate Editor are acknowledged.

References

EURAREA Consortium. (2004). EURAREA Project Final Reference Volume. Enhancing Small-Area Estimation Techniques to Meet European Needs. Office for National Statistics, London. Available from http://www.statistics.gov.uk/eurarea.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society*, Series A, 162, 227-245.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. New York: Springer-Verlag.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1988). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40.

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

Shen, W., and Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society*, Series B, 60, 455-471.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 81-85
Statistics Canada, Catalogue No. 12-001

Handling survey nonresponse in cluster sampling

Jun Shao [1]

Abstract

In surveys under cluster sampling, nonresponse on a variable is often dependent on a cluster level random effect and, hence, is nonignorable. Estimators of the population mean obtained by mean imputation or reweighting under the ignorable nonresponse assumption are then biased. We propose an unbiased estimator of the population mean by imputing or reweighting within each sampled cluster or a group of sampled clusters sharing some common feature. Some simulation results are presented to study the performance of the proposed estimator.

Key Words: Nonignorable nonresponse; Random-effect-based nonresponse; Imputation; Collapsing clusters.

1. Introduction

Nonresponse exists in most survey problems. The probability of having a nonrespondent in a survey item (variable) y typically depends on the unobserved value of y, which creates a great challenge in handling nonrespondents. Commonly used procedures for handling nonresponse (such as reweighting and imputation) are all based on the assumption that nonresponse is ignorable conditional on an auxiliary variable. More precisely,

$$P(y \text{ is a respondent} \mid y, z) = P(y \text{ is a respondent} \mid z), \quad (1)$$

where z is an auxiliary variable whose values are observed for all sampled units in the survey. That is, conditional on z, the value of y and its response status are statistically independent. Assumption (1) is referred to as the unconfounded response mechanism by Lee, Rancourt and Särndal (1994). Using the terminology in Rubin (1976), nonresponse under (1) is ignorable conditional on z.

There are situations in which it is difficult to find a variable z to satisfy (1). The purpose of this article is to study a method of handling nonresponse when cluster sampling is used, assuming that a variable z satisfying (1) is not available. In cluster sampling, sampling is carried out in two stages; the first stage sampled units are clusters containing units that are sampled in the second stage. Cluster sampling is used because of economic considerations. It is necessary when no reliable list of the second stage units in the population is available (for example, there is no complete list of people but a list of households is available). Under cluster sampling, the variable of interest y may be decomposed as $y = \mu + b + e$, where μ is an unknown overall mean of y, b is a cluster level random effect (all units in the same cluster share the same random effect b), and e is a within-cluster random effect. In many cases, the dependence of the value of y and its response status is through the unobserved cluster level random effect b:

$$P(y \text{ is a respondent} \mid y, b) = P(y \text{ is a respondent} \mid b), \quad (2)$$

i.e., if b were observed, then we would have assumption (1) with $z = b$. For example, suppose that clusters are households and a single person completes survey forms for all sampled persons in a household. It is likely that the response probability depends on the household level variable b, not on the within household variable e.

Assumption (2) was first used by Wu and Carroll (1988) in a health problem where the clusters have a longitudinal (repeated-measure) structure. They called (2) informative censoring (missing) and proposed a method under some parametric assumptions on the probability $P(y \text{ is a respondent} \mid b)$ and the distribution of y. Later, Little (1995) called this type of missing mechanism the nonignorable random-coefficient-based missing mechanism. Thus, assumption (2) will be referred to as nonignorable random-effect-based response mechanism. Since b is not observed, response mechanism (2) is actually nonignorable.

For survey data, it is difficult to impose any parametric model on the distribution of y. Furthermore, it is also difficult to fit a parametric model for the response mechanism under (2), since b is not observed. After introducing some details on the sampling design and our assumptions, we propose in Section 2 a method for the estimation of the population mean of y under response mechanism (2), without requiring a parametric model for the response mechanism. It is assumed that y follows a random (cluster) effect model, but there is no parametric assumption on the distribution of y. Results from a simulation study are presented in Section 3 for examining the performance of the proposed estimator. Some discussions are given in the last section.

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

2. Main results

Let S be a sample of clusters of size n from a population P. Within the i^{th} sampled cluster, let S_i be the second stage sample of size $m_i \geq 2$ from a population P_i. For sampled unit $j \in S_i$, a survey weight w_{ij} is constructed (from the specification of the sampling design) so that when there is no nonresponse, $\hat{Y} = \sum_{i\in S}\sum_{j\in S_i} w_{ij} y_{ij}$ is an unbiased estimator of the population total Y on any variable y, *i.e.*, $E_s(\hat{Y} - Y) = 0$, where y_{ij} is the y-value of unit j in cluster i $Y = \sum_{i\in P}\sum_{j\in P_i} y_{ij}$, and E_s is the expectation with respect to repeated sampling.

Let y be the variable of interest. We adopt an imputation model approach, *i.e.*, we assume that each y_{ij} in the population is a random variable with

$$y_{ij} = \mu_i + b_i + e_{ij}, \tag{3}$$

where μ_i is an unknown parameter, b_i is an unobserved cluster level random effect with mean 0 and a finite variance, e_{ij} is an unobserved within cluster random effect with mean 0 and a finite variance, and b_i's and e_{ij}'s are independent. Note that the distribution of y_{ij} may vary with (i, j).

Let δ_{ij} be the response indicator for y_{ij} ($\delta_{ij} = 1$ if y_{ij} is a respondent and $\delta_{ij} = 0$ if y_{ij} is a nonrespondent). We adopt the approach in Shao and Steel (1999), *i.e.*, δ_{ij} is defined for every unit in the population and nonresponse mechanism is part of the model. Let $\boldsymbol{\delta}_i$ be the vector containing δ_{ij}, $j \in S_i$, and $\boldsymbol{y}_i$ be the vector containing y_{ij}, $j \in S_i$. We assume the following nonignorable random-effect-based response mechanism: for every sample,

$$P_m(\boldsymbol{\delta}_i \mid b_i, \boldsymbol{y}_i) = P_m(\boldsymbol{\delta}_i \mid b_i), \quad i \in S, \tag{4}$$

where P_m is the probability with respect to the model and $P_m(\xi \mid \eta)$ denotes the conditional distribution of ξ given η. That is, conditional on b_i, $\boldsymbol{y}_i$ and $\boldsymbol{\delta}_i$ are independent. (Unconditionally, they may be dependent.) We assume that the stochastic mechanism with respect to the model is independent of the sampling mechanism so that $E_s E_m(X) = E_m E_s(X)$ as long as X is integrable, where E_m is the expectation with respect to P_m.

Furthermore, we assume that

$$\text{for any } i \in S, \text{ at least one } \delta_{ij} \text{ is 1.} \tag{5}$$

That is, each cluster has at least one respondent. Without this assumption (or some other assumption), the population total Y may not be estimable. More discussion is given in Section 4.

