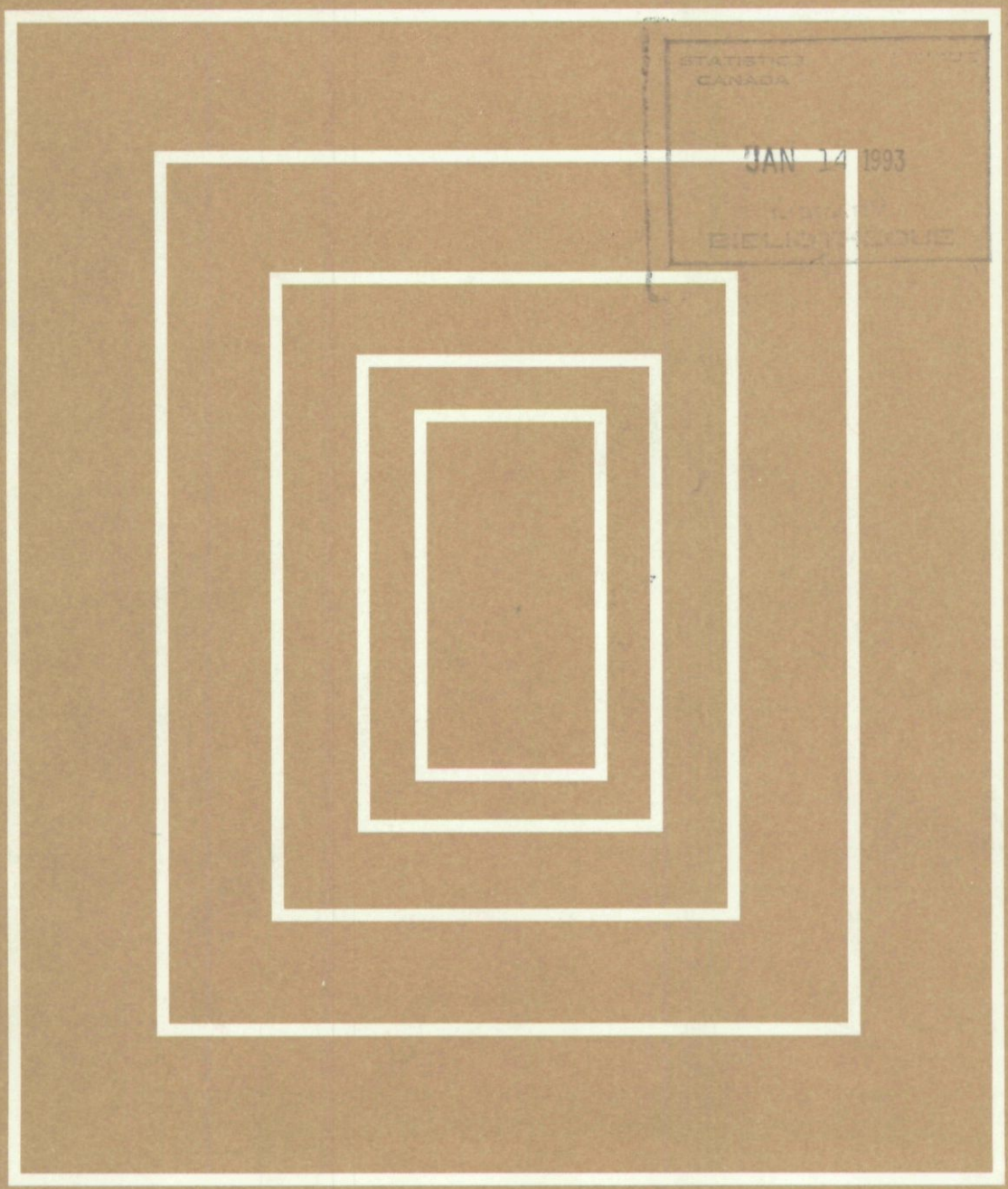
Catalogue 12-001

Survey Methodology

A Journal of Statistics Canada

December 1992 Volume 18 Number 2



Statistics Canada Statistique Canada

Canada

Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

December 1992 Volume 18 Number 2

Published by authority of the Minister
responsible for Statistics Canada

December 1992

Price: Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

The Survey Methodology Journal 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

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

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 18, Number 2, December 1992

CONTENTS

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 177–178
Statistics Canada

In This Issue

In August of 1991 a symposium in honour of Professor V.P. Godambe on the occasion of his 65th birthday was held at the University of Waterloo. Papers presented at this symposium were in the areas of foundations of inference, theory of estimation, and theory of survey sampling, all areas in which Professor Godambe has an interest and to which he has made significant contributions. The special section **Inference with Survey Data** in this issue, which is dedicated to Professor Godambe, contains some of the sampling related papers from the symposium. As a group these papers discuss many important issues for inference with survey data such as the role of modelling, robustness, complex survey designs, resampling methods, and the effects of imputation.

Royall considers model based estimation for finite population parameters. He describes the conflict between designs which provide model efficiency and those which are robust to model failure. Robustness is achieved through balanced samples. He presents a class of models for which the optimal sample is already balanced so that, for models in that class, there is no conflict between robustness and efficiency.

Smith and Njenga discuss model based and randomization based inference for sample surveys and suggest a robust non-parametric modelling approach to inference. Based on simulations using both real and synthetic data, they conclude that their estimator of a regression coefficient is robust to violations of assumptions of linearity and homoscedasticity, has good efficiency, and has reasonable conditional and unconditional properties.

Rao, Wu, and Yue review recent developments in resampling methods for complex survey designs, particularly the jackknife, balanced repeated replication, and the bootstrap. In a simulation study using a synthetic population they evaluate and compare variance estimators and confidence intervals for the population median.

Mantel considers model assisted estimation of a finite population mean based on a sample survey. He suggests that models should be extended so that the finite population mean is a known function of the optimal census based estimate of a model parameter. The extended model is then a compromise between model efficiency and finite population relevance.

Krieger and Pfeffermann discuss maximum likelihood estimation of model parameters. They describe various approaches in the literature and consider the problem of informative designs. They propose the use of weighted distributions where the weights are modelled as functions of the covariates and of the variable of interest. The approach performs reasonably well in a simulation study.

In the final paper of this special section Särndal considers the problem of variance estimation when imputation is used to complete a data set. Overall variance is derived as the sum of a sampling variance and an imputation variance. The suggested variance estimator is a design based estimator of the sampling variance with a model based correction for bias and a model based estimator of the imputation variance. Some examples and an empirical evaluation are presented.

Armstrong and Wu formulate the problem of sample allocation for a general two-phase survey design as a constrained programming problem. By exploiting its mathematical structure, they propose a solution that consists of iterations between two subproblems that are computationally much simpler. They provide empirical results showing that the proposed method works very well.

Couper and Groves examine whether experienced interviewers achieve higher response rates than inexperienced interviewers, controlling for differences in survey design and attributes of the population assigned to them. After demonstrating that the relationship is positive and curvilinear, they attempt to explain the mechanisms by which experienced interviewers achieve these rates and elaborate the nature of the relationship.

Lahiri and Wang propose new estimators for the "cost weights" and "relative importances" which are needed to construct the U.S. Consumer Price Index Numbers. The proposed estimators are composite estimators that combine information from relevant sources. A numerical comparison with four rival estimators is also presented.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 179–185
Statistics Canada

Robustness and Optimal Design Under Prediction Models for Finite Populations

RICHARD M. ROYALL[1]

ABSTRACT

In many finite population sampling problems the design that is optimal in the sense of minimizing the variance of the best linear unbiased estimator under a particular working model is bad in the sense of robustness – it leaves the estimator extremely vulnerable to bias if the working model is incorrect. However there are some important models under which one design provides both efficiency and robustness. We present a theorem that identifies such models and their optimal designs.

KEY WORDS: Balanced sample; Bias protection; Model failure; Working model.

1. INTRODUCTION

The "ratio estimator" of a finite population total $T = y_i + \ldots + y_N$ is $\hat{T} = N\bar{x}\bar{y}_s/\bar{x}_s$, where $\bar{x} = (x_1 + \ldots + x_N)/N$ is the known population mean of an auxiliary variable and $\bar{x}_s$ and $\bar{y}_s$ are sample means. This is the best linear unbiased (BLU) estimator of T under the model M:

$$E(Y_i) = \beta x_i,$$

$$\text{cov}(Y_i, Y_j) = \begin{cases} \sigma^2 x_i & i = j \\ 0 & \text{else.} \end{cases}$$

This estimator is biased under alternative models having different regression functions, in general, but protection against bias under specific alternatives can be assured by careful choice of the sample, as will be described below.

Throughout this paper we will be concerned with populations for which a particular model, such as M, is believed to apply, at least to a satisfactory degree of approximation. Our inferences will be made with reference to this model. For example, we will call an estimator $\hat{T}$ unbiased only if $E_M(\hat{T} - T) = 0$. On the other hand, we recognize that the model is an approximation and that it might be seriously wrong. Thus we describe it as a **working model**, and seek sampling and estimation procedures that are robust in the sense of performing well, not only under that working model, but also under alternative models that might better describe the relationships between variables in our population.

We denote by $M(\delta_0,\delta_1, \ldots \delta_J : v)$ the general polynomial regression model:

$$E(Y_i) = \sum_{j=0}^{J} \delta_j \beta_j x_i^j$$

[1] Richard M. Royall, Johns Hopkins University, Baltimore, MD 21205 U.S.A.

$$\operatorname{cov}(Y_i, Y_j) = \begin{cases} v_i \sigma^2 & i = j, \\ 0 & \text{else} \end{cases}$$

where δ_j is a zero-one indicator of whether the regressor x^j is included in the model. The best linear unbiased estimator under this model is denoted by $\hat{T}(\delta_0, \ldots, \delta_J : v)$. Thus our first model was $M(0, 1 : x)$, and $\hat{T}(0, 1 : x)$ is the ratio estimator.

Royall and Herson (1973) showed that $\hat{T}(0, 1 : x)$ remains unbiased under $M(\delta_0, \ldots, \delta_J : v)$ for any vector $(\delta_0, \ldots, \delta_J)$ of zeroes and ones, and any $v_1, \ldots, v_N$, if the sample is **balanced** on $x, x^2, \ldots, x^J$:

$$\sum_s x_i^j / n = \sum_1^N x_i^j / n \quad j = 1, 2, \ldots, J.$$

This means that in a balanced sample $\hat{T}(0, 1 : x)$ is robust in the sense that it remains unbiased under regression models that are much more general than the working model $M(0, 1 : x)$. Royall and Herson (1973, sec. 4.5) also detailed how approximate balance ensures the approximate unbiasedness of $\hat{T}(0, 1 : x)$. Furthermore they showed that in a balanced sample this estimator retains not only its unbiasedness but also its **optimality** under a wide variety of polynomial regression models, including $M(1 : 1)$, $M(1, 1 : x)$, and $M(0, 1, 1 : x^2)$. Specifically, the estimator is optimal under any polynomial regression model of degree J or less, provided only that the model's variance function is expressible as a linear combination of the regressors.

The robustness of the ratio estimator in balanced samples is achieved at a high cost in efficiency under the original working model $M(0, 1 : x)$. Under this model the sample that minimizes the variance consists of the n units whose x-values are largest, and the efficiency of a balanced sample is only $\bar{x}/\max_s(\bar{x}_s)$. (Royall and Herson 1973).

For the linear regression estimator, theoretical results have been established that are quite analogous to those sketched above for the ratio estimator, but with one important difference. The estimator is $\hat{T}(1, 1 : 1) = N[\bar{y}_s + b(\bar{x} - \bar{x}_s)]$, where $b = \sum_s (x_i - \bar{x}_s) y_i / \sum_s (x_i - \bar{x}_s)^2$. It is the optimal (BLU) estimator under the constant variance linear regression model, $M(1,1 : 1)$. When the sample is balanced, this estimator is robust, remaining unbiased (and optimal) under the same broad class of polynomial regression models as the ratio estimator. But unlike the ratio estimator, the regression estimator achieves robustness in balanced samples at **no cost in efficiency** - the variance under the working model $M(1, 1 : 1)$ is minimized in balanced samples, where $\bar{x}_s = \bar{x}$. This phenomenon occurs because the error variance $E(\hat{T} - T)^2$ is the sum of a constant and a term proportional to $(\bar{x} - \bar{x}_s)^2 \operatorname{var}(b)$. Minimizing $\operatorname{var}(b)$ requires maximizing $\sum_s (x_i - \bar{x}_s)^2$, but this term is eliminated altogether in samples with $\bar{x}_s = \bar{x}$.

Are there other models under which the same sample that minimizes the variance of the BLU estimator can also protect against bias under a wide range of alternative models? In particular, are there such models for problems requiring non-constant variance functions? We show that the answer is positive, giving a theorem that characterizes a family of models with the desired property and identifies the corresponding optimal samples. The results in this paper integrate and generalize those of Kott (1984) and Tallis (1986). They are also closely related to the work of Pereira and Rodrigues (1983) and Tam (1986), as well as that of Isaki and Fuller (1982).

2. BASIC RESULTS

It is convenient to shift to vector and matrix notation, in which Y is the population vector $(Y_1, Y_2, \ldots, Y_N)'$ and the model $M(X : V)$ specifies that $E(Y) = X\beta$ and $\text{var}(Y) = V\sigma^2$, where X is an $N \times p$ matrix of regressors, V is diagonal, and the vector β and the scalar σ^2 are unknown. For a given sample s of n units we list the sample units first, so that

$$Y = \begin{pmatrix} Y_s \\ Y_r \end{pmatrix}, \quad X = \begin{pmatrix} X_s \\ X_r \end{pmatrix}, \quad V = \begin{pmatrix} V_s & 0 \\ 0 & V_r \end{pmatrix},$$

where Y_r is the $(N - n)$-vector corresponding to the non-sample units, *etc.* We let 1_s and 1_r denote vectors $(1, \ldots 1)'$ of lengths n and $(N - n)$.

The population total is $T = 1_s'Y_s + 1_r'Y_r$. After the sample s is observed, the first component, $1_s'Y_s$, is known. The BLU estimator of T is obtained by adding to this known quantity the BLU predictor of $1_r'Y_r$:

$$\hat{T}(X : V) = 1_s'Y_s + 1_r'X_r\hat{\beta}(X : V),$$

where $\hat{\beta}(X : V) = (X_s'V_s^{-1}X_s)^{-1} X_s'V_s^{-1}Y_s$. The error variance is

$$\text{var}(\hat{T}(X : V) - T) = 1_r'(X_r'A_s^{-1}X_r + V_r)1_r\sigma^2,$$

where $A_s = X_s'V_s^{-1}X_s$. These formulas simplify when the vector $V1$ is in the linear manifold generated by the columns of X, which we denote by $\mathfrak{M}(X)$.

Lemma 1. If $V1 \in \mathfrak{M}(X)$ then

$$\hat{T}(X : V) = 1'X\hat{\beta}(X : V)$$

and under $M(X : V)$

$$\text{var}(\hat{T}(X : V) - T) = (1'XA_s^{-1}X'1 - 1'V1)\sigma^2.$$

Proof: The estimator simplifies because $V1 \in \mathfrak{M}(X)$ means that $V1 = Xc$ for some vector c, so that $X_s'1_s = X_s'V_s^{-1}X_sc$, from which we have $1_s'X_s\hat{\beta} = c'X_s'V_s^{-1}Y_s = 1_s'Y_s$. The variance formula follows from $\text{cov}(\hat{T}, T) = \text{cov}(1'X\hat{\beta}, 1_s'Y_s) = 1'XA_s^{-1}X_s'1_s = 1'Xc = 1'V1$.

Lemma 1 shows that for models with $V1 \in \mathfrak{M}(X)$, the sample affects the variance only through A_s^{-1}. This simplifies both the study of how the variance depends on the sample and the search for efficient samples.

The collection of samples that satisfy

$$1_s'W_s^{-1/2}X_s/n = 1'X/1'W^{1/2}1,$$

where W is an $N \times N$ matrix, will be denoted by $B(X : W)$. When W is the identity matrix, I, $B(X : I)$ is the collection of samples that are balanced on the columns of X. Royall and Herson (1973) proved that BLU estimators under a wide family of polynomial regression models are greatly simplified in balanced samples:

Theorem 1. Under $M(X : V)$ with $V1 \in \text{M}(X)$, if $s \in B(X : I)$ then

$$\hat{T}(X : V) = (N/n)1_s'Y_s$$
$$\text{var}(\hat{T}(X : V)) = [(N/n) - 1]1'V1\sigma^2. \tag{1}$$

The next theorem shows that if $V = I$ then the variance in (1) is the minimum possible, *i.e.* balanced samples $B(X : I)$, are optimal if $I1 \in \mathfrak{M}(X)$; it also identifies optimal samples for a class of models with more general variance structure.

Theorem 2. Under $M(X : V)$ if both $V1$ and $V^{1/2}1 \in \mathfrak{M}(X)$, then

$$\text{var}(\hat{T}(X : V) - T) \geq \left[(1'V^{1/2}1)^2/n - 1'V1\right]\sigma^2;$$

the bound is achieved if and only if $s \in B(X : V)$, in which case

$$\hat{T}(X : V) = (1'V^{1/2}1)(1_s'V_s^{-1/2}Y_s)/n.$$

Proof: Since $V1 \in \mathfrak{M}(X)$, the quantity to be minimized is $a'A_s^{-1}a$, where $a = X'1$ (Lemma 1). Now $V^{1/2}1 \in \mathfrak{M}(X)$ implies that there is a p-vector c_1 for which $V^{1/2}1 = Xc_1$ and, since V is diagonal, this ensures that $V_s^{1/2}1_s = X_sc_1$ for every sample s. From this it follows that $c_1'A_sc_1 = n$, and the desired inequality then follows from Schwarz's:

$$(a'A_s^{-1}a)(c_1'A_s\, c_1) = (a'A_s^{-1}a) \cdot n \geq (a'c_1)^2.$$

The necessary and sufficient condition for equality is $a' = kc_1'A_s$, where $k = 1'V^{1/2}1/n$. This is equivalent to $s \in B(X : V)$ because $c_1'A_s = 1_s'V_s^{-1/2}X_s$. The simple forms for the estimator $\hat{T}(X : V)$ and its variance are then easily obtained algebraically.

The formulas in Theorem 2 are familiar in conventional (randomization-based) sampling theory. The BLU estimator $\hat{T}(X : V)$ takes the simple form of the Horvitz-Thompson estimator $\hat{T}_{HT} = \sum_s y_i/\pi_i$, when π_i, the inclusion probability for unit i, is proportional to $v_i^{1/2}$. And the variance bound is the one established by Godambe and Joshi (1965, Theorem 6.1) for the model-based expectation of the random sampling variance.

Suppose that we have, for a working model $M(X : V)$ that satisfies the conditions of Theorem 2, an optimal sample s and BLU estimator $\hat{T}$. If we now consider a more general model $M(X, Z : V)$ with additional regressor(s) Z, the results of Theorem 2 continue to apply so long as the sample belongs to $B(Z : V)$ as well as to $B(X : V)$. Our sample and estimator remain optimal under the more general model, and the variance is unchanged. That is, we can maintain optimality under our working model (minimum variance sample and BLU estimator) and also protect against bias caused by the additional regressor(s) Z by imposing the additional constraint $B(Z : V)$ on the sample. This procedure not only protects our estimator from bias under $M(X, Z : V)$, it ensures that our sample and estimator both remain **optimal** under the more general model. Of course unbiasedness is ensured under the even more general model $M(X, Z : W)$, where W is any covariance matrix.

3. EXAMPLES

Four models have been particularly prominant in finite population sampling theory. In the polynomial regression model notation of section 1 these are $M(1 : 1)$, $M(1,1 : 1)$, $M(0,1 : x)$, and $M(0,1 : x^2)$. Optimal estimators under the first three models are the expansion, regression and ratio estimators, respectively. The optimal estimator under the fourth model,

$\hat{T}(0,1 : x^2) = \sum_s y_i + (N - n)\bar{x}_r \sum_s (y_i/nx_i)$, is approximated by the mean-of-ratios estimator $\hat{T}_{HT} = N\bar{x}\sum_s(y_i/nx_i)$ when the sampling fraction n/N is small.

One approach to finding a practical sampling and estimation strategy under one of these four working models is to use the best linear unbiased estimator under the model, while ensuring robustness by choosing a sample in which the estimator remains unbiased under more general polynomial regression models. For the first two models, $M(1 : 1)$ and $M(1, 1 : 1)$, we have seen that this strategy produces bias-robustness for free, at no cost in efficiency under the working model. Under both of these models bias protection requires simple (unweighted) balance; but the models satisfy the conditions of Theorem 2 with $V = I$, which implies that simple balance is optimal.

For the other two models, however, there is tension between robustness and efficiency. In section 1 we noted that under $M(0, 1 : x)$ the ratio estimator is optimal, and while the optimal sample consists of the n units maximizing $\bar{x}_s$, protection from bias under $M(1, 1 : x)$ requires a sample where $\bar{x}_s$ is not maximized but set equal to the population mean, $\bar{x}$. The situation under $M(0, 1 : x^2)$ is similar: the optimal sample is again the one where the sample mean $\bar{x}_s$ is maximized, but protection of the optimal estimator against bias under polynomial regression models requires an "overbalanced" sample, in which the sample mean equals $\sum_r x_i^2 / \sum_r x_i$ (Scott, Brewer and Ho 1978).

Under both of these models, $M(0, 1 : x)$ and $M(0, 1 : x^2)$, robustness can be achieved at a smaller cost in efficiency by starting with a more general working model. Theorem 2 shows the way. Consider first the model $M(0, 1 : x^2)$. If we use $\hat{T}(0, 1 : x^2)$ in an over-balanced sample, the error variance is $\{(N\bar{x})^2/n - \sum x_i^2 + \sum_s (x_i - \bar{x}_s)^2\}\sigma^2$. But if we use the more general working model $M(0, 1, 1 : x^2)$ and estimator $\hat{T}(0, 1, 1 : x^2)$, the theorem shows that any sample in which $\bar{x}_s = \sum x_i^2 / \sum x_i$ is optimal, yielding the minimum variance $\{(N\bar{x})^2/n - \sum x_1^2\}\sigma^2$. Now bias protection against even more general polynomial regression models can be obtained at no cost in efficiency by imposing the additional constraints of Condition $B(X : V)$ *i.e.* $\sum_s x_i^{j-1}/n = \sum_1^N x_i^j / \sum_1^N x_i \; j = 0, 3, \ldots, J$. Under these constraints on the sample, collectively called π-balance, $T(0, 1, 1 : x^2)$ is the mean-of-ratios estimator (Kott 1984). This sample and estimator remain optimal under all models of the form $M(\delta_0, 1, 1, \delta_3, \ldots, \delta_J : x^2)$.

Balanced samples $B(X : V)$ do not always exist. The above example illustrates this; when n becomes so large that $n/N > N(\bar{x}^2)/\sum x_i^2$ there can be no π-balanced sample, because otherwise the variance formula would become negative. Note that the condition $n/N > N(\bar{x}^2)/\sum x_i^2$ implies that $\max(x_i) > N\bar{x}/n$, so that in such populations there is no probability sampling plan with inclusion probability proportional to x.

To generalize the other model, $M(0.1 : x)$, so that the theorem will apply we can add a regressor, $x^{1/2}$:

$$E(Y_i) = \beta_{1/2} x_i^{1/2} + \beta_1 x_i$$

$$\text{var}(Y_i) = \sigma^2 x_i.$$

According to Theorem 2 any sample satisfying

$$\sum_s x_i^{1/2} \Big/ n = \sum_1^N x_i \Big/ \sum_1^N x_i^{1/2} \tag{2}$$

is optimal under this model, yielding the best linear unbiased estimator $\sum x_i^{1/2} \sum_s x_i^{-1/2} y_i / n$ and the minimum variance, $\{(\sum x_i^{1/2})^2/n - N\bar{x}\}\sigma^2$. This variance compares favorably with

that of the ratio estimator in a balanced sample, $N\bar{x}(N/n - 1)\sigma^2$. Now optimality of the sample and the estimator if in fact $E(Y_i) = \beta_0 + \beta_{1/2}x_i^{1/2} + \beta_1 x_i + \beta_2 x_i^2$ can be maintained (with no increase in variance) by imposing the additional conditions on the sample:

$$\sum_s x_i^{-1/2} \Big/ n = N \Big/ \sum_1^N x_i^{1/2}$$
$$\sum_s x_i^{3/2} \Big/ n = \sum_1^N x_i^2 \Big/ \sum_1^N x_i^{1/2}. \tag{3}$$

These conditions, (2) and (3), give the BLU estimator the simple form:

$$\sum_1^N x_i^{1/2} \sum_s (y_i/x_i^{1/2})/n,$$

which is of course the Horvitz-Thompson estimator for a probability-proportional-to-$x^{1/2}$ sampling plan.

4. PROBABILITY SAMPLING

The results in Section 2 are important in relation to an unobserved regressor Z. If Z were, like X, known for all population units, then we could use $M(X, Z : V)$ as the working model and $\hat{T}(X, Z : V)$ as the estimator in the first place. But suppose that we are unaware of the importance of Z and are using the working model $M(X : V)$ and the estimator $\hat{T}(X : V)$ when in fact $M(X, Z : V)$ applies. In this context we will refer to a sample from $B(X : V)$ as "balanced on X." Although we can choose a sample that is balanced on X, we cannot ensure that it will be balanced on Z, and if it is not, then our estimator is biased:

$$E(\hat{T}(X : V) - T) = \left[(1/n)(1'V^{1/2}1)(1_s'V_s^{-1/2}Z_s) - 1'Z\right]\gamma.$$

where γ is the Z-coefficient: $EY = X\beta + Z\gamma$.

Random sampling can help to provide protection against biases like this. If we use a probability sampling plan with inclusion probabilities, $\pi_i = nv_i^{1/2}/1'V^{1/2}1$, $i = 1, 2, \ldots, N$, then we will have balance on Z in expectation:

$$E_\pi 1_s'V_s^{-1/2}Z_s/n = 1'Z/1'V^{1/2}1,$$

the subscript π indicating that the expectation is with respect to the random sampling plan, not a prediction model. Furthermore, if our sampling plan is one under which $\mathrm{var}_\pi(1'V_s^{-1/2}Z_s/n)$ approaches zero as n grows, then the probability that we will draw a sample that is badly unbalanced, say one in which $|1_s'V_s^{-1/2}Z_s/n - 1'Z/1'V^{1/2}1| > \delta$, can be made small by taking a large enough sample, n. That is, probability sampling can provide balance on Z "in probability."

The strength of this result is in its scope–it applies for any matrix Z of regressors whatsoever. In particular it applies for the matrix X of regressors in our working model, as well as for

overlooked regressors. The weakness of course is that it applies to the sample selection process, not to a result of that process. The sample actually drawn will, with predictable frequency, be badly unbalanced on the known regressors X. If balance on X is important in a particular study, it should not be left to chance (This was documented empirically by Royall and Cumberland 1981). Restricted random sampling plans which guarantee that the selected sample will be balanced on X, such as Wallenius's "basket method" (1980), might represent a reasonable compromise strategy.

It sometimes happens that a regressor Z that is ignored when the sample is selected becomes available afterwards, as in the case of post-stratification for example. If it is determined that the selected sample is badly balanced on Z, then probability sampling has failed to provide the expected protection against bias under $M(X, Z : V)$; if it is too late to draw another sample, then to protect against the bias we must use an estimator that is unbiased under this model. That is, probability sampling does not guarantee approximate balance on Z; it only ensures that we have a good chance at approximate balance. It justifies confidence that a given sample is reasonably well balanced, in the absence of evidence to the contrary. It does not justify ignoring evidence of imbalance when it occurs.

Note that under the above probability sampling plan the estimator $(1' V^{\frac{1}{2}} 1)(1_s' V_s^{-\frac{1}{2}} Y_s)/n$, which is $\hat{T}(X : V)$ if both $V1$ and $V^{\frac{1}{2}}1$ belong to $\mathfrak{M}(X)$ and s is in $B(X : V)$, is unbiased with respect to the probability distribution generated by the sampling plan. But if the sample actually selected is not balanced on X (*i.e.* if s is not in $B(X : V)$) then this estimator is not unbiased under $M(X : V)$.

REFERENCES

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations - I. *Annals of Mathematical Statistics*, 36, 1707-1723.

ISAKI, C.T., and FULLER, W.A. (1987). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P.S. (1984). A fresh look at bias-robust estimation in a finite population. In *Proceedings of the Section Survey Research Methods, American Statistical Association*, 176-178.

PEREIRA, C.A., and RODRIGUES, J. (1983). Robust linear prediction in finite populations. *International Statistical Review*, 51, 293-300.

ROYALL, R.M., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 73, 66-77.

SCOTT, A.J., BREWER, K.R.W., and HO, W.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73, 359-361.

TALLIS, G.W. (1986). On the optimality of balanced sampling. *Statistics and Probability*, 4, 141-144.

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

WALLENIUS, K.T. (1980). Statistical methods in sole source contract negotiation. *Journal of Undergraduate Mathematics and Applications*, 0, 35-47.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 187–208
Statistics Canada

Robust Model-Based Methods for Analytic Surveys

T.M.F. SMITH and E. NJENGA[1]

ABSTRACT

This paper reviews the idea of robustness for randomisation and model-based inference for descriptive and analytic surveys. The lack of robustness for model-based procedures can be partially overcome by careful design. In this paper a robust model-based approach to analysis is proposed based on smoothing methods.

KEY WORDS: Analytic surveys; Robustness; Smoothing methods.

1. INTRODUCTION

The concept of robustness in finite population inference from both the randomisation and model-based viewpoints is examined. In his seminal paper on a unified theory of sampling from finite populations Godambe (1955) not only proved his famous non-existence theorem but also made suggestions for robust finite population inference. He proposed a superpopulation model for the unit variables y_i and suggested that strategies, that is the choice of both design and estimator, should be based on the model expectation of the sampling variance. He then imposed p-unbiasedness to obtain optimum strategies. These ideas were amplified in several papers including Godambe (1982) and Godambe and Thompson (1977). The results obtained include the apparent optimality of πps sampling and the Horvitz-Thompson (1952) estimator. But the inefficiency of this strategy in multipurpose surveys is well known so we find these results on optimality and robustness less convincing than the apparently negative results on the foundations of inference.

The lack of robustness of many model-based procedures is well known, see Hansen *et al.* (1983), and much of the work of Royall and his colleagues, for example Royall and Herson (1973a,b) has been devoted to constructing robust model-based strategies. After reviewing this work we propose a robust model-based method for estimating many complex statistics employed in the multivariate analysis of survey data which adjusts for the effects of selection. Our proposal is not a strategy but is a procedure which can be employed for the analysis of survey data after the sample is drawn.

2. FORMAL STRUCTURE

In order to examine robustness we must first structure finite population inference in the formal manner pioneered by Godambe (1955). We consider a population of N units with label set $U = \{1, 2, \ldots, N\}$. Attached to unit i is a vector of values, $\underset{\sim}{y}_i$, which will be measured on the sample units, and $\underset{\sim}{y}_U = (\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_N)$ denotes the finite population matrix of values. A sample, s, is a subset of U drawn according to some rule. We are concerned here with rules based only on prior information, $\underset{\sim}{z}_i$, available on all the units in the population. Let $\underset{\sim}{z}_U$ denote the prior information for the whole population, and let $p(s \mid \underset{\sim}{z}_U)$ denote the sampling rule.

[1] T.M.F. Smith, University of Southampton, United Kingdom; E. Njenga, Kenyatta University, Kenya.

Since the rule does not depend on $\underset{\sim}{y}_U$ it is uninformative. If $p(s \mid \underset{\sim}{z}_U)$ is a random sampling rule then it determines a probability distribution over ζ, the set of all samples, which is the basis for randomisation inference. The sample data comprises $d_s = \{(i,\underset{\sim}{y}_i): i \epsilon s\}$. Let y_s denote the matrix of sample values, then an estimator is a function of the data, d_s, and of the prior information, $\underset{\sim}{z}_U$, which includes auxiliary information. We denote by E_p, V_p, expectations and variances with respect to the distribution $p(s \mid \underset{\sim}{z}_U)$.

In a model-based approach it is further assumed that the population values $\underset{\sim}{y}_U$ are random variables. A major problem with this approach is to specify a parametric probability model for the joint distribution of all these random variables, which must be based on all the prior information including that on the structures of, and relationships between, the units in the population. So models must reflect hierarchical groupings (clusters) and block groupings (strata), as well as correlations between the variables. This structure is potentially so complex that attention is usually restricted to means and covariance matrices. In general let $f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U;\underset{\sim}{\lambda})$ denote the conditional finite population distribution, where $\underset{\sim}{\lambda}$ is a vector of unknown parameters. For predictive inference about finite population values, such as totals, this is a sufficient specification. For analytic inference about parameters in the marginal distribution of $\underset{\sim}{y}$ we must additionally specify the marginal distribution of the prior values $\underset{\sim}{z}_U$. Let $f(\underset{\sim}{z}_U;\underset{\sim}{\phi})$ denote this distribution, then the marginal distribution of $\underset{\sim}{y}_U$ is

$$f(\underset{\sim}{y}_U;\underset{\sim}{\theta}) = \int f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U;\underset{\sim}{\lambda}) f(\underset{\sim}{z}_U;\underset{\sim}{\phi}) d\underset{\sim}{z}_U, \tag{2.1}$$

where $\underset{\sim}{\theta} = g(\underset{\sim}{\lambda},\underset{\sim}{\phi})$ is the parameter of analytic interest.

Applying the sampling rule to the population generates the data, d_s. The joint distribution of the data, d_s, and prior values, $\underset{\sim}{z}_U$, is

$$\begin{aligned} f(d_s,\underset{\sim}{z}_U;\underset{\sim}{\lambda},\underset{\sim}{\phi}) &= p(s \mid \underset{\sim}{z}_U) \int f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U;\underset{\sim}{\lambda}) f(\underset{\sim}{z}_U;\underset{\sim}{\phi}) d\underset{\sim}{y}_{\bar{s}} \\ &= p(s \mid \underset{\sim}{z}_U)\, f(\underset{\sim}{y}_s \mid \underset{\sim}{z}_U;\underset{\sim}{\lambda})\, f(\underset{\sim}{z}_U;\underset{\sim}{\phi}), \end{aligned} \tag{2.2}$$

where $\bar{s}$ denotes units not in s. This distribution is the basis of a model-based approach to inference. We let E_m, V_m, denote expectations and variances with respect to the model.