If we assume ignorable nonresponse, *i.e.*, $P_m(\delta_{ij} = 1 \mid y_{ij}) = P_m(\delta_{ij} = 1)$, then a commonly used procedure is to impute each nonrespondent by the mean $\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij} y_{ij} / \sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}$, which leads to the following estimator of Y:

$$\begin{aligned}\hat{Y}_r &= \sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij} y_{ij}, \ \tilde{w}_{ij} \\ &= w_{ij}\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij} \Big/ \sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right).\end{aligned} \tag{6}$$

Under assumptions (3)-(5),

$$\begin{aligned}&E_s E_m(\hat{Y}_r) \\ &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij}(\mu_i + b_i + e_{ij})\right) \\ &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij}\mu_i\right) + E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij} b_i\right),\end{aligned} \tag{7}$$

where the last equality follows from

$$\begin{aligned}E_m(\delta_{ij}\tilde{w}_{ij} e_{ij}) &= E_m[E_m(\delta_{ij}\tilde{w}_{ij} e_{ij} \mid b_i)] \\ &= E_m[E_m(\delta_{ij}\tilde{w}_{ij} \mid b_i) E_m(e_{ij} \mid b_i)] = 0\end{aligned} \tag{8}$$

under (4). The first term in (7) is equal to

$$E_s E_m\left[\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\mu_i\right)\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right) \Big/ \left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right)\right]$$

which is approximately equal to (when n is large)

$$\begin{aligned}&\frac{E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\mu_i\right) E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right)}{E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right)} \\ &= \frac{E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\mu_i E_m(\delta_{ij})\right) E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right)}{E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij} E_m(\delta_{ij})\right)}.\end{aligned}$$

Note that

$$E_s E_m(Y) = E_m(Y) = \sum_{i\in P}\sum_{j\in P_i}\mu_i = E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\mu_i\right).$$

Hence, either $\mu_i = \mu$ for all i or $E_m(\delta_{ij})$ does not depend on (i, j) implies that the expectation of the first term in (7) is approximately equal to the expectation of Y. However, $E_m(\delta_{ij}\tilde{w}_{ij} b_i) \neq 0$ in general, because δ_{ij} and b_i

are dependent. Thus, the second term in (7) is not 0 and, hence, $\hat{Y}_r$ defined by (6) is biased under the nonignorable random-effect-based nonresponse. This bias does not go away asymptotically as $n \to \infty$ and/or $m_i \to \infty$ for all *i*.

Recognizing that the problem with $\hat{Y}_r$ is that imputation is done over the entire sample whereas the nonresponse depends on a cluster level random effect, we can find an unbiased estimator by performing imputation within each cluster. This would have been a natural way of imputing if the cluster random effect b_i were observed. If we impute a nonrespondent y_{ij} in cluster i by the cluster mean $\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{j\in S_i}\delta_{ij}w_{ij}$, then the resulting estimator is

$$\hat{Y}_c = \sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\bar{w}_{ij}y_{ij},$$

with

$$\bar{w}_{ij} = w_{ij}\left(\sum_{j\in S_i} w_{ij} \Big/ \sum_{j\in S_i}\delta_{ij}w_{ij}\right). \qquad (9)$$

Assumption (5) ensures that $\bar{w}_{ij}$ is well defined. Note that

$$\begin{aligned} E_sE_m(\hat{Y}_c) &= E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\bar{w}_{ij}\,\mu_i\right) + E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\bar{w}_{ij}\,b_i\right) \\ &= E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,\mu_i\right) + E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,b_i\right) \\ &= E_m(Y), \end{aligned}$$

where the first equality follows from assumption (3) and the fact that, under assumption (4), result (8) still holds with $\tilde{w}_{ij}$ replaced by $\bar{w}_{ij}$, the second equality follows from the definition of $\bar{w}_{ij}$ and the fact that μ_i and b_i do not depend on *j*, and the last equality follows from $E_m(b_i) = 0$. Hence, $\hat{Y}_c$ is an unbiased estimator of *Y*.

Since imputation is done within each cluster, the estimator defined by (9) seems inefficient when some cluster sample sizes m_i are very small. This worry, however, is not necessary in the case where $w_{ij} = w_i$ for all *j* (*e.g.*, the second stage sampling is with equal probability). When $w_{ij} = w_i$ for all *j*, imputation leading to $\hat{Y}_c$ in (9) is actually done in a much larger class, a group of clusters sharing something in common. Let $\bar{\delta}_i = m_i^{-1}\sum_{j\in S_i}\delta_{ij}$ be the response rate within cluster *i* and let

$$G_l = \left\{i \in S : m_i = m, \bar{\delta}_i = k/m\right\}, \quad l = (k, m), k \le m. \qquad (10)$$

For each $l = (k,m)$, G_l in (10) is the group of sample clusters having the same $m_i = m$ and $\bar{\delta}_i = k$. If $w_{ij} = w_i$ for all *j*, then, for $i \in G_l$ with $l = (k, m)$,

$$\begin{aligned} \bar{w}_{ij} &= w_{ij}\left(\sum_{j\in S_i} w_{ij} \Big/ \sum_{j\in S_i}\delta_{ij}w_{ij}\right) \\ &= w_i\left(\sum_{j\in S_i} w_i \Big/ \sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_i/\bar{\delta}_i \\ &= w_i/(k/m) \\ &= w_i\left(\sum_{i\in G_l} m_iw_i\right)\Big/\left(\sum_{i\in G_l}\frac{k}{m}m_iw_i\right) \\ &= w_i\left(\sum_{i\in G_l} m_iw_i\right)\Big/\left(\sum_{i\in G_l}\bar{\delta}_i m_iw_i\right) \\ &= w_i\left(\sum_{i\in G_l}\sum_{j\in S_i} w_i\right)\Big/\left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_{ij}\left(\sum_{i\in G_l}\sum_{j\in S_i} w_{ij}\right)\Big/\left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}\right). \end{aligned}$$

Therefore, imputation leading to $\hat{Y}_c$ in (9) is actually done within each group G_l when $w_{ij} = w_i$ for all *j*, *i.e.*, a nonrespondent in S_i is imputed by the sample mean of the respondents in G_l, $\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}$.

When w_{ij} varies with *j* for some *i*'s, some additional conditions are needed in order to combine clusters. A discussion is given in Section 4.

We end this section with a discussion of variance estimation, since most surveys require a variance estimator for each point estimator. A variance formula or its approximation (as $n \to \infty$) for $\hat{Y}_c$ can be derived, which may require more details on the sampling design. When the first stage sample size *n* is large, $m_i \le m$ for all *i* and a fixed integer *m*, and n/N is small, where *N* is the size of *P*, we can apply the adjusted jackknife method as described in Rao and Shao (1992). More precisely, we can follow the following steps.

1. Create *n* jackknife replicates, where the i^{th} replicate is obtained by deleting the i^{th} cluster and adjusting the weights to $w_{kj}^{(i)}, k \ne i, i = 1, \ldots, n$, according to the sampling design. For example, if the first stage sampling is a stratified sampling, then $w_{kj}^{(i)} = w_{kj}$ if *k* and *i* are not in the same stratum and $w_{kj}^{(i)} = n_h w_{kj}/(n_h - 1)$ if *k* and *i* are in the same stratum *h*, where n_h is the stratum size.