An implication of (2.2) is that the sampling rule, $p(s \mid \underset{\sim}{z}_U)$, must be completely known to the person making the inference, as must the values of $\underset{\sim}{z}_U$. Absence of knowledge may render $p(s \mid \underset{\sim}{z}_U)$ informative about the unobserved values $\underset{\sim}{y}_{\bar{s}}$, see Scott (1977), Sugden and Smith (1984), in which case it cannot be taken outside the integral in (2.2).

In this general set-up, embracing both random selection and modelling of values, randomisation inference corresponds to the case where the values $\underset{\sim}{y}_U$ are unknown constants and the model distribution becomes degenerate at the point $\underset{\sim}{y}_U$. The only probability remaining is that in $p(s \mid \underset{\sim}{z}_U)$, and this distribution over the set ζ of all possible samples is the basis of randomisation inference. Note that the randomisation distribution is completely specified by knowledge of the sampling rule and of the prior values, $\underset{\sim}{z}_U$. It does not depend on any unknown parameters or on the survey values, $\underset{\sim}{y}_U$. This renders $p(s \mid \underset{\sim}{z}_U)$ uninformative because there is less information in $p(s \mid \underset{\sim}{z}_U)$ than in $\underset{\sim}{z}_U$ itself. This accounts for the negative nature of Godambe's results about randomisation inference.

In contrast model-based inference depends solely on the model component of (2.2), since $p(s \mid \underset{\sim}{z}_U)$ contains no information about $\underset{\sim}{y}_{\bar{s}}$. Predictive inferences about $\underset{\sim}{y}_{\bar{s}}$ are made using the conditional distribution, $f(\underset{\sim}{y}_u \mid \underset{\sim}{y}_s,\underset{\sim}{z}_U;\underset{\sim}{\lambda})$, independent of the randomisation distribution, $p(s \mid \underset{\sim}{z}_U)$. The sampling rule is still important at the design stage, for it affects efficiency and robustness, but it has no rôle to play at the inference stage. Random sampling also provides

a guarantee that the sampling rule is in fact uninformative, providing a scientifically acceptable sampling procedure. Model-based inferences may not be robust, however, because they may depend strongly on the choice of model, as demonstrated by many authors including Hansen *et al.* (1983).

A compromise solution is to employ both components of (2.2), the model and the randomisation distribution, in the choice of estimator. This was proposed by Godambe (1955) as a positive response to his negative results. He proposed using as a criterion the model expectation of the randomisation variance, namely $E_m\ V_p(t_s)$, where t_s is an estimator of a finite population total T. To find an optimum solution in a particular class of models Godambe restricted the choice of t_s to the class of p-unbiased estimators. This restriction has been much criticized and subsequently several authors, including Brewer (1979), Särndal (1980), Isaki and Fuller (1982), Little (1983), have proposed replacing exact unbiasednesses by some form of approximate unbiasedness. This is usually expressed in the form of asymptotic design unbiasedness which requires the construction of a hypothetical sequence of finite populations with sizes tending to infinity. Although one may feel unhappy with this mathematical construction the suggestion that strategies, chosen before drawing the sample, should be based on considerations of the average under a model of a repeated sampling procedure is perfectly acceptable. The controversial issue is the choice of distribution for making inferences after the sample has been drawn.

3. ROBUSTNESS

Robustness is not a well defined concept in statistics. The Encyclopedia of Statistical Sciences, (Kotz and Johnson 1988), states that:

> *"a robust procedure performs well not only under ideal conditions but also under departures from the ideal."*

It goes on to say that both the nature of departures from the ideal and the meaning of *"performs well"* must be specified. With this broad definition in mind we now examine robustness for randomisation and model-based inference for finite population totals. The general perception is that randomisation inference is robust and that model-based inference is not.

Godambe's negative results can be interpreted to mean that randomisation inference is impossible in general. This is certainly true for heterogeneous populations, such as Royall's axe, ass and box of horseshoes, or for populations with a few very extreme values, but for homogeneous populations the evidence overwhelmingly shows that randomisation inference is not only possible but also works in a well defined sense.

Employing randomisation inference implies abandoning certain statistical principles, such as the likelihood principle, and replacing them by an appeal to the central limit theorem. The assertion is that under repeated random sampling using the specified rule $p(s \mid z_U)$

$$\frac{t_s - T}{\hat{V}_p(t_s)} \sim N(0,1), \tag{3.1}$$

for any t_s which is approximately p-unbiased for T, where both N and n are large, but n/N is small. Although proved formally only under SRS and related schemes, empirical evidence shows that the randomisation coverage properties of 95% confidence intervals of the form

$$t_s \pm 1.96\sqrt{\hat{V}_p(t_s)}, \tag{3.2}$$

where $\hat{V}_p(t_s)$ is a consistent estimator of $V_p(t_s)$, are approximately correct except for extreme designs or heterogeneous populations.

Godambe and Thompson (1977) express their views about this approach in the following terms.

> *"The use of such a confidence interval may be interpreted as follows:*
>
> *I: We are fairly sure a priori that y belongs to that subset of R^N for which the interval covers $T(y)$ for 95% of all possible samples.*
>
> *II: There is no way that the sampled y-values, in conjunction with whatever other information we may have about the population, have altered the conviction in I. Thus even after sampling we believe that if the design were implemented again and again on this population the interval would cover $T(y)$ approximately 95% of the time.*
>
> *The robustness of the interval arises of course from the fact that only very weak and essentially informal conditions are required for the validity of its interpretation in the sense of I and II."*

Very similar views are expressed by Hansen *et al.* (1983).

> *"For probability-sampling designs the computed confidence intervals, for samples large enough, are valid in the sense that the randomization probability that the confidence intervals contain the value being estimated is equal to or greater than the nominal confidence coefficient, independent of the distribution of the characteristics among the elements of the population from which the sample is drawn."*
>
> *"Robustness is usually understood to mean that inferences made from a sample are insensitive to violations of the assumptions that have been made. In principle, and ordinarily in fact, robustness is achieved in probability-sampling surveys by the use of sampling with known probabilities (i.e., randomization) and consistent estimators, and using a large enough sample that the central limit theorem applies, so that the estimates can be regarded as approximately normally distributed."*

Note that this concept of robustness does not appear to require any specification of ideal conditions or of departures from the ideal. Random sampling and consistent estimation are all that is required. Brewer and Särndal (1983) are quite explicit:

> *"Probability sampling methods are robust by definition; since they do not appeal to a model, there is no need to discuss what happens under model breakdown."*

How can a statistical procedure be so robust?

The reason is that the entire procedure is under the control of the statistician, no attempt is made to introduce "nature" into the structure. The randomisation distribution has a known form and does not depend on unknown parameters. There is no need to make an inference about $p(s \mid z_U)$. Similarly the framework for inference is chosen by the statistician, it is repeated sampling using $p(s \mid z_U)$. Different statisticians may use different sampling rules and estimators but the procedure represented by (3.1) gives approximately correct coverage properties in every case, and so is robust. This is an example of criterion robustness. However, any given procedure may not be efficient for the totals of some variables. We have already highlighted the well known inefficiency of the Horvitz-Thompson estimator which occurs when

the survey variable is negatively correlated with the size variable. The search for efficiency robustness over a wide range of variables leads frequently to the recommendation that the design should be a stratified SRS design, see, for example, Godambe (1982), Hansen *et al.* (1983).

In model-based inference the statistician is playing the game of modelling "nature". Probability distributions such as $f(y_U \mid z_U;\underset{\sim}{\lambda})$ are chosen by the statistician but their true form is unknown, as also are the values of the parameters. If an estimator, t_s of T, is chosen then its expected value and variance will depend on the choice of model. Deviations from the model may lead to changes in the mean and variance and hence to changes in confidence intervals based on applying the central limit theorem to the model residuals. In model-based inference the robustness due to the central limit theorem is more limited than that in randomisation inference since it applies only to the residuals. Some model deviations can be controlled by choosing an appropriate design, as in Royall and Herson (1973a,b), but there can never be complete robustness. The framework for inference is also completely different. Instead of employing the unconditional distribution based on repeated sampling model-based inference employs the conditional distribution given the selected sample s.

Can these two positions ever be reconciled? Before sampling, when choosing strategies, they can. Both schools of thought have the same prior information, z_U, and both use models to suggest designs and estimators and choose strategies based on the overall mean squared error

$$E_m E_p (t_s - T)^2. \tag{3.3}$$

Randomisers usually impose a constraint such as approximate p-unbiasedness while modellers may impose approximate model unbiasedness and the two positions can be reconciled by choosing a sample design such that the model-unbiased estimator is also p-unbiased. This strategy utilizes the full structure of (2.2) and gets the best of both worlds.

After sampling there appears to be little hope of reconciliation. The two frameworks for inference are quite different, one being based on an unconditional distribution the other on a conditional distribution. Royall and Cumberland (1981) have demonstrated convincingly how much difference this can make. Incidentally they have also demonstrated the lack of robustness of some of the conventional model-based variance estimators.

One case where reconciliation is possible occurs in stratified sampling. Both randomisers and modellers have converged on stratified sampling as a robust design, and for SRS within strata model-based and p-based inferences coincide. This provides evidence for one of the few positive results in sample surveys:

Theorem: Stratification is a good thing.

Proof: See Cochran (1977, Ch.5).

Stratification allows us to look at the problem of robustness more closely. If both a randomiser and a modeller adopt the same stratification, and both also adopt the same SRS design within strata, then for a given sample they will both make identical inferences. Now suppose on the basis of further analysis or evidence it is agreed that an extra level of stratification should have been used. How does this affect the respective inferences? The modeller now has to say that the original model was misspecified and hence that inferences from that model would be biased. Both the estimator and the variance of the original model would be wrong. The randomiser, however, can say that the extra information is interesting, and could be used to post-stratify the original results, but that it can also be ignored if necessary because the original inferences are still valid in the sense defined in (3.2). All that has happened is a possible loss of efficiency. In one case the original inference is condemned as not being robust, in the other case the same

inference is apparently robust. The modellers bias, when averaged over repeated samples, is transformed for the randomiser into a component of sampling variance, or a loss of efficiency. So if initially randomisers and modellers start from the same position then deviations from that position are interpreted differently. In one case it is a bias in the other case a variance. Can this really be called robust in one case and not robust in the other?

4. ANALYTIC INFERENCE

In analytic inference the target for inference is no longer a known function of the finite population values, $\underset{\sim}{y}_U$, so that even if $n = N$ there is still residual uncertainty in the inference. Examples are tests of hypotheses, where the null hypothesis of no difference is meaningless in a fixed finite population. Possible targets for inference are the parameters $\underset{\sim}{\lambda}$, $\underset{\sim}{\phi}$, of the model (2.2), or functions of them such as $\underset{\sim}{\theta}$ in (2.1). Other targets are the parameters in finite populations related to the given finite population in some known way, perhaps through a spatial or time series structure. Methods for analytic inference have recently been reviewed by Skinner *et al.* (1989).

The starting point for analytic inference is the specification of the superpopulation model which aims to show how the finite population is related to the superpopulation. A common assumption is that the finite population is generated as IID random variables from a superpopulation. Whether this can be justified for populations with structure, such as clustering or stratification, is debatable. In this paper we assume that it is true, at least within broadly defined strata. With this assumption a SRS from the finite population is itself an IID sample from the superpopulation and inferences can be made directly from the sample to the superpopulation. If the sample is not a SRS, but is drawn using a design $p(s \mid \underset{\sim}{z}_U)$ which uses the information in $\underset{\sim}{z}_U$, then the achieved sample is no longer an IID sample from the superpopulation. This is the problem of selection and the effect of selection must be taken into account in the final inference.

The superpopulation model establishes a hierarchy,

$$\text{superpopulation} \supset \text{finite population} \supset \text{sample}.$$

If the finite population is IID from the superpopulation then finite population parameters, such as means, are related to the corresponding superpopulation parameters by

$$\bar{y}_U = E_m(\bar{y}_U) + O_p(N^{-1/2}). \tag{4.1}$$

Since N is usually very large an inference about $\bar{y}_U$ is a good approximation to an inference about $E_m(\bar{y}_U)$. Inferences about $\bar{y}_U$ using the p-weights associated with the sampling rule $p(s \mid \underset{\sim}{z}_U)$ are the basis of the randomisation approach to analytic inference. Note that this approach depends strongly on the IID assumption for the finite population.

For more complex analyses, such as logistic regression analysis, the pseudo-MLE approach in Skinner *et al.* (1989, sec. 3.4.4.) and Binder (1983) can be used to define both the finite population parameter of interest and the randomisation estimator. The finite population parameter is usually defined through an estimating equation, see Godambe (1960) and Godambe and Thompson (1986). As in Section 3 confidence intervals are based on the unconditional distribution generated by repeated random sampling.

Model-based analytic inference is based on the complete model of the survey population $\underset{\sim}{y}_U$, the design variables $\underset{\sim}{z}_U$, and the sample selection rule $p(s \mid \underset{\sim}{z}_U)$, that is

$$f(\underset{\sim}{y}_U, \underset{\sim}{z}_U, s; \underset{\sim}{\lambda}, \underset{\sim}{\phi}) = f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U; \underset{\sim}{\lambda}) f(\underset{\sim}{z}_U; \underset{\sim}{\phi}) p(s \mid \underset{\sim}{z}_U). \tag{4.2}$$

For random sampling rules the selection scheme leaves the conditional distribution $f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U; \underset{\sim}{\lambda})$ unchanged, but changes the marginal distribution of $\underset{\sim}{z}_U$ from $f(\underset{\sim}{z}_U; \underset{\sim}{\phi})$ before selection to

$$g_s(\underset{\sim}{z}_U; \underset{\sim}{\phi}) = f(\underset{\sim}{z}_U; \underset{\sim}{\phi}) p(s \mid \underset{\sim}{z}_U) \tag{4.3}$$

after selection. Thus inferences about $\underset{\sim}{\lambda}$ are unaffected by selection but inferences about ϕ, and hence about $\theta = g(\underset{\sim}{\lambda}, \underset{\sim}{\phi})$, the parameters of the marginal distribution $f(\underset{\sim}{y}_U; \underset{\sim}{\theta})$, are affected by selection. For these latter inferences the sample data cannot be treated as though it were a SRS from the superpopulation model.

If we assume that the superpopulation distributions are multivariate normal then

(i) $E(\underset{\sim}{y} \mid \underset{\sim}{z})$ is linear in $\underset{\sim}{z}$, and

(ii) $V(\underset{\sim}{y} \mid \underset{\sim}{z}) = \underset{\sim}{K}$, independent of $\underset{\sim}{z}$.

Under these assumptions of linearity and homoscedasticity a model-based estimator of the covariance matrix, $\underset{\sim}{\Sigma}_{yy}$, of $\underset{\sim}{y}$ is given by

$$\hat{\underset{\sim}{\Sigma}}_{yy} = \underset{\sim}{V}_{yys} + \underset{\sim}{b}_{yz} \quad (\underset{\sim}{V}_{zzu} - \underset{\sim}{V}_{zzs}) \underset{\sim}{b}_{yz}^T, \tag{4.4}$$

as shown in Skinner *et al.* (1989 Section 6.4), where $\underset{\sim}{V}_{yys}$, $\underset{\sim}{V}_{zzs}$, $\underset{\sim}{b}_{yz}$ are sample covariance matrices and a matrix of regression coefficients based on treating the sample data as IID from the conditional distribution $f(\underset{\sim}{y}_U \mid \underset{\sim}{z}_U; \underset{\sim}{\lambda})$. We call (4.4) the Pearson adjusted estimator after Pearson (1903).

Theoretical and empirical studies by Pfeffermann and Holmes (1985), Holmes (1987) and Njenga (1990), have shown that model-based inferences from (4.4) are not robust to departures from the assumptions of linearity and homoscedasticity. Nathan and Holt (1980) proposed a p-weighted version of (4.4) as a more robust alternative. This estimator is formed by replacing all the equally weighted sums in (4.4) by the corresponding p-weighted sums. The resulting estimator is called the probability weighted maximum likelihood estimator (*pwml*). The properties of this estimator have been studied empirically and theoretically in Holmes (1987), Njenga (1990) and in Skinner, Holt and Smith (1989, Ch.8). It was found to have similar unconditional properties to alternative p-weighted estimators, such as the Horvitz-Thompson estimator of $\underset{\sim}{\Sigma}_{yy}$, and superior conditional properties. In the simulation study in Section 6 the *pwml* estimator is taken to represent the entire class of p-weighted estimators. Since the p-weighted version of $\underset{\sim}{V}_{zzs}$ in (4.4) is a design consistent estimator of $\underset{\sim}{V}_{zzu}$ the resulting estimator is a design consistent estimator of $\underset{\sim}{\Sigma}_{yy}$. We now investigate a new robust model-based procedure.

5. A NONPARAMETRIC MOMENT-BASED ESTIMATOR

In this section we attempt to overcome the lack of robustness of model-based estimators such as (4.4) which depend strongly on assumptions of linearity and homoscedasticity. If the finite population is realized as IID observations from the superpopulation and if interest centres on the superpopulation parameters $\underset{\sim}{\mu}_y, \underset{\sim}{\Sigma}_{yy}$ in the marginal distribution of $\underset{\sim}{y}$, then the approach we adopt uses the fact that the sample data are IID from the conditional distribution $f(\underset{\sim}{y} \mid \underset{\sim}{z})$

while the design variables z_U are an IID sample of size N from the marginal distribution of z. For simplicity we assume that only one design variable has been used, such as a measure of size, so that z is a scalar random variable.

We assume that the conditional mean and covariance matrix of y given z are smooth functions of z of unknown form. Let

$$E(y \mid z) = \mu(z), \tag{5.1}$$

$$V(y \mid z) = \Sigma_{yy}(z). \tag{5.2}$$

These parametric functions can be estimated using some form of nonparametric estimation such as linear smoothing. Examples of linear smoothing methods are kernel estimation, see, for example, Gasser and Muller (1979), local regression, see, for example, Cleveland (1979), and smoothing splines, see, for example, Silverman (1985). We propose estimating the functions in (5.1) term by term using the kernel estimator

$$\hat{\mu}(z) = \sum_{j \in s} W_k(z,z_j) y_j. \tag{5.3}$$

We constrain the sum of the weights to be unity so that the estimator is a weighted average and employ the Gaussian kernel with k being the bandwidth. These estimators have been extensively studied and a recent review is Gasser and Engel (1990).

The structure in (5.1) and (5.2) implicitly assumes that we can write

$$y_j = \mu(z_j) + \epsilon_j, \quad j \in s, \tag{5.4}$$

so that

$$\hat{\epsilon}_j = y_j - \hat{\mu}(z_j), \quad j \in s. \tag{5.5}$$

Thus

$$\hat{\epsilon}_j \hat{\epsilon}_j^T = (y_j - \hat{\mu}(z_j))(y_j - \hat{\mu}(z_j))^T \tag{5.6}$$

is an estimator of $\Sigma_{yy}(z_j)$. Applying a linear smoother to each term $\sigma_{ab}(z_j)$ of $\Sigma_{yy}(z_j)$ gives

$$\hat{\sigma}_{ab}(z) = \sum_{j \in s} W_h(z,z_j) \hat{\epsilon}_{ja} \hat{\epsilon}_{jb}, \tag{5.7}$$

where $W_h(z,z_j)$ is a kernel with band width h which will usually be wider than the band width k chosen for the estimation of the conditional mean, (5.3).

The estimates of the marginal moments then employ the standard results that

$$\mu_y = E_z(\mu(z)), \tag{5.8}$$

$$\Sigma_{yy} = E_z(\Sigma_{yy}(z)) + V_z(\mu(z)). \tag{5.9}$$

Now

$$\underset{\sim}{\mu}_y = \int \underset{\sim}{\mu}(z) f(z) dz,$$

and our proposed estimator is

$$\hat{\underset{\sim}{\mu}}_y = \int \hat{\underset{\sim}{\mu}}(z) \hat{f}(z) dz. \tag{5.10}$$

Since N is large we propose using the empirical p.d.f. (Parzen 1962), given by

$$d\hat{F}(z) = \hat{f}(z) = 1/N, \quad \text{if} \quad z = z_j, \quad j = 1, \ldots, N, \tag{5.11}$$

$$= 0 \quad , \quad \text{otherwise} .$$

Substituting in (5.10) gives the estimator

$$\hat{\underset{\sim}{\mu}}_y = N^{-1} \sum_{j=1}^{N} \hat{\underset{\sim}{\mu}}(z_j). \tag{5.12}$$

To estimate $\underset{\sim}{\Sigma}_{yy}$ we adopt a similar procedure for the first term of (5.9). The second term can be written

$$V_z(\underset{\sim}{\mu}(z)) = \int (\underset{\sim}{\mu}(z) - \underset{\sim}{\mu}_y)(\underset{\sim}{\mu}(z) - \underset{\sim}{\mu}_y)^T f(z) dz. \tag{5.13}$$

For our estimator we propose

$$\hat{V}_z(\underset{\sim}{\mu}(z)) = N^{-1} \sum_{j=1}^{N} (\hat{\underset{\sim}{\mu}}(z_j) - \hat{\underset{\sim}{\mu}}_y)(\hat{\underset{\sim}{\mu}}(z_j) - \hat{\underset{\sim}{\mu}}(y))^T. \tag{5.14}$$

Thus the proposed estimator of is $\underset{\sim}{\Sigma}_{yy}$ is

$$\hat{\underset{\sim}{\Sigma}}_{yy} = N^{-1} \left[\sum_{j=1}^{N} \{\hat{\underset{\sim}{\Sigma}}_{yy}(z_j) + (\hat{\underset{\sim}{\mu}}(z_j) - \hat{\underset{\sim}{\mu}}_y)(\hat{\underset{\sim}{\mu}}(z_j) - \hat{\underset{\sim}{\mu}}_y)^T \} \right]. \tag{5.15}$$

Njenga (1990) examines the asymptotic statistical properties of these estimators.

One of the main reasons for estimating $\underset{\sim}{\Sigma}_{yy}$ is to carry out some form of multivariate analysis, such as a regression analysis between two or more of the components of $\underset{\sim}{y}$. In the next section we report the results of a simulation study in which the simple regression coefficient between two y-variables is estimated from stratified random samples with different sampling fractions.

6. ESTIMATING A REGRESSION COEFFICIENT A SIMULATION STUDY

Let $\underset{\sim}{y} = (y_1, y_2)^T$ with mean $\underset{\sim}{\mu}_y = (\mu_1, \mu_2)^T$ and covariance matrix

$$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

We are interested in estimating a function of $\underset{\sim}{\Sigma}_{yy}$, the simple linear regression coefficient,

$$B_{12} = \sigma_{12}/\sigma_2^2. \tag{6.1}$$

The elements of $\underset{\sim}{\Sigma}_{yy}$ will be estimated using:

(i) the Pearson adjusted estimator of $\underset{\sim}{\Sigma}_{yy}$ based on (4.4),

(ii) the probability weighted version of (4.4),

(iii) a kernel estimator based on (5.14).

The corresponding estimators of B_{12}, or of its finite population equivalent B_{12U}, are denoted $\hat{B}_{12,ml}$, $\hat{B}_{12,pwml}$ and $\hat{B}_{12,nw}$ respectively. The estimator $\hat{B}_{12,ml}$ is indexed by *"ml"* because it is also the MLE under a multivariate normal model. The estimator $B_{12,nw}$ is indexed *"nw"* after Nadaraya (1964) and Watson (1964). The first two estimators were chosen because of their good performance in previous simulation studies, see Skinner *et al.* (1989, Ch.8).

We carried out three types of simulation study. In the first simulation study we generated a multivariate normal population to compare the performance of the new estimator with the maximum likelihood estimator which is optimal for this population. In the second simulation study we generated a quadratic homoscedastic population to compare the estimators when only the linearity assumption is violated. In the last simulation study we compared the estimators when the structure of the population is unknown, *i.e.* we used a 'real' population. In these simulation studies we carried out both conditional and unconditional analyses. The former allow us to assess whether a particular estimator is good in some samples and poor for others whereas the latter averages over all possible samples for a particular design.

The new estimator uses the Gaussian Kernel

$$W_k(z_i,z_j) = c_i \exp\{-(z_i - z_j)^2/2k^2\}, \quad i \in U, \quad j \in s,$$

where $c_i = 1/\sum_{j\in s} \exp\{-(z_i - z_j)^2/2k^2\}$. A simulation with different values of the band width k showed that the mean squared error was relatively constant for a wide range of values of k and that this was achieved by trading off bias against variance. We selected values for k that gave relatively small values for the bias for each stratified sample design.

Since the 'real' population available to us was 6,962 observations from the 1975 UK Family Expenditure Survey we constructed all three populations to be of this size with mean vector and covariance matrix

$$\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_z \end{bmatrix}, \quad \underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1z} \\ & \sigma_2^2, & \sigma_{2z} \\ & & \sigma_z^2 \end{bmatrix}.$$

The actual values of $\underset{\sim}{\Sigma}$ are shown in Table 6.1.

The design variable is based on the expenditure on food, the independent variable is the total income and the dependent variable is the total expenditure. This finite population was stratified into five strata according to increasing values of the design variable, such that the first stratum contains 1,393 units with lowest values of z, second, third, fourth contain 1,392 units each and the fifth contains the last 1,393 units with the highest z values.

Table 6.1
Parameter Values from the Real Population

Variable		S.D.	Correlation matrix		
y_1	Expenditure on all items	0.668	1		
y_2	Total income	0.849	0.75	1	
z	Expenditure on food	0.658	0.41	0.28	1

Table 6.2
Stratified Sample Designs

Sample design		n_1	n_2	n_3	n_4	n_5	Symbol
D1	Proportional allocation	20	20	20	20	20	Δ
D2	Increasing allocation	5	9	16	30	40	∇
D3	U-shaped allocation	40	8	4	8	40	+

The sample designs used were based on those used by Holt, Smith and Winter (1980). Denote a stratified random sampling design by $(n_1 \ldots n_5)$ with n_h units selected from the h^{th} stratum, $h = 1, \ldots, 5$, then the designs are shown in Table 6.2, together with the symbols used in the plots.

For the various stratified sample designs we selected 1,000 independent samples of size 100 from the finite population. The sampling distribution of the various statistics under investigation were estimated from these 1,000 repeated samples. We obtain the unconditional results by averaging the statistics under investigation over all the 1,000 samples.

To assess the conditional properties of the estimators the 1,000 samples were divided into 20 groups of 50 samples each according to increasing values of $\Delta_{zz}^F = (S_{zzs} - S_{zz})/S_{zz}$ for the *nw* and *ml* estimators where

$$S_{zz} = N^{-1}\sum_U (z_i - \bar{z}_U)^2, \quad S_{zzs} = n^{-1}\sum_s (z_i - \bar{z}_s)^2,$$

$$\bar{z}_U = N^{-1}\sum_U z_i, \quad \bar{z}_s = n^{-1}\sum_s z_i,$$

and of $\Delta_{zz}^{*F} = (S_{zzs}^* - S_{zz})/S_{zz}$ for the *pwml* estimators where

$$S_{zzs}^* = \sum_s w_i (z_i - \bar{z}_s^*)^2, \quad \bar{z}_s^* = \sum_s w_i z_i, \quad w_i = (N\pi_i)^{-1} \quad \text{and} \quad \pi_i$$

denotes the probability of including the i^{th} unit in the sample such that the first group contained the 50 samples with the smallest values of Δ_{zz}^F (or Δ_{zz}^{*F}) and so on up to the 20th group which contains the 50 samples with the largest values of Δ_{zz}^F (or Δ_{zz}^{*F}). We assume that the variation in Δ_{zz}^F (or Δ_{zz}^{*F}) within each group is small. The conditional distribution of the various estimators given Δ_{zz}^F (or Δ_{zz}^{*F}) can then be plotted.

The biases, standard deviations and mean square errors reported in simulation studies 1 and 2 are computed around the value of B_{12U} in the finite population generated from the model. This enables them to be compared with the values generated from the real finite population in simulation study 3.

Table 6.3

Unconditional Absolute Biases of the Three Estimators of B_{12}
$N = 6{,}962$, $n = 100$ True Value $B_{12} = 0.595$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0003	0.0003	0.0185
D2	0.0007	0.0019	0.0269
D3	0.0026	0.0018	0.0159

Table 6.4

Unconditional Standard Deviation of the Three Estimators of B_{12}

Sample design	Standard deviations		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0500	0.0500	0.0507
D2	0.0522	0.0693	0.0531
D3	0.0486	0.0710	0.0503

Table 6.5

Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0025	0.0025	0.0029
D2	0.0027	0.0048	0.0035
D3	0.0024	0.0050	0.0028

Simulation Study 1

In the first simulation study the 6,962 finite population values were generated from a multivariate normal distribution with correlation matrix given in Table 6.1. These data should be favourable to the estimator $\hat{B}_{12,ml}$.

The unconditional biases, standard deviations and mean squared errors are shown in Tables 6.3, 6.4 and 6.5.

As expected the estimator $\hat{B}_{12,ml}$ is best in terms of mean squared error. The new estimator $\hat{B}_{12,nw}$ does surprisingly well, it has a large bias but a similar standard deviation. The size of the bias for a very smooth (linear) population is consistent with the results in other studies, see Gasser and Engel (1990). A very wide bandwidth is needed to capture a very smooth function.