2. Re-impute the nonrespondents in the i^{th} jackknife replicate using the respondents in the i^{th} jackknife replicate, $i = 1, \ldots, n$.

3. Compute $\hat{Y}_{c,i}$ the same as $\hat{Y}_c$ but based on the i^{th} re-imputed jackknife replicate, $i = 1, \ldots, n$.

4. Compute the jackknife variance estimator for $\hat{Y}_c$ using a standard jackknife formula (*e.g.*, Shao and Tu 1995, Chapter 6). For example, if the first stage sampling is a stratified sampling with *H* strata, then a jackknife variance estimator is

$$v = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i \in S_h} \left(\hat{Y}_{c,i} - \frac{1}{n} \sum_{k \in S} \hat{Y}_{c,k} \right)^2,$$

where S_h is the sample from the h^{th} stratum and n_h is the size of S_h.

3. Simulation results

We now present some results from a simulation study to examine the performance of the estimators $\hat{Y}_r$ and $\hat{Y}_c$.

We create a finite population similar to the elementary school teacher population in Maricopa County, Arizona (Lohr 1999, pages 446-447). The finite population contains 311 clusters (schools). In each cluster, the second stage units are teachers. The cluster size (the number of teachers) varies from 6 to 59 and, hence, the first stage sampling is an unequal probability sampling with probability proportional to cluster size. The first stage sampling is with replacement and the sample size is 31. The second stage sampling is a simple random sampling of size 6 (for any cluster) without replacement.

For each teacher, the variable of interest is the minutes spent per week in school on preparation. The values of y_{ij} for this variable in the simulation are generated according to model (3), where μ_i is the mean minutes spent per week in school on preparation for the i^{th} school, b_i is a random effect of the i^{th} school, and e_{ij} is a random effect of the j^{th} teacher in the i^{th} school. The values of μ_i's are the sample means in the data set in Lohr (1999, pages 446-447), which vary from 25.52 to 42.18 with a mean of 33.76 and a median of 33.47. The value of b_i is generated according to $b_i = 8.31(X_i - 2)$, where X_i has the gamma distribution with shape parameter 2 and scale parameter 1. The value of e_{ij} is generated from the normal distribution with mean 0 and standard deviation 2.27. The b_i's and e_{ij}'s are independently generated. The values of $y_{ij} = \mu_i + b_i + e_{ij}$ are generated in each simulation run so that we can evaluate the biases and standard errors of estimators using joint probability under sampling and models (3)-(5).

For sampled units, nonrespondents are generated according to (4) and (5). That is, each sampled cluster has one respondent and the response status of the rest of the sampled units in each cluster are independently determined by $P(y_{ij} \text{ is missing} \mid b_i) = e^{b_i - 1}/(1 + e^{b_i - 1})$. The mean nonresponse probability is 33.76%.

For the estimation of the finite population mean, a simulation of 1,000 runs shows that, when $\hat{Y}_r$ is used, the bias, standard error, and root mean squared error are -2.89, 1.32, and 3.17, respectively, and the relative bias $E(\hat{Y}_r - Y)/E(Y)$ is -8.5%; when $\hat{Y}_c$ is used, the bias, standard error, and root mean squared error are 0.12, 1.81, and 1.82, respectively, and the relative bias $E(\hat{Y}_c - Y)/E(Y)$ is 0.3%. This simulation result supports our theory, *i.e.*, $\hat{Y}_c$ is approximately unbiased but $\hat{Y}_r$ is biased. In this case, $\hat{Y}_c$ has a larger standard error than $\hat{Y}_r$, but $\hat{Y}_r$ has a much larger root mean squared error than $\hat{Y}_c$ due to its large bias.

4. Discussions

Without the assumption that each sampled cluster has at least one respondent, the population total may not be estimable unless some other assumption is added. Under the nonresponse mechanism (4), when all observations in a cluster are nonrespondents, no information in that cluster can be recovered from observed data in other clusters unless some additional assumption is made. For example, one may assume that the population of clusters with no respondent is similar to that of clusters with 1 respondent, in which case one can collapse clusters by distributing the weights of clusters with 0 respondent to the weights of clusters with 1 respondent. Another approach is to assume a model so that we can extrapolate results to clusters with no respondent.

The results in Section 2 are given for mean imputation. Extensions to some other imputation methods are straightforward. For example, if random hot deck imputation is considered, then our result leads to imputation within clusters (or G_l's). When there is a covariate x whose values are all observed, our result can be extended to regression imputation with model (3) modified to $y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}$. For unit nonresponse, our result can also be applied to re-weighting, *i.e.*, adjusting weights within clusters (or G_l's).

Our method is imputation model based. We assume random-effect model (3) and random-effect-based response mechanism (4). If model (4) does not hold, then $E_m(\delta_{ij} \tilde{w}_{ij} e_{ij}) \neq 0$ and our estimator $\hat{Y}_c$ has a bias with a magnitude depending on the size of $|E_m(\delta_{ij} \tilde{w}_{ij} e_{ij})|$. Similarly, $\hat{Y}_c$ is not valid if model (3) does not hold.

It is shown in Section 2 that the condition $w_{ij} = w_i$ for all j ensures that imputation is done within each G_l that is the group of clusters with the same size and response rate. For two-stage sampling, this condition is satisfied when the last stage sampling is with equal probability (*e.g.*, simple random sampling without replacement). For three-stage sampling, model (3) should be replaced by $y_{ijk} = \mu_{ij} + b_{ij} + e_{ijk}$ and b_i in (4) should be replaced by b_{ij}. The survey weight w_{ijk} satisfies $w_{ijk} = w_{ij}$ as long as the last stage sampling is with equal probability and our result still holds. In two-stage sampling with w_{ij} varying with j, we may perform imputation within a group of clusters that have the same $E_m(\boldsymbol{y}_i \mid \boldsymbol{\delta}_i)$. For example, suppose that, in addition to (3)-(5), $\mu_i = \mu, b_i$'s are independent and identically distributed (iid), and conditional on b_i, the components of $\boldsymbol{\delta}_i$ are iid. Then $E_m(b_i \mid \boldsymbol{\delta}_i) = E_m(b_i \mid \bar{\delta}_i)$ depending only on

the size of the cluster m_i and $\overline{\delta}_i$. Hence we can perform imputation within each G_l defined by (10).

Acknowledgments

This work was partially supported by the NCI Grant CA53786 and NSF Grant DMS-0404535. The author would like to thank Mr. Lei Xu for programming in the simulation study and two referees for their helpful comments.

References

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Duxbury Press, New York.