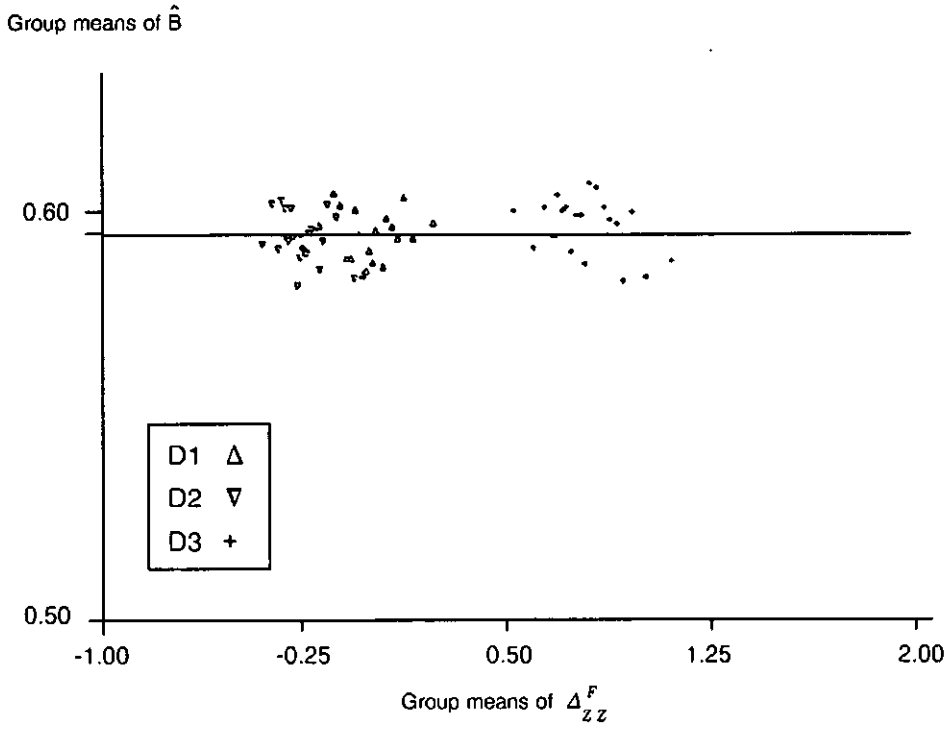


Figure 6.1 Scattergram of group means of $\hat{B}_{12,ml}$

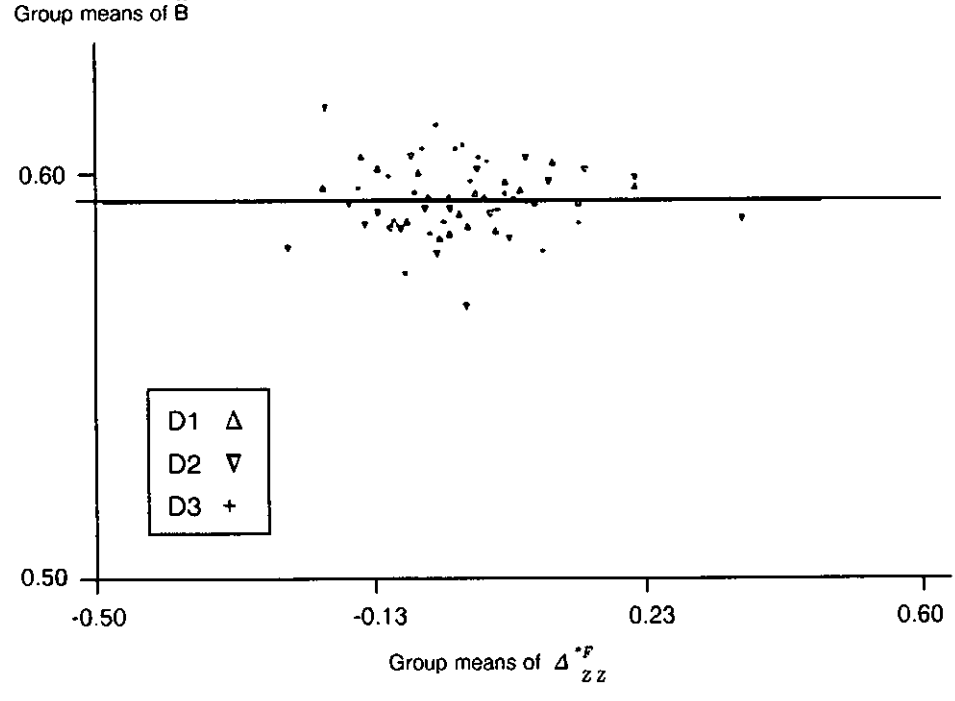


Figure 6.2 Scattergram of group means of $\hat{B}_{12,pwml}$

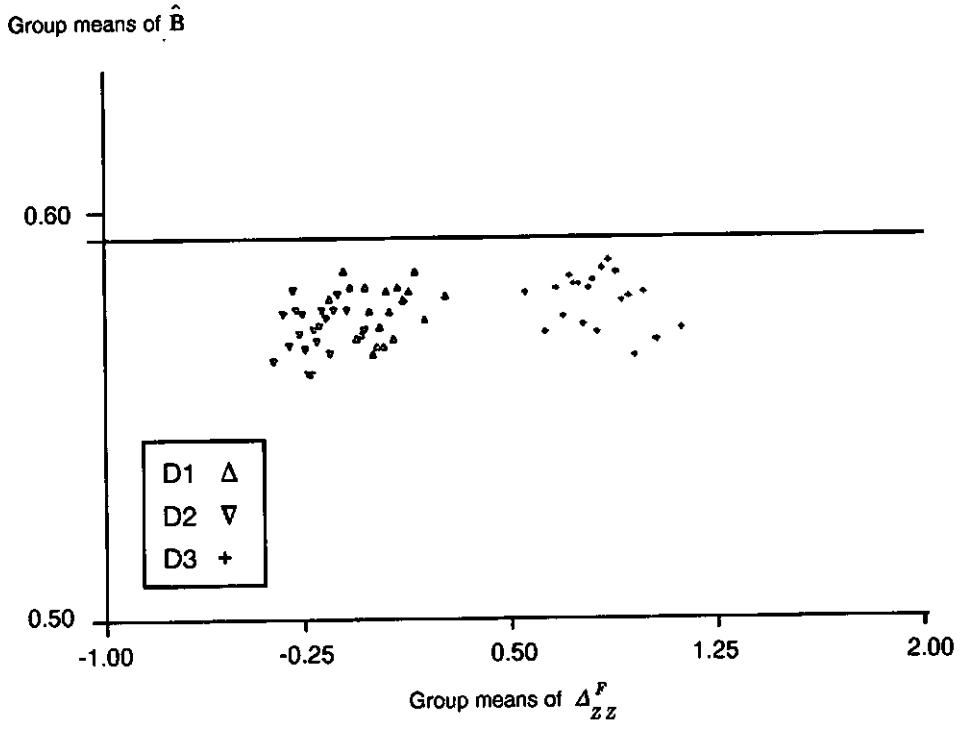


Figure 6.3 Scattergram of group means of $\hat{B}_{12,nw}$

The conditional plots are shown in Figures 6.1, 6.2 and 6.3. These plots show that there is no additional pattern to the bias beyond the absolute level of bias shown in Table 6.3. Previous studies have shown consistent patterns of bias for SRS estimators and simple p-weighted estimators, see Skinner *et al.* (1989, Chs. 7 and 8).

Simulation Study 2

Repeated sampling from a quadratic homoscedastic population

This simulation study is similar to one carried out by Holmes (1987). We generated 6,962 finite population values of (y_{1i},y_{2i},z_i) $i = 1 \ldots 6{,}962$ by first generating a value of z_i from the uniform distribution $U(0,10)$. Using this generated value of z_i the corresponding values of y_{1i} and y_{2i} are obtained from the relationships;

$$y_{2i} = m_2 + H_2 z_i + R_2 z_i^2 + \epsilon_{2i}$$

and

$$y_{1i} = m_1 + H_1 z_i + R_1 z_i^2 + \epsilon_{1i},$$

where ϵ_{2i} and ϵ_{1i} are random variables from normal distributions with mean zero and constant variance, and $R_1 \neq 0$, $R_2 \neq 0$. Following Holmes (1987) we chose the parameters in these expressions so that the regressions of y_1 and y_2 on z are monotonically increasing functions of z and the regression of y_1 on y_2 is approximately linear so that the regression coefficient B_{12} will be a meaningful parameter to estimate.

Table 6.6

Unconditional Standard Deviation of the Three Estimators of B_{12}
$N = 6{,}962$, $n = 100$ True Value $B_{12} = 0.857$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0119	0.0119	0.0171
D2	0.0923	0.0132	0.5556
D3	0.0124	0.0098	0.0104

Table 6.7

Unconditional Standard Deviation of the Three Estimators of B_{12}

Design	Standard deviations		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0877	0.0877	0.0877
D2	0.0972	0.1230	0.1150
D3	0.0785	0.1110	0.0797

Table 6.8

Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0078	0.0078	0.0080
D2	0.0180	0.0153	0.0164
D3	0.0063	0.0124	0.0065

The unconditional results of the three estimators of the regression coefficient are given in Tables 6.6, 6.7 and 6.8.

We see from the tables that the ml estimator is severely biased and very inefficient for the increasing allocation design D2, but is approximately unconditionally unbiased and efficient for the designs D1 and D3. The *pwml* estimator as expected is approximately unconditionally unbiased across all the sample designs considered. Though more biased than the *pwml* estimator, the *nw* estimator is less biased than the *ml* estimator for the unequal probability designs. We also see that the *nw* estimator is more efficient than *ml* for the design D2 and approximately equally efficient for design D3. It is also more efficient than the *pwml* estimator for the U-shaped design D3.

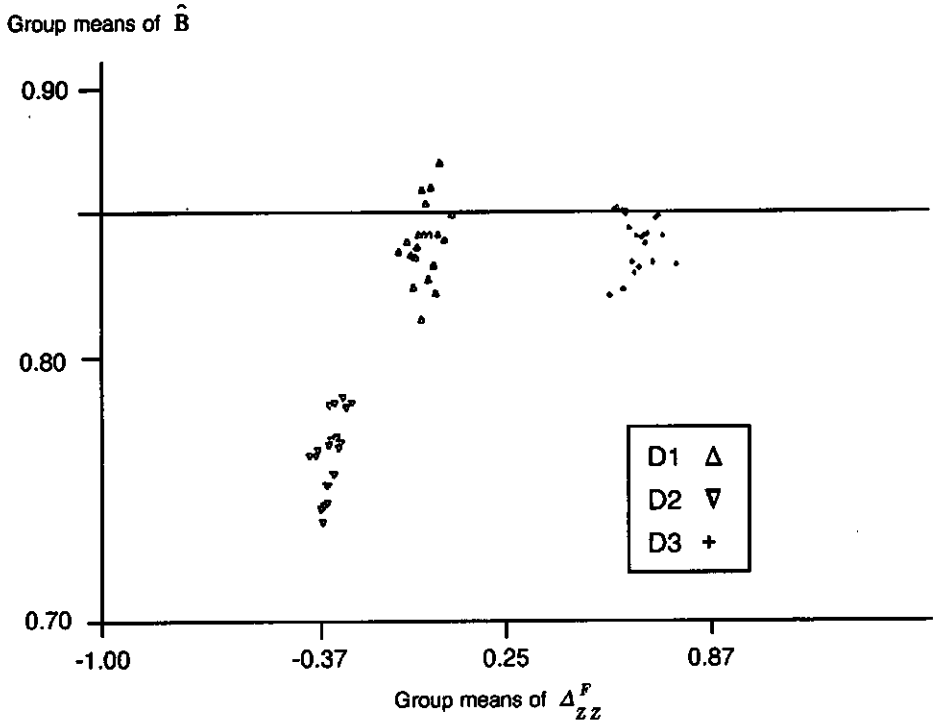


Figure 6.4 Scattergram of group means of $\hat{B}_{12,ml}$

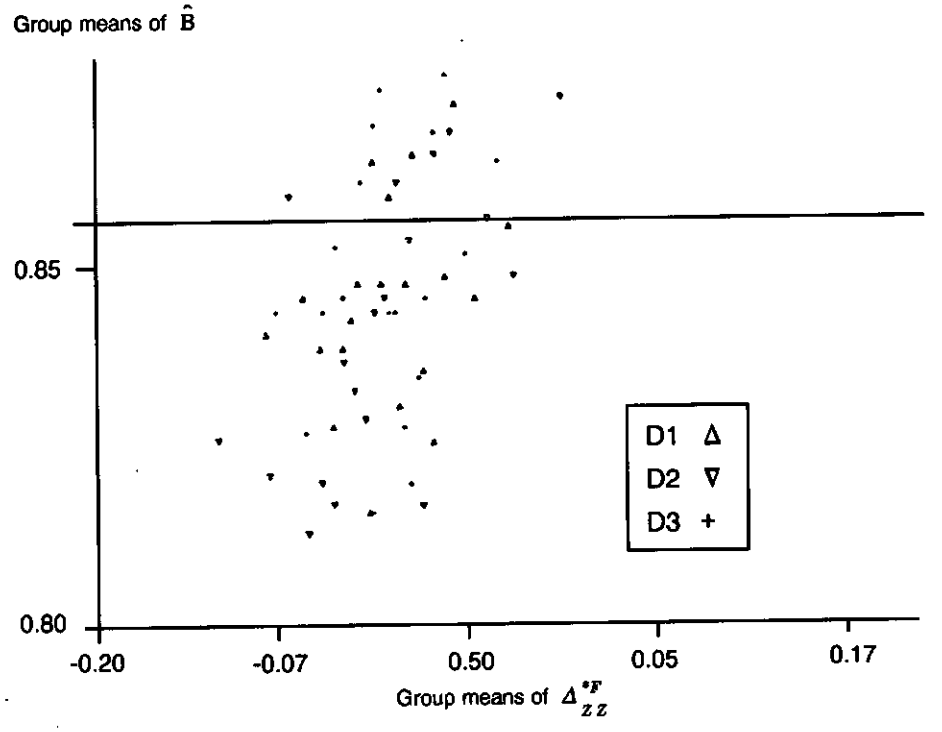


Figure 6.5 Scattergram of group means of $\hat{B}_{12,pwml}$

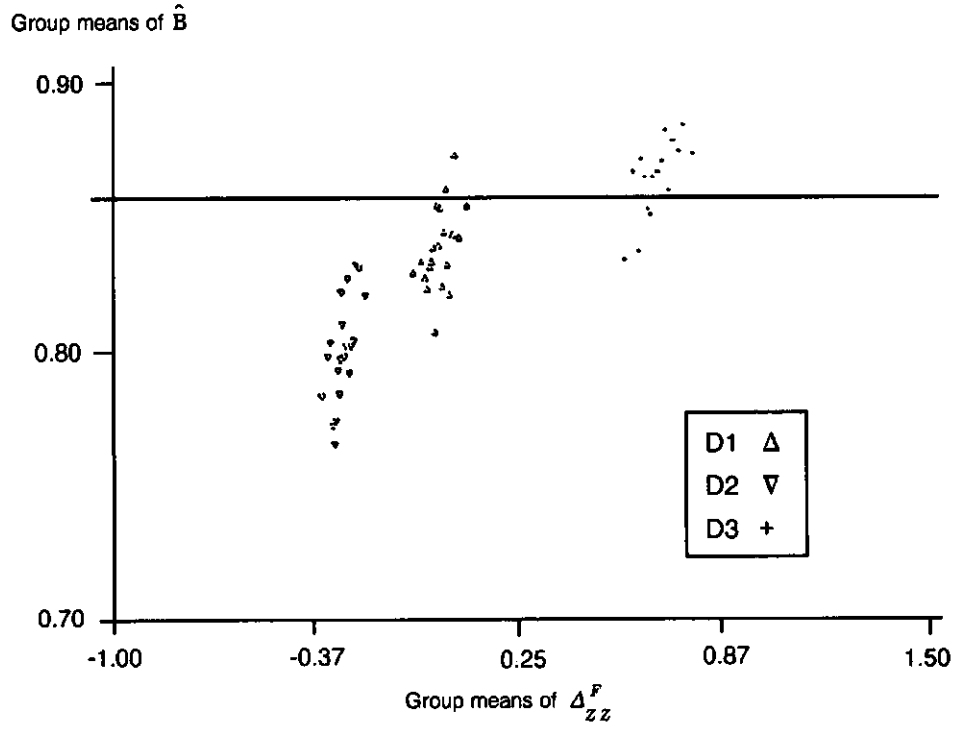


Figure 6.6 Scattergram of group means of $\hat{B}_{12,nw}$

The plots of the conditional analysis are shown in Figures 6.4, 6.5 and 6.6.

We see from Figure 6.4 that the *ml* estimator is approximately conditionally unbiased for the design D1 and D3, and has no additional conditional bias for the design D2. From Figure 6.5 we see that the *pwml* estimator has no additional conditional bias for any of the designs. We see from Figure 6.6 that the *nw* kernel estimator has only a small additional conditional bias within each of the three probability designs.

Simulation Study 3

Repeated sampling from a multivariate 'Real' population

In this simulation study we employ the 6,962 actual data points from the Family Expenditure Survey for the finite population. We consider the same variables as in section 3.1 and sample repeatedly from this population to investigate the robustness properties of the three regression estimators. We expect the real population to violate all the normality assumptions.

The unconditional results are shown in Tables 6.9, 6.10 and 6.11, and we see that the *nw* kernel estimator is the most efficient and is approximately unconditionally unbiased across all the probability designs. The *ml* estimator is less biased and more efficient than the *pwml* estimator for the unequal probability designs.

The plots of the conditional analyses are shown in Figures 6.7, 6.8 and 6.9.

We see from Figure 6.7 that the *ml* estimator is approximately conditionally unbiased for the designs D1 and D2 but has a slight conditional bias for design D3. From Figure 6.8 we see that the *pwml* estimator has no additional conditional bias for any of the designs. From Figure 6.9 we see that the *nw* kernel estimator is approximately conditionally unbiased for the three probability designs.

Table 6.9

Unconditional Absolute Biases of the Three Estimators of B_{12}
$N = 6{,}962$, $n = 100$ True Value $B_{12} = 0.595$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0245	0.0245	0.0056
D2	0.0260	0.0408	0.0060
D3	0.0128	0.0355	0.0072

Table 6.10

Unconditional Standard Deviation of the Three Estimators of B_{12}

Sample design	Standard deviation		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.111	0.111	0.111
D2	0.106	0.132	0.108
D3	0.111	0.122	0.111

Table 6.11

Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0130	0.0130	0.0121
D2	0.0120	0.0192	0.0117
D3	0.0125	0.0161	0.0123

We conclude from these simulation studies that the new estimator $\hat{\beta}_{12,nw}$ has performed well. When the assumptions of linearity and homoscedasticity are violated it appears to be robust across a variety of designs, to have good efficiency and to have reasonable conditional as well as unconditional properties. We know from previous studies that $\hat{\beta}_{12,pwml}$ performs as well as more conventional p-weighted estimators unconditionally and has far better conditional properties. The fact that in this study the new estimator $\hat{B}_{12,nw}$ apparently has better properties than the *pwml* estimator, which was chosen to represent the class of p-weighted estimators because of its performance in other simulation studies, suggests that it is an approach that could be considered in analytic studies of a small number of key parameters.