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

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

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

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

Wu, M.C., and Carroll, R.J. (1988). Estimation and comparisons of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 87-94
Statistics Canada, Catalogue No. 12-001

On an optimal controlled nearest proportional to size sampling scheme

Neeraj Tiwari, Arun Kumar Nigam and Ila Pant [1]

Abstract

The concept of 'nearest proportional to size sampling designs' originated by Gabler (1987) is used to obtain an optimal controlled sampling design, ensuring zero selection probabilities to non-preferred samples. Variance estimation for the proposed optimal controlled sampling design using the Yates-Grundy form of the Horvitz-Thompson estimator is discussed. The true sampling variance of the proposed procedure is compared with that of the existing optimal controlled and uncontrolled high entropy selection procedures. The utility of the proposed procedure is demonstrated with the help of examples.

Key Words: Controlled sampling; Non-preferred samples; Quadratic programming; High entropy variance.

1. Introduction

In many situations, some samples may be undesirable due to administrative inconvenience, long distance, similarity of units or cost considerations. Such samples are termed non-preferred samples and the technique for avoiding these samples is known as 'controlled selection' or 'controlled sampling'. This technique, originated by Goodman and Kish (1950) has received considerable attention in recent years due to its practical importance.

The technique of controlled sampling is most appropriate when financial or other considerations make it necessary to select a small number of large first stage units, such as hospitals, firms, schools *etc*., for inclusion in the study. The main purpose of controlled sampling is to increase the probability of selecting a preferred combination beyond that possible with stratified sampling, whilst simultaneously maintaining the initial selection probabilities for each unit of the population, thus preserving the property of a probability sample. This situation generally arises in field surveys where the practical considerations make selection of some units undesirable but it is necessary to follow probability sampling. Controls may be imposed to secure a proper distribution geographically or otherwise and to ensure adequate sample size for some subgroups of the population. Goodman and Kish (1950) considered the reduction of sampling variances of the key estimates as the principal objective of controlled selection, but they also cautioned that this might not always be attained. A real problem emphasizing the need for controls beyond stratification was also discussed by Goodman and Kish (1950, page 354) with the objective of selecting 21 primary sampling units to represent the North Central States. Hess and Srikantan (1966) used the data for the 1961 universe of nonfederal, short-term general medical hospitals in the United States to illustrate the applications of estimation and variance formulae for controlled selection. Waterton (1983) used the data available from a postal survey of Scottish school leavers carried out in 1977 to describe the advantages of controlled selection and compare the efficiency of controlled selection with multiple proportionate stratified random sampling (meaning the sampling scheme in which instead of one stratifying variable, many variables each of which is associated with the variable of interest y, are used by cross-classifying the population on the basis of these variables) and found the controlled selection to perform favourably.

Three different approaches have been advanced in the recent literature to implement controlled sampling. These are (i) using typical experimental design configurations, (ii) the method of emptying boxes and (iii) using linear programming approaches. While some researchers have used simple random sampling designs to construct the controlled sampling designs, one of the more popular strategies is the use of IPPS (inclusion probability proportional to size) sampling designs in conjunction with the Horvitz-Thompson (1952) estimator. To construct controlled simple random sampling designs, Chakrabarti (1963) and Avadhani and Sukhatme (1973) proposed the use of balanced incomplete block (BIB) designs with parameters $v = N$, $k = n$ and λ, where N is the population size and n is the sample size. Wynn (1977) and Foody and Hedayat (1977) used the BIB designs with repeated blocks for situations where non-trivial BIB designs do not exist. Gupta, Nigam, and Kumar (1982) studied controlled sampling designs with inclusion probabilities proportional to size and used BIB designs in conjunction with the Horvitz-Thompson estimator of the population total $Y(=\sum_{i=1}^{N} y_i$, where y_i is the value of the i^{th} unit of the population, U). Nigam, Kumar and Gupta (1984) used some configurations of different types of experimental designs, including BIB designs, to obtain controlled IPPS sampling plans with the

1. Neeraj Tiwari, Ila Pant, Department of Statistics, Kumaon University, S.S.J. Campus, Almora-263601, India. E-mail: kumarn_amo@yahoo.com; Arun Kumar Nigam, Institute of Applied Statistics & Development Studies, Lucknow-226017, India. E-mail: dr_aknigam@yahoo.com.

additional property $c\pi_i\pi_j \le \pi_{ij} \le \pi_i\pi_j$ for all $i \ne j = 1, ..., N$ and some positive constant c such that $0 < c < 1$, where π_i and π_{ij} denote first and second order inclusion probabilities, respectively. Hedayat and Lin (1980) and Hedayat, Lin, and Stufken (1989) used the method of 'emptying boxes' to construct controlled IPPS sampling designs with the additional property $0 < \pi_{ij} \le \pi_i\pi_j$, $i < j = 1, ..., N$. Srivastava and Saleh (1985) and Mukhopadhyay and Vijayan (1996) suggested the use of 't-designs' to replace simple random sampling without replacement (SRSWOR) designs to construct controlled sampling designs.

All the methods of controlled sampling discussed in the previous paragraph may be carried out manually with varying degrees of laboriousness, but none has exploited the advantage of modern computing. Using the simplex method in linear programming, Rao and Nigam (1990, 1992) proposed optimal controlled sampling designs that minimize the probability of selecting the non-preferred samples, while retaining certain properties of an associated uncontrolled plan. Utilizing the approach of Rao and Nigam (1990, 1992), Sitter and Skinner (1994) and Tiwari and Nigam (1998) used the simplex method in linear programming to solve multi-way stratification problems with 'controls beyond stratification'.

In the present article, we use quadratic programming to propose an optimal controlled sampling design which ensures that the probability of selecting non-preferred samples is exactly equal to zero, rather than minimizing it, without sacrificing the efficiency of the Horvitz-Thompson estimator based on an associated uncontrolled IPPS sampling plan. The idea of 'nearest proportional to size sampling designs', introduced by Gabler (1987), is used to construct the proposed design. The Microsoft Excel Solver of the Microsoft Office 2000 package is used to solve the quadratic programming problem. The applicability of the Horvitz-Thompson estimator to the proposed design is discussed. The true sampling variance of the estimate for the proposed design is empirically compared with the variances of the alternative optimal controlled designs of Rao and Nigam (1990, 1992) and uncontrolled high entropy selection procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In Section 3, some examples are considered to demonstrate the utility of the proposed procedure by comparing the probabilities of non-preferred samples and sampling variances of the estimates. Finally in Section 4, the findings of the paper are summarized.

2. The optimal controlled sampling design

In this section, we use the concept of 'nearest proportional to size sampling designs' to propose an optimal controlled IPPS sampling design that matches the original π_i values, satisfies the sufficient condition $\pi_{ij} \le \pi_i\pi_j$ for non-negativity of the Yates-Grundy (1953) form of the Horvitz-Thompson (HT) (1952) estimator of the variance and also ensures that the probability of selecting non-preferred samples is exactly equal to zero. Before coming to the proposed plan, we briefly describe the Misdzuno-Sen and Sampford IPPS designs which will be used in the proposed plan for obtaining the initial IPPS design $p(s)$.

2.1 The Midzuno-Sen and Sampford IPPS designs

To introduce the concept of IPPS designs, we assume that a known positive quantity, x_i, is associated with the value of the i^{th} unit of the population and there is reason to believe that the y_i's are approximately proportional to x_i's. Here x_i is assumed to be known for all units of the population and y_i is to be collected for sampled units. In IPPS sampling designs, π_i, the probability of including the i^{th} unit in a sample of size n, is np_i, where p_i is the single draw probability of selecting the i^{th} unit in the population (also known as the normal size measure of unit i), given by

$$p_i = \frac{x_i}{\sum_{j=1}^{N} x_j}, \quad i = 1, 2, ..., N.$$

We first describe the Midzuno-Sen IPPS scheme and then discuss Sampford's design.

The Midzuno-Sen (MS) (1952, 1953) scheme has a restriction that the probabilities of selecting the i^{th} unit in the population (p_i's) must satisfy the condition

$$\frac{1}{n}.\frac{n-1}{N-1} \le p_i \le \frac{1}{n}, \qquad i = 1, 2, ..., N. \quad (1)$$

If (1) is satisfied for the p_i values of the population under consideration, we apply the MS scheme to get an IPPS plan with the revised probabilities of selection, p_i^*'s, [also known as revised normal size measures] given by

$$p_i^* = n\,p_i\,.\,\frac{N-1}{N-n} - \frac{n-1}{N-n}, \qquad i = 1, 2, ..., N. \quad (2)$$

Now, supposing that the s^{th} sample consists of units $i_1, i_2, ..., i_n$, the probability of including these units in the s^{th} sample under the MS scheme is given by

$$p(s) = \pi_{i_1, i_2, ..., i_n} = \frac{1}{\binom{N-1}{n-1}} \left(p_{i_1}^* + p_{i_2}^* + ... + p_{i_n}^* \right). \quad (3)$$

However, due to restriction (1), the MS plan limits the applicability of the method to units that are rather similar in

size. Therefore, when the initial probabilities do not satisfy the condition of the MS plan, we suggest the use of Sampford's (1967) plan to obtain the initial IPPS design $p(s)$.

Using Sampford's scheme, the probability of including n units $i_1, i_2, ..., i_n$ in the s^{th} sample is given by

$$p(s) = \pi_{i_1, i_2, ..., i_n} = n K_n \lambda_{i_1} \lambda_{i_2} ... \lambda_{i_n} \left(1 - \sum_{u=1}^{n} p_{i_u}\right), \quad (4)$$

where $K_n = (\sum_{t=1}^{n} t L_{n-t} / n^t)^{-1}$, $\lambda_i = p_i/(1-p_i)$ for a set $S(m)$ of $m \le N$ different units, $i_1, i_2, ..., i_m$, and L_m is defined as

$$L_0 = 1,\ L_m = \sum_{S(m)} \lambda_{i_1} \lambda_{i_2} ... \lambda_{i_m} \quad (1 \le m \le N).$$

2.2 The proposed plan

Consider a population of N units. Suppose a sample of size n is to be selected from this population. The single draw selection probabilities of these N units of the population (p_i's) are known. Let S and S_1 denote respectively, the set of all possible samples and the set of non-preferred samples.

Given the selection probabilities for N units of the population, we first obtain an appropriate uncontrolled IPPS design $p(s)$, such as the Midzuno-Sen (1952, 1953) or Sampford (1967) design, as described in Section 2.1. After obtaining the initial IPPS design $p(s)$, the idea behind the proposed plan is to get rid of the non-preferred samples S_1 by confining ourselves to the set $S - S_1$ by introducing a new design $p_0(s)$ which assigns zero probability of selection to each of the non-preferred samples belonging to S_1, given by

$$p_0(s) = \begin{cases} \dfrac{p(s)}{1 - \sum_{s \in S_1} p(s)} & \text{for } s \in S - S_1 \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $p(s)$ is the initial uncontrolled IPPS sampling plan.

Consequently, $p_0(s)$ is no longer an IPPS design. So, applying the idea of Gabler (1987), we are interested in the 'nearest proportional to size sampling design' $p_1(s)$ in the sense that $p_1(s)$ minimizes the directed distance D from the sampling design $p_0(s)$ to the sampling design $p_1(s)$, defined as

$$D(p_0,\ p_1) = E_{p_0}\left[\frac{p_1}{p_0} - 1\right]^2 = \sum_{s \in S - S_1} \frac{p_1^2(s)}{p_0(s)} - 1 \quad (6)$$

subject to the following constraints:

$$\begin{aligned} &\text{(i)} && p_1(s) \ge 0, \\ &\text{(ii)} && \sum_{s \in S - S_1} p_1(s) = 1, \\ &\text{(iii)} && \sum_{s \ni i} p_1(s) = \pi_i, \\ &\text{(iv)} && \sum_{s \ni i,\ j} p_1(s) > 0 \quad \text{and} \\ &\text{(v)} && \sum_{s \ni i,\ j} p_1(s) \le \pi_i \pi_j. \end{aligned} \quad (7)$$

The ordering of the above five constraints is carried out in accordance with their necessity and desirability. Constraints (i) and (ii) are necessary for any probability sampling design. Constraint (iii), which requires that the selection probabilities in the old and new schemes remain unchanged, which ensures that the resultant design will be IPPS. This constraint is a very strong constraint and it affects the convergence properties of the proposed plan to a great extent. Constraint (iv) is highly desirable because it ensures unbiased estimation of the variance. Constraint (v) is desirable as it ensures the sufficient condition for non-negativity of the Yates-Grundy estimator of the variance.

The solution to the above quadratic programming problem, *viz.*, minimizing the objective function (6) subject to the constraints (7), provides us with the optimal controlled IPPS sampling plan that ensures zero probability of selection for the non-preferred samples. The proposed plan is as near as possible to the controlled design $p_0(s)$ defined in (5) and at the same time it achieves the same set of first order inclusion probabilities π_i, as for the original uncontrolled IPPS sampling plan $p(s)$. Due to the constraints (iv) and (v) in (7), the proposed plan also ensures the conditions $\pi_{ij} > 0$ and $\pi_{ij} \le \pi_i \pi_j$ for the Yates-Grundy estimator of the variance to be stable and non-negative.

The distance measure $D(p_0,\ p_1)$ defined in (6) is similar to the χ^2-statistic often employed in related problems and is also used by Cassel and S≡rndal (1972) and Gabler (1987). Other distance measures are also discussed by Takeuchi, Yanai and Mukherjee (1983). An alternative distance measure for the present discussion may be defined as

$$D(p_0,\ p_1) = \sum_s \frac{(p_0 - p_1)^2}{(p_0 + p_1)}. \quad (8)$$

When applied on different numerical problems considered by us, we found that the use of (8) gave similar results to (6) in convergence and efficiency and so we will give results using (6) as the distance measure.