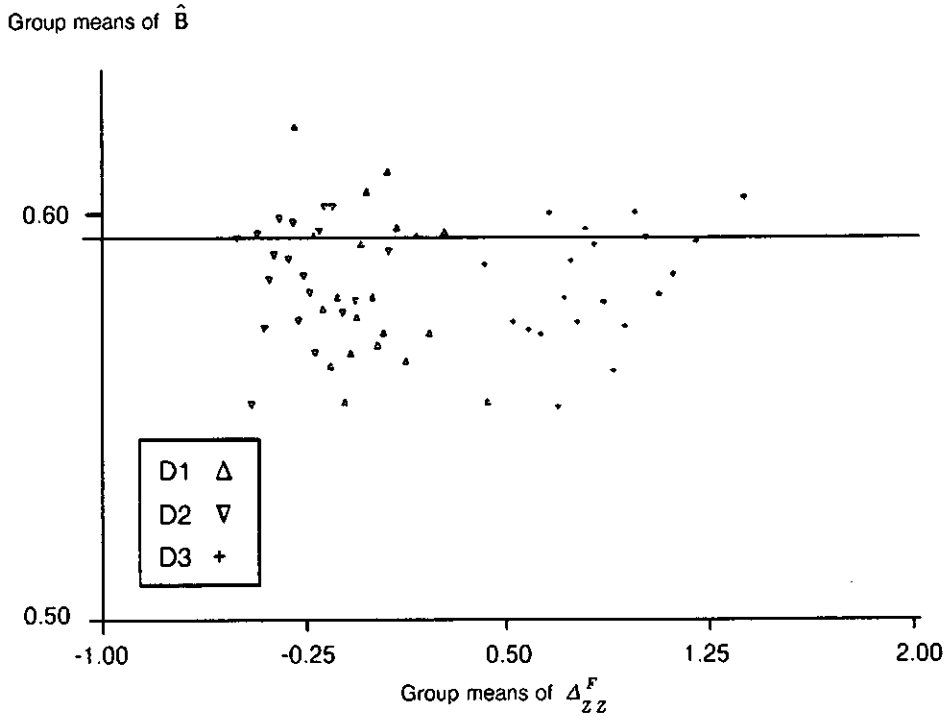


Figure 6.7 Scattergram of group means of $\hat{B}_{12,ml}$

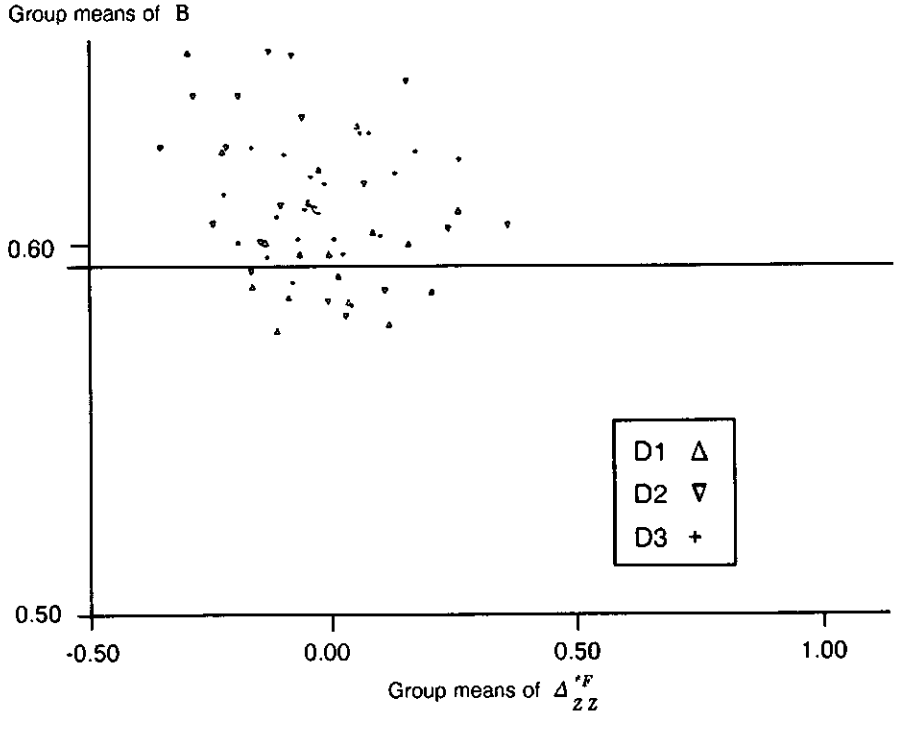


Figure 6.8 Scattergram of group means of $\hat{B}_{12,pwml}$

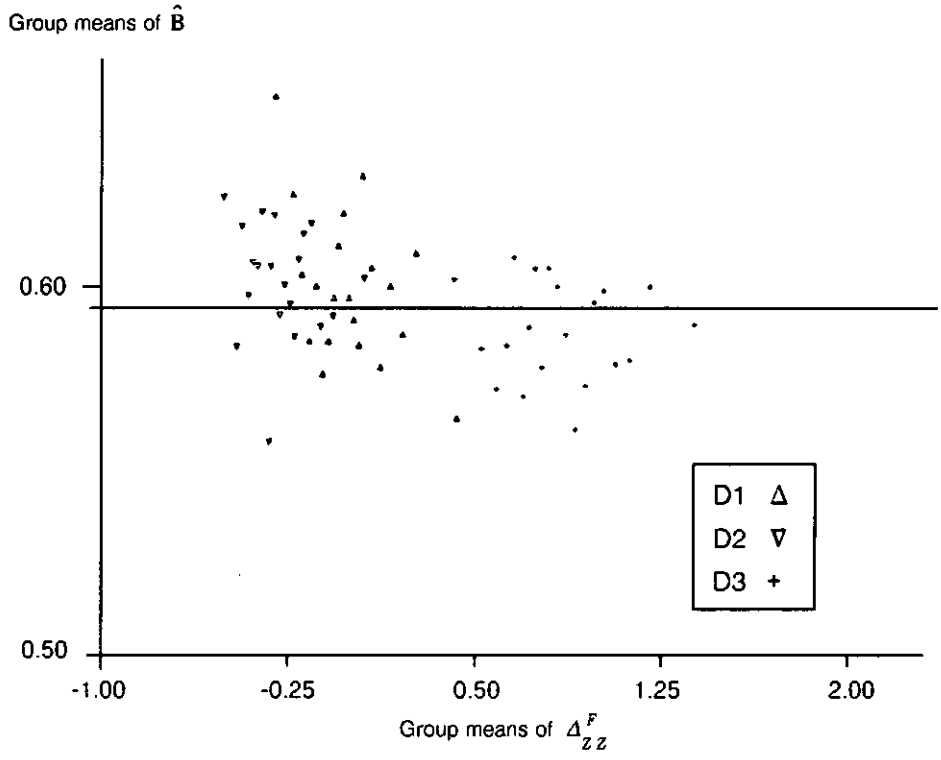


Figure 6.9 Scattergram of group means of $\hat{B}_{12,nw}$

ACKNOWLEDGEMENTS

The authors wish to thank an anonymous referee for many helpful comments which improved the presentation of the paper. E. Njenga was supported by a grant from the British Council.

REFERENCES

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

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

BREWER, K.R.W., and SÄRNDAL, C.-E. (1983). Six approaches to enumerative survey sampling. *Incomplete Data in Sample Surveys*, (Vol. 3). New York: Academic Press, 363-368.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GASSER, T., and MULLER, H.G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation*, (Eds. T. Gasser and M. Rosenblatt). New York: Springer-Verlag, 23-68.

GASSER, T., and ENGEL, J. (1990). The choice of weights in kernel regression estimation. *Biometrika*, 77, 377-381.

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

GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1211.

GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.

GODAMBE, V.P., and THOMPSON, M.E. (1977). Robust near optimal estimation in survey practice. *Bulletin of the International Statistical Institute*, 47, 129-146.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.

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

HOLT, D., SMITH, T.M.F., and WINTERS, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, Ser. A, 143, 474-487.

HOLMES, D. (1987). The effect of selection on the robustness of multivariate methods. Unpublished Doctoral thesis, University of Southampton, U.K.

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

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

KOTZ, S., and JOHNSON, N.L. (1988). *Encyclopedia of Statistical Sciences*, (Vol. 8). New York: John Wiley, 157.

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

NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability Application*, 9, 141-142.

NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, B, 42, 377-386.

NJENGA, E.G. (1990). Robust estimation of the regression coefficients in complex surveys. Unpublished Ph.D. thesis, University of Southampton.

PARZEN, E. (1962). On the estimation of the probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions Royal Society of London*, A, 200, 1-66.

PFEFFERMANN, D.J., and HOLMES, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, A, 148, 268-278.

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

ROYALL, R.M., and HERSON, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

ROYALL, R.M. and HERSON, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association*, 68, 890-893.

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

SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā*, C, 39, 1-9.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric curve fitting. *Journal of the Royal Statistical Society*, B, 47, 1-52.

SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, A, 359-372.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 209–217
Statistics Canada

Some Recent Work on Resampling Methods for Complex Surveys

J.N.K. RAO, C.F.J. WU and K. YUE[1]

ABSTRACT

Resampling methods for inference with complex survey data include the jackknife, balanced repeated replication (BRR) and the bootstrap. We review some recent work on these methods for standard error and confidence interval estimation. Some empirical results for non-smooth statistics are also given.

KEY WORDS: Balanced repeated replication; Bootstrap; Jackknife; Stratified multistage designs; Variance estimation.

1. INTRODUCTION

Standard sampling theory is largely devoted to estimation of mean square error (MSE) of unbiased or approximately unbiased estimators $\hat{Y}$ of a population total Y. An estimator of MSE, or a variance estimator, provides us with a measure of uncertainty in the estimator $\hat{Y}$. It is a common practice to assume that the estimator $\hat{Y}$ is approximately normally distributed and then use a two-sided confidence interval $\hat{Y} \pm z_{\alpha/2} s(\hat{Y})$ or a one-sided confidence interval $(\hat{Y} - z_\alpha s(\hat{Y}), \infty)$ or $(-\infty, \hat{Y} + z_\alpha s(\hat{Y}))$, where $s(\hat{Y})$ is the standard error of $\hat{Y}$ (*i.e.*, square root of estimated MSE) and z_α is the upper α-point of a $N(0, 1)$ variable. These intervals cover the true total Y with a probability of approximately $1 - \alpha$ in large samples, but the actual coverage probability could be significantly lower than $1 - \alpha$ in small samples or in highly clustered samples. For nonlinear statistics, such as ratios, regression or correlation coefficients, the well-known linearization (or Taylor expansion) method is often used (see Rao 1988 for detailed applications). Resampling methods, such as the jackknife, balanced repeated replication (BRR) and the bootstrap, are also being used, and in fact several agencies in the U.S.A and Canada have adopted the jackknife method of variance estimation for stratified multistage surveys. An advantage of the linearization method is that it is applicable to general sampling designs, but involves the derivation of a separate standard error formula, $s(\hat{\theta})$, for each nonlinear statistic, $\hat{\theta}$. On the other hand, resampling methods employ a single standard error formula for all statistics $\hat{\theta}$. However, the jackknife and the BRR methods are strictly applicable only to those stratified multistage designs in which clusters within strata are sampled with replacement or the first-stage sampling fraction is negligible. The bootstrap method of Rao and Wu (1987) works for more general designs, but it is computationally cumbersome and its properties for complex designs have not been fully investigated.

This paper provides an account of some recent work on resampling methods for complex surveys. Some empirical results on jackknife and bootstrap variance estimation for non-smooth statistics, such as the median, under stratified cluster sampling and stratified simple random sampling are also given.

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6. C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1. Kim Yue, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

2. STRATIFIED MULTISTAGE SAMPLING

Large-scale surveys often employ stratified multistage designs with large numbers of strata, L, and relatively few primary sampling units (clusters), $n_h(\geq 2)$, sampled within each stratum h. In fact, it is quite common to select $n_h = 2$ clusters within each stratum to permit maximum degree of stratification of clusters consistent with the provision of a valid variance estimator. We assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals Y_{hi}, $i = 1, \ldots, n_h$; $h = 1, \ldots, L$.

Let $w_{hik}(> 0)$ be the survey weight attached to the k-th sample element (ultimate unit) in the i-th sample cluster belonging to h-th stratum. Often, the basic weights w_{hik} are subjected to post-stratification adjustment to ensure consistency with known totals of post-stratification variables. For example, the Canadian Labour Force Survey uses a generalized regression estimator to ensure consistency. We shall, however, ignore this complication in the present paper. An estimator of the population total Y is of the form

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}, \tag{2.1}$$

where s denotes the sample of elements and y_{hik} is the value of a characteristic of interest, y, associated with the sample element $(hik)\in s$. We assume complete response on all items.

It is a common practice to sample the clusters with probabilities proportional to sizes (pps) and without replacement to increase the efficiency of the estimators compared to pps sampling with replacement and to avoid the possibility of selecting the same cluster more than once in the sample. However, at the stage of variance estimation the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement and subsampling done independently each time a cluster is selected. This approximation leads to overestimation of variance of $\hat{Y}$, but the relative bias is likely to be small if the first stage sampling fraction is small in each stratum.

Writing $\hat{Y}$ as

$$\hat{Y} = \sum_{h=1}^{L} \bar{r}_h, \tag{2.2}$$

with

$$r_{hi} = \sum_{k} (n_h w_{hik}) y_{hik}, \quad \bar{r}_h = \sum_{i} r_{hi}/n_h,$$

we note that the r_{hi} are independent and identically distributed (iid) random variables with the same mean, Y_h, and the same variance in each stratum h, under with replacement sampling of clusters. It therefore follows that an unbiased estimator of variance of $\hat{Y}$ is given by

$$s^2(\hat{Y}) = \sum_{h} s_{rh}^2/n_h, \tag{2.3}$$

with

$$(n_h - 1)s_{rh}^2 = \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2.$$

Under without-replacement sampling of clusters, $s^2(\hat{Y})$ will overestimate the true variance of $\hat{Y}$.

We are also often interested in estimating the population distribution function, $F(t)$, and the p-th quantile, $\theta = F^{-1}(p)$, $0 < p < 1$; in particular, the population median $\theta = F^{-1}(1/2)$. The survey estimator of $F(t)$ is given by

$$\hat{F}(t) = \sum_{(hik) \in s} \tilde{w}_{hik} a_{hik}, \tag{2.4}$$

where $\tilde{w}_{hik} = w_{hik} / \sum_s w_{hik}$ are the normalized weights ($\sum_s \tilde{w}_{hik} = 1$) and $a_{hik} = 1$ if $y_{hik} \leq t$, $a_{hik} = 0$ otherwise. The sample p-th quantile is obtained as

$$\hat{\theta} = \hat{F}^{-1}(p). \tag{2.5}$$

In practice, $\hat{\theta}$ is computed by first arranging the sampled values y_{hik} in an ascending order, say $\{y_{(hik)}\}$, and then cumulating the associated weights $\tilde{w}_{hik}$ until p is first crossed. The first $y_{(hik)}$ encountered after crossing p is taken as the sample p-th quantile, $\hat{\theta}$. Woodruff (1952) obtained confidence intervals for a quantile, and Rao and Wu (1987) obtained a simple variance estimator using Woodruff's interval (see also Kovar, Rao and Wu 1988, Francisco and Fuller 1991). Shao (1991) considered general L-statistics, including the sample Lorenz curve and the Gini coefficient, which are examples of smooth L-statistics, and the sample quantiles which are examples of non-smooth L-statistics.

Many nonlinear parameters of interest, such as population means, ratios, regression and correlation coefficients, can be expressed as smooth functions, $\theta = g(Y)$, of a vector of totals, $Y = (Y_1, \ldots, Y_q)'$, of suitably defined variates. An estimator of θ is given by $\hat{\theta} = g(\hat{Y})$. The linearization method may be used to estimate the variance of $g(\hat{Y})$, under any complex design (see Binder 1983 and Rao 1988).

3. RESAMPLING METHODS

Resampling methods, such as the jackknife and the bootstrap, are widely used in the iid case. Suitable modification/extensions of these methods have also been developed to handle survey data involving stratification and clustering. We now give a brief account of some recent work on three such methods: jackknife, balanced repeated replication and bootstrap, in the context of stratified multistage sampling.

3.1 Jackknife

For simplicity, assume $\hat{\theta} = g(\hat{Y})$, a smooth function of the estimated total $\hat{Y}$. Let $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$ be the estimator of θ obtained from the sample after omitting the data from the j-th sampled cluster in g-th stratum ($j = 1, \ldots, n_g$; $g = 1, \ldots, L$), where

$$\hat{Y}_{(gj)} = \sum_{\substack{(hik) \in s \\ h \neq g}} w_{hik} y_{hik} + \sum_{\substack{(gik) \in s \\ i \neq j}} \left\{ \frac{n_g}{n_g - 1} w_{gik} \right\} y_{gik}. \tag{3.1}$$

Note that $\hat{Y}_{(gj)}$ is obtained by changing the weight of (gik)-th element to $n_g w_{gik} / (n_g - 1)$, $i \neq j$, but retaining the original weights, w_{hik}, for $h \neq g$. A customary delete-1 cluster jackknife variance estimator of $\hat{\theta}$ is given by

$$s_J^2(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2. \tag{3.2}$$

Two variations of $s_J^2(\hat{\theta})$ are obtained by changing $\hat{\theta}$ in (3.2) to $\hat{\theta}_{(g.)} = \sum_j \hat{\theta}_{(gj)}/n_g$ and $\hat{\theta}_{(..)} = \sum_g \sum \hat{\theta}_{(gj)}/n$, where $n = \sum_g n_g$. In the linear case, $\hat{\theta} = \hat{Y}$, all the jackknife variance estimators reduce to the "correct" variance estimator, $s^2(\hat{Y})$, given by (2.3). Rao and Wu (1987) made a second order analysis of the resampling variance estimators when $\hat{\theta}$ is expressed as a smooth function of totals, $\hat{Y}$. Their main results on the jackknife are: (1) Different jackknife variance estimators are asymptotically equal to higher order terms, as the number of strata, L, increases. (2) In the important case of $n_h = 2$ for all h, the linearization variance estimator, $s_L^2(\hat{\theta})$, and any jackknife variance estimator are asymptotically equal to higher order terms, indicating that the choice between the two methods should depend more on operational considerations than on statistical criteria.

A drawback of the customary delete-1 jackknife method in the case of independent and identically distributed (i.i.d.) observations is that, unlike the bootstrap, it fails to provide a consistent variance estimator for non-smooth statistics, such as the median. Shao and Wu (1989), however, have shown that this deficiency of the delete-1 jackknife can be rectified by using a more general jackknife, called the delete-d jackknife, with the number of observations deleted, d, depending on a smoothness measure of the statistic. In particular, for the sample quantiles, the delete-d jackknife with d satisfying $n^{1/2}/d \to 0$ and $n - d \to \infty$ as $n \to \infty$ leads to consistent variance estimators in the case of i.i.d. observations. This result suggests that a similar effect might hold in the case of delete-1 cluster jackknife for stratified multistage sampling since all the sampled elements in a sampled cluster (gj) are deleted in computing $s_J^2(\hat{\theta})$ given by (3.2). At present we are studying this problem theoretically, but we performed a limited simulation study which suggests that the delete-1 cluster jackknife variance estimator $s_J^2(\hat{\theta})$ might perform quite well. We now report the results of the simulation study for the median, $\hat{\theta} = \hat{F}^{-1}(1/2)$.

For the simulation study, we generated stratified cluster samples $\{y_{hik}, k = 1, \ldots, M; i = 1, \ldots, n_h; h = 1, \ldots, L\}$ employing the nested error model $y_{hik} = \mu_h + a_{hi} + e_{hik}$ with $a_{hi} \overset{iid}{\sim} N(0,\sigma_{ah}^2)$ and $e_{hik} \overset{iid}{\sim} N(0,\sigma_{eh}^2)$, where the cluster size, M is assumed to be equal for all clusters (hi), and the intra-cluster correlations, $\sigma_{ah}^2/(\sigma_{ah}^2 + \sigma_{eh}^2) = \rho_h$, are assumed to be equal for all strata h (*i.e.*, $\rho_h = \rho$). The normalized survey weights are given by $\tilde{w}_{hik}$ with $w_{hik} = W_h/(n_h M)$ and W_h denotes the relative size of stratum h. The number of strata $L(= 32)$, strata means, μ_h, variances $\sigma_h^2 = \sigma_{ah}^2 + \sigma_{eh}^2$ and sizes W_h were chosen to correspond to real populations encountered in the US National Assessment of Educational Progress Study (Hansen and Tepping 1985). We generated 1,000 independent stratified cluster samples with $n_h = 2$ for each selected combination (ρ,M) and then computed the bias and relative bias of the jackknife variance estimator, $s_J^2(\hat{\theta})$, for the median: $\text{Bias}[s_J^2(\hat{\theta})] = \sum_t s_{Jt}^2(\hat{\theta})/1{,}000 - \text{MSE}(\hat{\theta})$, where $s_{Jt}^2(\hat{\theta})$ is the value of $s_J^2(\hat{\theta})$ for the t-th simulated sample ($t = 1, \ldots, 1{,}000$) and Rel. $\text{Bias}[s_J^2(\hat{\theta})] = \text{Bias}[s_J^2(\hat{\theta})]/\text{MSE}(\hat{\theta})$. We calculated $\text{MSE}(\hat{\theta})$ from an independent set of 10,000 stratified cluster samples for each (ρ,M): $\text{MSE}(\hat{\theta}) = \sum_t (\hat{\theta}_t - \hat{\theta}_.)^2/10{,}000$, where $\hat{\theta}_t$ is the value of $\hat{\theta}$ for the t-th simulated sample, $\hat{\theta}_. = \sum \hat{\theta}_t/10{,}000$ and $t = 1, \ldots, 10{,}000$.

Table 1 reports the simulated values of bias and relative bias (in brackets) of the jackknife variance estimator for selected combinations of ρ and M. First, we note that for the special case of stratified simple random sampling ($\rho = 0$, $M = 1$), the relative bias is very large (116%) thus confirming the inconsistency of $s_J^2(\hat{\theta})$ in this case. Second, we observe that both the bias and relative bias decrease as M increases for a given ρ. Moreover, for a given cluster

Table 1

Bias and % Relative Bias (in Brackets) of Jackknife Variance Estimator for the Median Under Stratified Cluster Sampling ($n_h = 2$, $L = 32$) and Selected Values of Equal Intra-Cluster Correlation, ρ, and Equal Cluster Size, M

ρ	M				
	1	10	20	30	50
0	7.5(116)	.28(41)	.09(29)	.04(15)	.01(15)
0.05	-	.22(27)	.09(18)	.05(12)	.03 (8)
0.10	-	.28(28)	.10(14)	.06 (9)	.02 (3)
0.20	-	.31(22)	.11(10)	.08 (8)	.03 (3)
0.30	-	.32(18)	.11 (7)	.07 (5)	.01 (1)
0.50	-	.44(17)	.15 (6)	.11 (5)	.04 (2)

size M, the bias generally increases with ρ, but the relative bias in fact decreases because MSE $(\hat{\theta})$ is increasing faster than the bias as ρ increases. It is indeed gratifying that the relative bias is no more than 10% for $M \geq 30$ and $\rho \geq 0.10$ or $M \geq 20$ and $\rho \geq 0.20$.

3.2 Balanced Repeated Replication (BRR)

Balanced repeated replication (BRR) was proposed by McCarthy (1969) for the important special case of $n_h = 2$ clusters per stratum. A set of R balanced half-samples (replications) is formed by deleting one cluster from the sample in each stratum. This set may be defined by a $R \times L$ design matrix (δ_h^r), $1 \leq r \leq R$, $1 \leq h \leq L$ with $\delta_h^r = +1$ or -1 according as whether the first or second sample cluster in the h-th stratum is in the r-th half-sample, and $\sum_r \delta_h^r \delta_{h'}^r = 0$ for all $h \neq h'$, *i.e.* the columns of the matrix are orthogonal. A minimal set of R balanced half-samples may be constructed from Hadamard matrices ($L + 1 \leq R \leq L + 4$) by choosing any L columns, excluding the column of $+1$'s.

Let $\hat{\theta}^{(r)}$ be the estimator of θ obtained from the r-th half-sample. Note that $\hat{\theta}^{(r)}$ is obtained from $\hat{\theta}$ by changing the weight of (hik)-th element to $2w_{hik}$ or 0 according as the (hi)-th cluster is selected or not selected in the half-sample. A BRR variance estimator of $\hat{\theta}$ is given by

$$s^2_{\mathrm{BRR}}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.3}$$

Several variations of $s^2_{\mathrm{BRR}}(\hat{\theta})$ are also available; for example, $\hat{\theta}$ may be changed to $\hat{\theta}(\cdot) = \sum_r \hat{\theta}^{(r)}/R$. In the linear case, $\hat{\theta} = \hat{Y}$, all the BRR variance estimators reduce to the "correct" variance estimator, $s^2(\hat{Y})$, as in the case of the jackknife.

Krewski and Rao (1981) established the consistency of $s^2_J(\hat{\theta})$ and $s^2_{\mathrm{BRR}}(\hat{\theta})$ for smooth statistics $\hat{\theta} = g(\hat{Y})$, as L increases. Rao and Wu (1985) made a second order analysis and showed that $s^2_{\mathrm{BRR}}(\hat{\theta})$ and $s^2_L(\hat{\theta})$ are not asymptotically equivalent to second order terms, unlike $s^2_J(\hat{\theta})$ and $s^2_L(\hat{\theta})$. Shao and Wu (1992) established the consistency of $s^2_{\mathrm{BRR}}(\hat{\theta})$ for the quantiles, $\hat{\theta} = \hat{F}^{-1}(p)$.

The BRR method has been extended to the case of $n_h = p > 2$ clusters per stratum for p prime or power of prime (Gurney and Jewett 1975), but the number of replications, R, needed is much larger than in the case of $n_h = 2$. In many survey designs n_h's are not equal. To accommodate the general case of unequal n_h, Gupta and Nigam (1987) and Wu (1991)

advocated the use of mixed-level orthogonal arrays of strength two for drawing balanced replicates, where n_h is the number of symbols in the h-th column of the array. Orthogonality of the array guarantees that the replicates drawn are balanced. Unlike the case of equal n_h, the adjustment of survey weights is more complicated. A correct method was given by Wu (1991). From his formula (6), two separate adjustments should be applied to the sampled and unsampled units in each replicate. Simple algebra on Wu's equation (6) shows that w_{hik} is changed to $w'_{hik} = [1 + (n_h - 1)^{1/2}]w_{hik}$ or $w''_{hik} = [1 - (n_h - 1)^{1/2}]w_{hik}$ according as the (hik)-th element is selected or not selected in the replicate. (Note that $w'_{hik} = 2$ and $w''_{hik} = 0$ for $n_h = 2$). The remaining calculation of $\hat{\theta}^{(r)}$ and $s^2_{\mathrm{BRR}}(\hat{\theta})$ are the same as in (3.3). Furthermore, these modified survey weights can be applied to $\hat{\theta} = \hat{F}^{-1}(p)$ and more general $\hat{\theta} = T(\hat{F})$, where T is a functional of $\hat{F}$. All we need to do is to change w_{hik} in (2.4) to w'_{hik} or w''_{hik} according as the (hik)-th element is selected or not selected in the r-th replicate to get $\hat{F}^{(r)}$ of F for the r-th replicate, and $\hat{\theta}^{(r)} = T(\hat{F}^{(r)})$. The calculation of the BRR variance estimator is the same as in (3.3).

There are two problems with the use of mixed orthogonal arrays. First, the array size can be large for general n_h. Second, orthogonal arrays do not exist for any combination of n_h's. A practical solution is to group the n_h sample psu's in stratum h into two to four groups of psu's and then apply the method to the groups by treating the groups as units in the BRR method. This extension is called the grouped BRR method. As shown by Wu (1991), its efficiency loss can be relatively small, compared to the full BRR, if the groupings are done judiciously. For example, more groups are needed if n_h is large and the units within the stratum are more heterogeneous. For $n_h = 2$, 3 or 4, many mixed orthogonal arrays have been constructed (see, for example, Dey 1985 and Wang and Wu 1991). If n_h can only take 2 or 4, saturated orthogonal arrays for any combination can be easily constructed as in Wu (1989). That is, the number of replications can be as small as possible. It is therefore possible to compile a large collection of mixed orthogonal arrays for practical use if n_h is restricted to 2, 3 or 4.

The BRR method and extensions considered thus far only take one unit (psu) per stratum for each replicate. If n_h is large, say more than 3, Sitter (1992) proposed the use of orthogonal multi-arrays to allow the number of resampled units per stratum to be greater than one. It may require fewer replicates and it can cover cases where orthogonal arrays of strength two are not available; for example, $n_h = 6$.

3.3 Bootstrap

The bootstrap method for the iid case has been extensively studied (Efron 1982). Rao and Wu (1987) provided an extension to stratified multistage designs, but covering only smooth statistics $\hat{\theta} = g(\hat{Y})$. They required that, in order to have valid variance estimation in the case of small n_h, some scale adjustment, similar to those in Section 3.2, is necessary. What they did not realize is that the scale adjustment should be made on the survey weights w_{hik} rather on the y_{hik} values directly, which is what they proposed. As a result, their method cannot be extended to cover the quantile $\theta = F^{-1}(p)$. We now present a general method that covers smooth as well as non-smooth statistics for arbitrary sizes, n_h. It works as follows: (i) Draw a simple random sample of m_h clusters with replacement from the n_h sample clusters, independently for each h. Let m^*_{hi} be the number of times (hi)-th sample cluster is selected $(\sum_i m^*_{hi} = m_h)$. Define the bootstrap weights

$$w^*_{hik} = \left[\{1 - (m_h/(n_h - 1))^{1/2}\} + (m_h/(n_h - 1))^{1/2}(n_h/m_h)m^*_{hi}\right]w_{hik}. \qquad (3.4)$$

If the (hi)-th cluster is not selected in the bootstrap sample, $m_{hi}^* = 0$ and the second term of (3.4) vanishes. If m_h is chosen to be less than or equal to $n_h - 1$, then the bootstrap weights w_{hik}^* are all positive if $w_{hik} > 0$ for all $(hik) \in s$ Calculate θ^*, the bootstrap estimator of θ, using the weights w_{hik}^* in the formula for $\hat{\theta}$. The bootstrap median, for example, is calculated as before using the normalized bootstrap weights $\tilde{w}_{hik}^* = w_{hik}^* / \sum_s w_{hik}^*$, provided all $w_{hik}^* > 0$. (ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimates $\theta_{(1)}^*, \ldots, \theta_{(B)}^*$.

The bootstrap variance estimator $s_{\text{BOOT}}^2(\hat{\theta}) = E_*(\theta^* - E_*\theta^*)^2$, is approximated by

$$\tilde{s}_{\text{BOOT}}^2(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} [\theta_{(b)}^* - \hat{\theta}]^2. \tag{3.5}$$

A variation of (3.5) is obtained by changing $\hat{\theta}$ to $\theta_{(\cdot)}^* = \sum_b \theta_{(b)}^*/B$. In the linear case, $s_{\text{BOOT}}^2(\hat{\theta})$ reduces to the "correct" variance estimator $s^2(\hat{Y})$.

Rao and Wu (1987) obtained bootstrap-t confidence intervals for smooth functions, $\theta = g(Y)$, by approximating the distribution of $t = (\hat{\theta} - \theta)/s_J(\hat{\theta})$ by its bootstrap counterpart $t^* = (\theta^* - \hat{\theta})/s_J(\theta^*)$, where $s_J^2(\theta^*)$ is obtained from(3.2) with w_{hik} changed to w_{hik}^*. A two-sided $(1 - \alpha)$-level confidence interval for θ is then given by $\{\hat{\theta} - t_U^* s_J(\hat{\theta}), \hat{\theta} - t_L^* s_J(\hat{\theta})\}$, where t_L^* and t_U^* are the lower and upper $\alpha/2$-points of t^* obtained from the bootstrap histogram of $t_{(1)}^*, \ldots, t_{(B)}^*$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Empirical work by Kovar, Rao and Wu (1988) for smooth functions indicates that the bootstrap-t interval with $m_h = n_h - 1$ tracks the error rates in both the lower and upper tails better than the jackknife interval $\{\hat{\theta} - z_{\alpha/2} s_J(\hat{\theta}), \hat{\theta} + z_{\alpha/2} s_J(\hat{\theta})\}$, but the total error rate is not distinguishable from the latter, *i.e.*, for two-sided intervals, they exhibit similar performance in terms of actual coverage probability. If a variance stabilizing transformation can be found, such as the $\tanh^{-1}$ transformation on the estimated correlation coefficient, then the problem of uneven error rates in the two tails for the jackknife interval seems to be corrected. This suggests that the jackknife interval, or any other normal-theory interval, based on such transformations can be useful when the transformations are known, while the bootstrap provides an alternative when such transformations do not exist or are unknown.

We now present the results of a limited simulation study on the performance of the proposed bootstrap method in the case of the median. Employing the Hansen-Tepping basic population 1 with $L = 32$ strata (see Kovar *et al.* 1988, Sections 3 and 6 for details), we generated 500 independent stratified simple random samples with $n_h = 5$ and then computed the relative bias and coefficient of variation (relative stability) of the Woodruff-based variance estimator with $\alpha = 0.1$ (see Kovar *et al.* 1988, eq. (2.8)), the BRR variance estimator (3.3) and the bootstrap variance estimator (3.5) and its variation obtained by changing $\hat{\theta}$ to $\theta_{(\cdot)}^*$. We used $m_h = n_h - 1$ and $n_h - 3$ and $B = 500$ bootstrap replicates for each sample, while the BRR replicates were obtained from an orthogonal array with 250 runs. The true MSE of $\hat{\theta}$ was approximated by selecting 10,000 independent stratified random samples. We also calculated the error rates in each tail (nominal rate of 5% in each tail) and standardized lengths of the normality-based confidence interval using the BRR variance estimator, the Woodruff interval and the bootstrap interval obtained from the percentile method using the bootstrap histogram of $\theta_{(1)}^*, \ldots, \theta_{(B)}^*$ for each sample.

Table 2 reports the simulated values of the relative bias, coefficient of variation, lower (L) and upper (U) error rates, and standardized lengths. First, we note that the bootstrap variance estimator (3.5) has a larger relative bias and a slightly larger coefficient of variation (CV) than

Table 2

% Relative Bias and % CV of Variance Estimator and Error Rates and Standardized Lengths of Confidence Intervals (Nominal Level of 5% in Each Tail) for the Median Under Stratified Simple Random Sampling $L = 32$, $n_h = 5$)

Method	% Rel. Bias	% CV	Error Rate		St. Length
			L	U	
Woodruff	4.2	47	4.2	5.6	0.997
BRR	3.1	31	5.0	5.0	1.004
Bootstrap*:					
$m_h = 4$	12.6 (7.5)	52 (48)	5.0	5.2	0.987
$m_h = 2$	13.0 (7.8)	54 (49)	5.0	4.8	0.988

* Results for the variation of the bootstrap variance estimator are given in the brackets.

its variation obtained by changing $\hat{\theta}$ to $\theta^*_{(.)}$: Relative bias of 12.6% *vs.* 7.5% and CV of 52% *vs.* 48% for $m_h = n_h - 1 = 4$. On the other hand, the BRR variance estimator has the smallest relative bias (3.1%) and the smallest CV (31%), while the Woodruff-based variance estimator has a smaller relative bias (4.2%) and a comparable CV (47%). Secondly, the lower and upper error rates are close to the nominal level (5%) for the bootstrap and the BRR intervals, while the error rates are slightly uneven for the Woodruff interval ($L = 4.2\%$ and $U = 5.6\%$). Finally, we note that the standardized lengths are roughly equal for all the methods. Overall, the bootstrap variance estimator and the bootstrap intervals based on the percentile method did not exhibit better performance relative to either the BRR variance estimator and the associated normality-based interval or the Woodruff-based variance estimator and the Woodruff interval.

ACKNOWLEDGEMENT

J.N.K. Rao's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

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

DEY, A. (1985). *Orthogonal Fractional Factorial Designs*. New Delhi: Wiley Eastern.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

GURNEY, M., and JEWETT, R.S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, 70, 819-821.

HANSEN, M., and TEPPING, B.J. (1985). Estimation for variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25-45.

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

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

RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics, Vol. 6*, (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: Elsevier Science, 427-447.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., and WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data. *Bulletin of the International Statistical Institute*.