While all the other controlled sampling plans discussed by earlier authors attempt to minimize the selection

probabilities of the non-preferred samples, the proposed plan completely excludes the possibility of selecting non-preferred samples by ensuring zero probability for them and at the same time it also ensures the non-negativity of the Yates-Grundy estimator of the variance. However, in some situations a feasible solution to the quadratic programming problem, satisfying all the constraints in (7), may not exist. Constraint (v) may then be relaxed. This may not guarantee the non-negativity of the Yates-Grundy form of the variance estimator. However, since the condition $\pi_{ij} \leq \pi_i \pi_j$ is sufficient for non-negativity of the Yates-Grundy estimator of the variance but not necessary for $n > 2$, as pointed out by Singh (1954), there will still be a possibility of obtaining a non-negative estimator of the variance. After relaxing the constraint (v) in (7), if the Yates-Grungy estimator of the variance comes out to be negative, an alternative variance estimator may be used. This has been demonstrated in Example 5 in Section 3. If even after relaxing constraint (v), a feasible solution of the quadratic programming problem is not found, constraint (iv) may also be relaxed and consequently an alternative variance estimator in place of the Yates-Grundy form of the HT variance estimator may be used. The effect of relaxing these constraints on efficiency of the proposed design is difficult to study, as after relaxing the non-negativity constraint (v) the Yates-Grundy estimator of the variance does not provide accurate results. Using the Yates-Grundy estimator of the variance, for some problems the variance estimate is smaller after relaxing constraint (v) [as in the case of Examples 2(a), 2(b) and 3(a) in Section 3] while for other problems it is larger [as in the case of Example 1(a), 1(b), 3(b), 4(a) and 4(b) in Section 3]. Relaxing a constraint leading to an increased variance estimate may be due to the inability of the Yates-Grundy form of the variance estimator to estimate the true sampling variance correctly, when the non-negativity condition is not satisfied.

The proposed method may also be considered superior to the earlier methods of optimal controlled selection in the sense that setting some samples to have zero selection probability is different from associating a cost with each sample and then trying to minimize the cost, the technique used in earlier approaches of controlled selection. The technique employed by the earlier authors for controlled selection was a crude approach giving some samples very high cost and others very low.

One limitation of the proposed plan is that it becomes impractical when $\binom{N}{n}$ is very large, as the process of enumeration of all possible samples and formation of the objective function and constraints becomes quite tedious. This limitation also holds for the optimum approach of Rao and Nigam (1990, 1992) and other controlled sampling approaches discussed in Section 1. However, with the advent of faster computing techniques and modern statistical packages, there may not be much difficulty in using the proposed procedure for moderately large populations. On the basis of the size of populations that we have considered in the empirical evaluation, we found that the proposed method can easily handle the controlled selection problems up to a population of 12 units and a sample of size 5. The proposed method may be used to select a small number of first-stage units from each of a large number of strata. This involves a solution of a series of quadratic programming problems, each of a reasonable size, provided the set of non-preferred samples is specified separately in each stratum.

As in the case of linear programming, there is no guarantee of convergence of a quadratic programming problem. Kuhn and Tucker (1951) have derived some necessary conditions for the optimum solution of a quadratic programming algorithm but no sufficient conditions exist for convergence. Therefore unless the Kuhn-Tucker conditions are satisfied in advance, there is no way of verifying whether a quadratic programming algorithm converges to an absolute (global) or relative (local) optimum. Also, there is no way to predict in advance that the solution of a quadratic programming problem exists or not.

2.3 Comparison of sampling variance of the estimate

To estimate the population mean $\bar{Y}(= N^{-1}\sum_{i=1}^{N} y_i)$ based on a sample s of size n, we use the HT estimator of $\bar{Y}$ defined as

$$\hat{\bar{Y}}_{\mathrm{HT}} = \sum_{i \in s} \frac{Y_i}{N\pi_i}. \qquad (9)$$

Sen (1953) and Yates and Grundy (1953) showed independently that for fixed size sampling designs, $\hat{\bar{Y}}_{\mathrm{HT}}$ has the variance

$$V(\hat{\bar{Y}}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{i<j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2, \quad (10)$$

and an unbiased estimator of $V(\hat{\bar{Y}}_{\mathrm{HT}})$ is given as

$$\hat{V}(\hat{\bar{Y}}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{i<j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \quad (11)$$

Constraint (v), when used in the proposed plan, ensures the non-negativity of the variance estimator (11).

To demonstrate the utility of the proposed procedure, we use the empirical examples given in Section 3 to compare the true sampling variance of the HT estimator for the proposed procedure obtained through (10) with variances of the HT estimator using the optimal controlled plan of Rao and Nigam (1990, 1992) and those of two uncontrolled high entropy (meaning the absence of any detectable pattern or

ordering in the selected sample units) procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In what follows, we reproduce the expressions for the variances of these two high entropy procedures.

The expression for variance of $\hat{\bar{Y}}_{\mathrm{HT}}$ correct to $O(N^{-2})$ using the procedure of Goodman and Kish (1950) is given as

$$V(\hat{\bar{Y}}_{\mathrm{HT}})_{\mathrm{GK}} = \frac{1}{nN^2}\left[\sum_{i\in U} p_i A_i^2 - (n-1)\sum_{i\in U} p_i^2 A_i^2\right] - \frac{n-1}{nN^2} \times\left[2\sum_{i\in U} p_i^3 A_i^2 - \sum_{i\in U} p_i^2 \sum_{i\in U} p_i^2 A_i^2 - 2\left(\sum_{i\in U} p_i^2 A_i\right)^2\right], \quad (12)$$

where $A_i = Y_i / p_i - Y$, $Y = \sum_{i=1}^{N} Y_i$ and U denotes the finite population of N units.

Recently, Brewer and Donadio (2003) derived the π_{ij}-free formula for the high entropy variance of the HT estimator. They showed that the performance of this variance estimator, under conditions of high entropy, was reasonably good for all populations. Their expression for the variance of the HT estimator is given by

$$V(\hat{\bar{Y}}_{\mathrm{HT}})_{\mathrm{BD}} = \frac{1}{N^2}\sum_{i\in U} \pi_i(1-c_i\pi_i)\left(\frac{Y_i}{\pi_i^{-1}} - \frac{Y}{n^{-1}}\right)^2, \quad (13)$$

where $c_i = (n-1)/\{n-(2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1}\sum_{k\in U}\pi_k^2\}$ for all $i \in U$, which appears to perform better than the other values of c_i suggested by them.

3. Examples

In this section, we consider some empirical examples to demonstrate the utility of the proposed procedure and compare it with the existing procedures of optimal controlled sampling. In the present discussion, we begin with the Midzuno-Sen (1952, 1953) IPPS design to demonstrate our procedure, as it is relatively easy to compute the probability of drawing every potential sample under this scheme. However, if the conditions of the Midzuno-Sen scheme are not satisfied, we demonstrate that other IPPS sampling without replacement procedures, such as the Sampford (1967) procedure, may also be used to obtain the initial IPPS design $p(s)$. The true sampling variance of the HT estimator under the proposed plan is also compared with that of the existing procedures of optimal controlled selection and uncontrolled high entropy selection procedures given by (12) and (13).

Example 1: Let us consider a population consisting of six villages, borrowed from Hedayat and Lin (1980). The set S of all possible samples consists of 20 samples each of size $n = 3$. Due to the considerations of travel, organization of fieldwork and cost considerations, Rao and Nigam (1990) identified the following 7 samples as non-preferred samples:

123; 126; 136; 146; 234; 236; 246

(a). The Y_i and p_i values associated with the six villages of the population are:

Y_i:	12	15	17	24	17	19
p_i:	0.14	0.14	0.15	0.16	0.22	0.19

Since the p_i values satisfy the condition (1), we apply the MS scheme (3) to get an IPPS plan with the revised normal size measures (p_i^*'s) given by (2).

Applying the method discussed in Section 2 and solving the resulting quadratic programming problem with the Microsoft Excel Solver of Microsoft Office 2000 package, we obtain the controlled IPPS plan given in Table 1.

Table 1 Optimal controlled IPPS plan corresponding to Midzuono-Sen (MS) and Sampford's (SAMP) schemes for Example 1

s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]	s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]
124	0.14	0.09	245	0.03	0.12
125	0.03	0.05	256	0.13	0.14
134	0.00	0.00	345	0.02	0.06
135	0.09	0.03	346	0.20	0.10
145	0.03	0.06	356	0.06	0.06
156	0.13	0.07	456	0.06	0.16
235	0.09	0.05			

This plan matches the original π_i values, satisfies the condition $\pi_{ij} \le \pi_i\pi_j$ and ensures that the probability of selecting non-preferred samples is exactly equal to zero. Obviously, due to the fulfillment of the condition $\pi_{ij} \le \pi_i\pi_j$, we can apply the Yates-Grundy form of the HT variance estimator for estimating the variance of the proposed plan.

We have also solved the above example, using plan (3) of Rao and Nigam (1990, page 809) with specified π_{ij}'s taken from the Sampford's plan [to be denoted by RN3] and their plan (4) [to be denoted by RN4]. Using the RN3 plan, the probability of non- preferred samples (ϕ) comes out to be 0.155253 and using the RN4 plan with $c = 0.005$, ϕ comes out to be zero, whereas the proposed plan always ensures zero probability to non-preferred samples.

The values of the true sampling variance of the HT estimator $[V(\hat{\bar{Y}}_{\mathrm{HT}})]$ for the proposed plan, the RN3 plan, the RN4 plan, the Randomized Systematic IPPS sampling plan of Goodman and Kish (1950) [to be denoted by GK] and the uncontrolled high entropy sampling plan of Brewer and Donadio (2003) [to be denoted by BD] are produced in the first row of Table 2. It is clear from Table 2 that the

proposed plan yields almost the same value of variance of the HT estimator as yielded by the RN4 plan. The value of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan is slightly higher than those obtained from the RN3, GK and BD plans. This increase in variance may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 2 Values of the true sampling variance of the HT estimator $[V(\hat{\bar{Y}}_{HT})]$ for the Proposed, RN3, RN4, GK and BD plans

$V(\hat{\bar{Y}}_{HT})$	RN3	RN4	GK	BD	PROPOSED PLAN
Ex1(a)					
$N = 6, n = 3$	2.93	4.02	3.03	2.92	4.06
Ex 1(b)					
$N = 6, n = 3$	4.76	5.07	4.89	4.15	4.78
Ex 2(a)					
$N = 7, n = 3$	4.48	5.01	4.61	4.45	3.56
Ex 2(b)					
$N = 7, n = 3$	11.97	14.52	12.25	11.44	9.49
Ex 3(a)					
$N = 8, n = 3$	4.85	4.29	4.96	4.86	3.90
Ex 3(b)					
$N = 8, n = 3$	7.29	8.43	7.74	7.37	8.17
Ex 4(a)					
$N = 8, n = 4$	3.19	3.46	3.23	3.15	3.75
Ex 4(b)					
$N = 8, n = 4$	2.41	2.53	2.54	2.38	2.25
Ex 5					
$N = 7, n = 4$	3.08	3.93	3.12	3.07	5.10

(b). Now suppose that the p_i values for the above population of 6 units are as follows:

p_i: 0.10 0.15 0.10 0.20 0.27 0.18

Since these values of p_i do not satisfy the condition (1) of the MS plan, we apply the Sampford (1967) plan to get the initial IPPS design $p(s)$ using (4).

Applying the method discussed in Section 2 and solving the resultant quadratic programming problem, we obtain the controlled IPPS plan given in Table 1. This plan again ensures zero probability to non-preferred samples and satisfies the non-negativity condition for the Yates-Grundy form of the HT variance estimator. This example was also solved by the RN3 and RN4 plans. The value of ϕ for the RN3 plan is 0.064135 and the value of ϕ for the RN4 plan with $c = 0.005$ is zero. The proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are produced in the second row of Table 2. The proposed plan appears to perform better than the RN4 and GK plans and quite close to other plans considered by us.

Further examples were constructed to analyze the performance of the proposed plan. The populations with Y_i and p_i values and the set of non-preferred samples for each population are summarized in the Appendix. The p_i values for Examples 2(a), 3(a) and 4(a) satisfy the condition (1) of Midzuno-Sen plan and hence for these examples the Midzuno-Sen IPPS plan is used to obtain the initial IPPS design $p(s)$. However, for Examples 2(b), 3(b) and 4(b) the p_i values do not satisfy this condition and therefore we apply the Sampford IPPS plan to obtain the initial IPPS design. The probabilities of non-preferred samples (ϕ) for these examples using the RN3 plan, the RN4 plan and the proposed method are produced in Table 3. Table 3 shows that while the RN3 and RN4 plans only attempt to minimize the probability of non-preferred samples, the proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan for the population summarized in the Appendix are given in Table 2. From Table 2 we conclude that for all the empirical problems considered by us, the proposed plan appears to perform better than or quite close to the RN3, RN4, GK and BD plans. The increase in variance of the estimate for the proposed plan in some cases may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 3 The probabilities of non-preferred samples using RN3, RN4 and Proposed plans

Probability of non-preferred samples (ϕ)	RN3 PLAN	RN4 PLAN	Proposed Plan
Example 2(a) $N = 7, n = 3$	0.06	0 ($c = 0.5$)	0
Example 2(b) $N = 7, n = 3$	0.05	0 ($c = 0.5$)	0
Example 3(a) $N = 8, n = 3$	0.12	0 ($c = 0.005$)	0
Example 3(b) $N = 8, n = 3$	0.17	0 ($c = 0.005$)	0
Example 4(a) $N = 8, n = 4$	0.05	0 ($c = 0.005$)	0
Example 4(b) $N = 8, n = 4$	0.13	0 ($c = 0.005$)	0
Example 5 $N = 7, n = 4$	0.30	0.1008 ($c = 0.5$)	0

Example 5: We now consider one more example to demonstrate the situation where the proposed plan fails to provide a feasible solution satisfying all the constraints in (7). In such situations, we have to drop a constraint in (7) to obtain a feasible solution of the related quadratic programming problem.

Consider a population of seven villages. Suppose a sample of size $n = 4$ is to be drawn from this population. There are 35 possible samples, out of which the following 14 are considered as non-preferred:

1234; 1236; 1246; 1346; 1357; 1456; 1567;
2345; 2346; 2456; 2567; 3456; 3567; 4567.

Suppose that the following p_i values are associated with the seven villages:

p_i: 0.14 0.13 0.15 0.13 0.16 0.15 0.14 .

Since the p_i values satisfy condition (1), we apply the MS plan (3) to obtain the initial IPPS design $p(s)$ and solve the quadratic programming problem by the method discussed in Section 2. However, no feasible solution of the related quadratic programming problem exists in this case. Consequently, we drop constraint (v) in (7) for this particular problem to obtain a feasible solution of the quadratic programming problem. The probabilities of non-preferred samples using the RN3 plan, the RN4 plan and the Proposed plan for this empirical problem are given in the last row of Table 3. The proposed plan again matches the original π_i values and ensures the probability of selecting the non-preferred samples exactly equal to zero. However, due to non-fulfillment of the condition $\pi_{ij} \le \pi_i \pi_j$ for this example, the non-negativity of the Yates-Grundy estimator of the variance is not ensured. The values of the true variance, $V(\hat{\bar{Y}}_{\mathrm{HT}})$, for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are produced in the last row of Table 2. The value of $V(\hat{\bar{Y}}_{\mathrm{HT}})$ for this empirical example using the proposed plan does not appear to be satisfactory. For such problems where constraint (v) is not satisfied, we suggest the use of alternative variance estimators in place of the Yates-Grundy variance estimator.

We have also solved one more example with $N = 9$ and $n = 4$ using both the Midzuno-Sen and Sampford's methods for obtaining the initial IPPS design $p(s)$. The details of these solutions are omitted for brevity and can be obtained from the authors.

4. Conclusion

We have proposed a quadratic programming approach to solve the controlled sampling problems ensuring zero probability to non-preferred samples. The concept of 'nearest proportional to size sampling designs' of Gabler (1987) is used to obtain the proposed plan. The approach is simple in concept and is very flexible in allowing for a range of different objective functions as well as in permitting a variety of constraints. The only limitation of the procedure is that it cannot be applied to large populations, as the computational process becomes quite tedious for large populations. The utility of the proposed procedure is demonstrated with the help of examples and its true sampling variance is empirically compared with that of existing controlled sampling plans and uncontrolled high entropy sampling procedures. The proposed plan performs suitably.

Acknowledgements

The authors are grateful to an Associate Editor and two referees for their valuable suggestions and constructive comments on an earlier version of this paper, which led to considerable improvement in presentation of this work.

Appendix

The populations for Example 2-4 with Y_i and p_i values and the set of non-preferred samples.

Example 2. $N = 7$, $n = 3$.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467.

Y_i:	12	15	17	24	17	19	25
(a). p_i:	0.12	0.12	0.13	0.14	0.20	0.15	0.14
(b). p_i:	0.08	0.08	0.16	0.11	0.24	0.20	0.13

Example 3. $N = 8$, $n = 3$.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467;
128; 178; 248; 458; 468; 478; 578.

Y_i:	12	15	17	24	17	19	25	18
(a). p_i:	0.10	0.10	0.11	0.12	0.18	0.13	0.12	0.14
(b). p_i:	0.05	0.09	0.20	0.15	0.10	0.11	0.12	0.18

Example 4. $N = 8$, $n = 4$.

Non-preferred samples: 1234; 1236; 1238; 1246; 1248; 1268; 1346;
1348; 1357; 1456; 1468; 1567; 1568; 1678;
2345; 2346; 2456; 2468; 2567; 2568; 2678;
3456; 3468; 3567; 3678; 4567; 4678; 5678.

Y_i:	12	15	17	24	17	19	25	18
(a). p_i:	0.11	0.11	0.12	0.13	0.17	0.12	0.11	0.13
(b). p_i:	0.09	0.09	0.18	0.11	0.12	0.14	0.17	0.10

References

Avadhani, M.S., and Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *International Statistical Review*, 41, 175-182.

Brewer, K.R.W., and Donadio, M.E. (2003). The high-entropy variance of the Hortivz-Thompson Estimator. *Survey Methodology*, 29, 189-196.

Cassel, C.M., and Särndal, C.-E. (1972). A model for studying robustness of estimators and informativeness of labels in sampling with varying probabilities. *Journal of Royal Statistical Society*, Series B, 34, 279-289.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *Journal of Indian Statistical Association*, 1, 78-85.

Foody, W., and Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *Annals of Statistics*, 5, 932-945.

Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics-Theory & Methods*, 16(4), 1117-1131.

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

Gupta, V.K., Nigam, A.K. and Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, 69, 191-196.

Hedayat, A., and Lin, B.Y. (1980). Controlled probability proportional to size sampling designs. Technical Report, *University of Illinois at Chicago*.

Hedayat, A., Lin, B.Y. and Stufken, J. (1989). The construction of IPPS sampling designs through a method of emptying boxes. *Annals of Statistics*, 17, 1886-1905.

Hess, I., and Srikantan, K.S. (1966). Some aspects of probability sampling technique of controlled selection. *Health Serv. Res. Summer 1966*, 8-52.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *Journal of American Statistical Association,* 47, 663-85.

Kuhn, H.W., and Tucker A.W. (1951). Non-linear programming. *Proceedings of Second Berkely Symposium on Mathematical Statistics and Probability*, 481-492.

Midzuno, H. (1952). On the sampling system with probability proportional to sums of sizes. *Annals of Institute of Statistics & Mathematics*, 3, 99-107.

Mukhopadhyay, P., and Vijayan, K. (1996). On controlled sampling designs. *Journal of Statistical Planning & Inference*, 52, 375-378.

Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society*, B, 46, 564-571.

Rao, J.N.K., and Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.

Rao, J.N.K., and Nigam, A.K. (1992). 'Optimal' controlled sampling: A unified approach. *International Statistical Review*, 60, 89-98.

Sampford, M.R. (1967). On sampling with replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.

Singh, D. (1954). On efficiency of sampling with varying probabilities without replacement. *Journal of Indian Society of Agricultural Statistics*, 6, 48-57.

Sitter, R.R., and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20, 65-73.

Srivastava, J., and Saleh, F. (1985). Need of *t*-designs in sampling theory. *Utilitas Mathematica*, 28, 5-17.

Takeuchi, K., Yanai, H. and Mukherjee, B.N. (1983). The Foundations of Multivariate Analysis. 1st Ed. New Delhi: Wiley Eastern Ltd.

Tiwari, N., and Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning & Inference*, 69, 89-100.

Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Wynn, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics*, 5, 414-418.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society*, B, 15, 253-261.

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 22, No. 3, 2006

Volume 22, No. 4, 2006

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 4, December/décembre 2006

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", *etc*.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω ; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.