SHAO, J. (1991). *L*-statistics in complex survey problems. Technical Report, University of Ottawa, Ottawa.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SHAO, J., and WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics*, 20 (to appear).

SITTER, R.R. (1992). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, (to appear).

WANG, J.C., and WU, C.F.J. (1991). An approach to the construction of asymmetrical orthogonal arrays. *Journal of the American Statistical Association*, 86, 450-456.

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

WU, C.F.J. (1989). Construction of 2^m4^n designs via a grouping scheme. *Annals of Statistics*, 17, 1880-1885.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 219–224
Statistics Canada

An Estimating Function Approach to Finite Population Estimation

HAROLD J. MANTEL[1]

ABSTRACT

Godambe and Thompson (1986) define and develop simultaneous optimal estimation of superpopulation and finite population parameters based on a superpopulation model and a survey sampling design. Their theory defines the finite population parameter, θ_N, as the solution of the optimal estimating equation for the superpopulation parameter θ; however, some other finite population parameter, ϕ, may be of interest. We propose to extend the superpopulation model in such a way that the parameter of interest, ϕ, is a known function of θ_N, say $\phi = f(\theta_N)$. Then ϕ is optimally estimated by $f(\theta_s)$, where θ_s is the optimal estimator of θ_N, as given by Godambe and Thompson (1986), based on the sample s and the sampling design.

KEY WORDS: Estimating functions; Generalized linear estimator; Finite population parameter.

1. ESTIMATION OF A MEAN

The problem discussed in this paper is the estimation of a finite population parameter such as the mean based on a sample survey. There is also a hypothesized superpopulation regression model relating the variable of interest to some known covariables. The objective is an estimation procedure which has good properties with respect to both the sampling design and the hypothesized model. The approach here is based on the work of Godambe and Thompson (1986).

We suppose that we have a finite population of labeled individuals $\boldsymbol{P} = \{i: i = 1, \ldots, N\}$. With each individual i is associated an unknown variable y_i and a vector of covariables, $\boldsymbol{x}_i$. The vector $\boldsymbol{x}_i$ may be known for all $i \in \boldsymbol{P}$ or only for i in the sample and the population mean $\bar{\boldsymbol{x}}_N$ would be known. Letting E_m denote expectation with respect to the superpopulation model, the model assumptions are:

(i) y_i and y_j are independent for $i \neq j$

(ii) $E_m(y_i) = \boldsymbol{x}_i^T\beta$ for some unknown real vector β

(iii) $E_m(y_i - \boldsymbol{x}_i^T\beta)^2 = \sigma^2 v_i$, $i = 1, \ldots, N$, for known v_i and some unknown σ^2.

Following Godambe and Thompson (1986) we define a finite population parameter $\hat{\beta}_N$ as the solution of the linearly optimal estimating equation

$$g^* = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i/v_i = 0, \tag{1}$$

that is,

$$\hat{\beta}_N = (\boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{y}_N, \tag{2}$$

[1] H.J. Mantel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

where $y_N^T = (y_1, \ldots, y_N)$, V_N is a diagonal matrix with entries $v_1, \ldots, v_N$, and X_N is a matrix with N rows, the ith row being x_i^T.

Now $\hat{\beta}_N$ is unknown. Godambe and Thompson (1986) defined and developed simultaneous optimal estimation of β and $\hat{\beta}_N$ based on the model and the sampling design. We will denote the data from a sample survey by $\chi_s = \{(i, y_i), i \epsilon s\}$.

For simultaneous estimation of β and $\hat{\beta}_N$ we consider estimating functions $\boldsymbol{h}(\chi_s, \beta)$ such that $E_p(\boldsymbol{h}) = \boldsymbol{g}^*$ in (1), where E_p denotes expectation with respect to the sampling design. A function $\boldsymbol{h}^*$ in this class is called optimal if for all other $\boldsymbol{h}$ in the class $E_m E_p\{\boldsymbol{h}\boldsymbol{h}^T\} - E_m E_p\{\boldsymbol{h}^*\boldsymbol{h}^{*T}\}$ is non-negative definite. Theorem 1 of Godambe and Thompson (1986) shows that the optimal function $\boldsymbol{h}^*$ is given by

$$\boldsymbol{h}^*(\chi_s, \beta) = \sum_{i \epsilon s} (y_i - x_i^T \beta) x_i / \pi_i v_i, \tag{3}$$

where π_i is the probability under the sampling design that individual i is included in the sample s. We will denote the root of this function by $\hat{\beta}_s$, that is,

$$\hat{\beta}_s = (X_s^T \Pi_s^{-1} V_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} V_s^{-1} y_s, \tag{4}$$

where y_s is the vector of y_is for $i \epsilon s$, Π_s and V_s are diagonal matrices with entries π_i and v_i respectively, $i \epsilon s$, and X_s is the matrix with rows x_i^T, $i \epsilon s$.

So far we have discussed only estimation of β or $\hat{\beta}_N$. Our problem was to estimate $\bar{y}_N$, the population mean of the y_is. One possibility is to use a generalized regression estimator,

$$\bar{y}_{\text{GREG}} = \bar{x}_N^T \hat{\beta}_s + \mathbf{1}_s^T \Pi_s^{-1} (y_s - X_s \hat{\beta}_s)/N, \tag{5}$$

where $\mathbf{1}_s$ is a vector of 1's whose length is the size of the sample s. This estimator is discussed, for example, by Särndal, Swensson and Wretman (1992). The first part of the estimator gives good model properties while the second part gives good design properties. However, the model and design justifications of $\bar{y}_{\text{GREG}}$ in (5) do not depend on the particular form of $\hat{\beta}_s$, and there is no immediately apparent reason why $\hat{\beta}_s$ in (5) could not be replaced by a purely model based estimator of β. The design optimality of $\hat{\beta}_s$ is apparently irrelevant.

The estimator we will propose here more closely integrates the hypothesized model with the finite population parameter $\bar{y}_N$. Since $\hat{\beta}_N$ in (2) is optimally estimated by $\hat{\beta}_s$ in (4), functions of $\hat{\beta}_N$ are optimally estimated by the same function of $\hat{\beta}_s$. If $\bar{y}_N = \boldsymbol{u}^T \hat{\beta}_N$ for some vector $\boldsymbol{u}$ then we would estimate $\bar{y}_N$ by $\boldsymbol{u}^T \hat{\beta}_s$. Such a $\boldsymbol{u}$ exists if and only if $V_N \mathbf{1}_N$ is in the column space of X_N, in which case, with $V_N \mathbf{1}_N = X_N \boldsymbol{a}$, we may take $\boldsymbol{u} = X_N^T V_N^{-1} X_N \boldsymbol{a}/N = \bar{x}_N$. The idea then is that if $V_N \mathbf{1}_N$ is not in the column space of X_N, we will add it. In doing so we lose something of model efficiency, though the augmented model remains valid in light of the original model. We relax model efficiency to gain some sort of finite population relevance. As an interesting special case we note that when the model variances do not depend on i our approach leads to including an arbitrary constant term in the regression model.

The approach taken here seems quite similar to that of Little (1983) who suggests model based estimation restricting attention to models that yield asymptotically design consistent estimators. Alternatively, Isaki and Fuller (1982) suggest restricting to designs for which the model based estimator is asymptotically design consistent.

2. COMPARISON TO THE GENERALIZED REGRESSION ESTIMATOR

Let W_N be the design matrix for the augmented model, that is

$$W_N = (V_N \mathbf{1}_N, X_N). \tag{6}$$

For the discussion of this section we assume that $V_N \mathbf{1}_N$ is not in the column space of X_N. Similarly, let W_s be the augmented form of X_s, and γ, $\hat{\gamma}_N$, and $\hat{\gamma}_s$ be the augmented forms of β, $\hat{\beta}_N$, and $\hat{\beta}_s$ respectively.

For convenience, we will refer to our estimator of the population mean as the augmented regression estimator,

$$\bar{y}_{\text{AREG}} = \bar{w}_N^T \hat{\gamma}_s. \tag{7}$$

We first show that $\bar{y}_{\text{AREG}}$ is also a type of generalized difference estimator. From (6), if $\boldsymbol{u}$ is a vector of appropriate length with the first entry equal to one and the rest zeros then $W_N \boldsymbol{u} = V_N \mathbf{1}_N$ and $W_s \boldsymbol{u} = V_s \mathbf{1}_s$. Then

$$\mathbf{1}_s^T \Pi_s^{-1} W_s \hat{\gamma}_s = \boldsymbol{u}^T W_s^T V_s^{-1} \Pi_s^{-1} W_s \hat{\gamma}_s = \boldsymbol{u}^T W_s^T V_s^{-1} \Pi_s^{-1} y_s = \mathbf{1}_s^T \Pi_s^{-1} y_s$$

and it follows that the second part of the generalized regression estimator in (5) with $\hat{\beta}_s$ replaced by $\hat{\gamma}_s$ is equal to 0.

Secondly, let us compare $\bar{y}_{\text{AREG}}$ in (7) to $\bar{y}_{\text{GREG}}$ in (5). A few tedious calculations give us that

$$\bar{y}_{\text{AREG}} = \bar{x}_N \hat{\beta}_s + (c_1/c_2) \mathbf{1}_s^T \Pi_s^{-1} (y_s - X_s \hat{\beta}_s)/N,$$

where

$$c_1 = \mathbf{1}_N^T (V_N \mathbf{1}_N - X_N (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s)$$

and

$$c_2 = \mathbf{1}_s^T \Pi_s^{-1} (V_s \mathbf{1}_s - X_s (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s).$$

Written in this way $\bar{y}_{\text{AREG}}$ appears very similar to $\bar{y}_{\text{GREG}}$ except for an adjusted weight for the second part. It does not seem possible to give an heuristic explanation of the weight (c_1/c_2). However, we note that c_1 is just the population sum of the residuals from a weighted regression of the v_i's onto the x_i's based on the sample s, and c_2 looks something like a Horvitz-Thompson estimator of c_1, except that the residuals also depend on the sample s. For large samples from large populations we would expect (c_1/c_2) to be close to 1.

In comparing $\bar{y}_{\text{AREG}}$ with $\bar{y}_{\text{GREG}}$ we may say that $\bar{y}_{\text{AREG}}$ is more design based and $\bar{y}_{\text{GREG}}$ is more model based. Of course, $\bar{y}_{\text{GREG}}$ is design consistent, but $\bar{y}_{\text{AREG}}$ has also a finite sample design justification in that $\hat{\gamma}_s$ is the solution of an estimating equation which is design unbiased for the parameter defining equation of $\hat{\beta}_N$. Parameter defining equations are discussed by Godambe and Thompson (1984, 1986).

3. VARIANCE ESTIMATION AND CONFIDENCE INTERVALS

A method of confidence interval construction which would be consistent with the general philosophy of estimating functions would be to construct an asymptotically multivariate normal pivotal based on h^* and an estimator of its variance. Approximate confidence regions for $\hat{\gamma}_N$ would then correspond to probability regions of the estimated multivariate normal distribution of this approximate pivotal. However, we are not interested in $\hat{\gamma}_N$ but in a non-injective function of $\hat{\gamma}_N$. We will adopt the more straight-forward approach of estimating the variance of $\bar{y}_{\text{AREG}}$ directly.

Särndal, Swensson, and Wretman (1989) have investigated variance estimation for $\bar{y}_{\text{GREG}}$ in (5) for the case that the second part is zero. As we have seen in section 2, our estimator $\bar{y}_{\text{AREG}}$ is precisely of that type. Their variance estimator may be written as

$$\hat{V}_g = \sum_{i\epsilon s} \sum_{j\epsilon s} \tilde{\Delta}_{ij} g_{is} \tilde{e}_{is} g_{js} \tilde{e}_{js}, \tag{8}$$

where $\tilde{\Delta}_{ij} = (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}$, π_{ij} is the design probability that both individuals i and j are included in the sample s, g_{is} is the ith element of the row vector $\tilde{w}_N^T(W_s^T V_s^{-1} \Pi_s^{-1} W_s)^{-1} W_s^T V_s^{-1}$, and $\tilde{e}_{is} = (y_i - x_i^T \hat{\gamma}_s)/\pi_i$. See Särndal, Swensson and Wretman (1989) for a detailed discussion of the model and design properties of $\hat{V}_g$ in (8). Note that $\bar{y}_{\text{AREG}}$ in (7) may be written as $\bar{y}_{\text{AREG}} = \sum_{i\epsilon s} g_{is} y_i/\pi_i$ and

$$\bar{y}_{\text{AREG}} - \bar{y}_N = \sum_{i\epsilon s} g_{is} \tilde{e}_{iN} = \tilde{w}_N^T(\hat{\gamma}_s - \hat{\gamma}_N),$$

where $\tilde{e}_{iN} = (y_i - w_i^T \hat{\gamma}_N)/\pi_i$. Now, with $V_N \mathbf{1}_N = W_N a$, we have $\tilde{w}_N^T = \mathbf{1}_N^T V_N V_N^{-1} W_N/N = a^T W_N^T V_N^{-1} W_N/N$, so that for large samples g_{is} will be near $1/N$ for $i\epsilon s$. The design variance of $\bar{y}_{\text{AREG}}$ is then approximately equal to

$$\sum_{i\epsilon P} \sum_{j\epsilon P} \Delta_{ij} \tilde{e}_{iN} \tilde{e}_{jN}/N^2,$$

where $\Delta_{ij} = (\pi_{ij} - \pi_i\pi_j)$, and this may be estimated by

$$\hat{V}_1 = \sum_{i\epsilon s} \sum_{i\epsilon s} \tilde{\Delta}_{ij} \tilde{e}_{is} \tilde{e}_{js}/N^2. \tag{9}$$

$\hat{V}_1$ in (9) was considered in early work on the general regression estimator, for example, Särndal (1981, 1982). Now $\hat{V}_g$ in (8) may be thought of as a version of $\hat{V}_1$ in (9) adjusted for the realized values of g_{is}, $i\epsilon s$. Särndal, Swensson and Wretman (1989) show that $\hat{V}_g$ in (8), as well as being design consistent for the design variance of $\bar{y}_{\text{AREG}}$, is often model unbiased or nearly model unbiased for the model mean squared error of $\bar{y}_{\text{AREG}}$.

Now approximate confidence intervals for $\bar{y}_N$ could be constructed based on a standard normal approximation to the distribution of $(\bar{y}_{\text{AREG}} - \bar{y}_N)/\{\hat{V}_g\}^{1/2}$. The justification of this procedure, from both a design and a model point of view, is asymptotic and the question of its appropriateness for particular finite samples must be addressed. One possibility is to compare

a set of confidence intervals obtained by this procedure to a set of purely model based intervals based on a further assumption of normality of errors and a t-statistic. If the two sets of intervals are wildly different there may be reason to doubt the validity of the jointly model and design based intervals, but more work is needed before this question can be answered satisfactorily.

An alternative approach to variance estimation in this framework is given by Binder (1983). The design variance of $\boldsymbol{h}^*$ as an estimator of $\boldsymbol{g}^*$ at $\hat{\gamma}_N$ could be estimated using standard design based techniques substituting $\hat{\gamma}_s$ for $\hat{\gamma}_N$, and then the variance of $\hat{\gamma}_s$ as an estimator of $\hat{\gamma}_N$ would be derived from a Taylor linearization of $\boldsymbol{h}^*$ about $\hat{\gamma}_N$. Taylor linearization could again be used to derive an estimator of the variance of a function of $\hat{\gamma}_s$ as an estimator of the same function of $\hat{\gamma}_N$.

4. AREAS FOR FURTHER RESEARCH

We have seen how the approach described here could be used for the estimation of finite population means or, more generally, for functions of linear regression parameters. It is natural to wonder whether and how the approach may be adapted to the estimation of other types of finite population parameters such as distribution functions and quantiles or to estimation for small areas.

Consider the special case of estimation of a distribution function at one point. There are two possible approaches to incorporate covariate information into a model. The first is to model the probability explicitly as a function of the covariates, an example is the logistic model. A second approach, which is common in the context of estimating a distribution function, as in Chambers and Dunstan (1986), Rao, Kovar and Mantel (1990), and others, is to model the residuals from a regression of the observed variable onto the covariables as being independent and identically distributed from some unknown distribution. The present approach requires that the parameter of interest be a function of the finite population parameter. Can this approach be adapted for the estimation of distribution functions or quantiles?

Another important problem in survey sampling is small area estimation, that is estimation of totals, means or proportions for subsets of the finite population. A good review is given in Platek, Rao, Särndal and Singh (1987). An obvious adaptation of the approach of Section 1 is to apply it separately within each domain of interest, what might be described as post-stratified generalized regression estimation. Note that this approach would require the totals of the covariates for each domain of interest. A very common approach in small area estimation is to borrow strength across areas via a model relating small areas to each other and to some covariates. A good review is given in Singh, Mantel and Thomas (1991). A very fruitful approach has been the empirical Bayes estimation based on random effects models which was introduced by Fay and Herriot (1979). Liang and Waclawiw (1990) discuss estimating functions for empirical Bayes models. Can the idea of modelling to borrow strength across small areas be formulated in such a way that the parameters of interest become functions of a population parameter?

ACKNOWLEDGEMENT

I am grateful to a referee and to the editor for helpful comments and suggestions.

REFERENCES

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

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

GODAMBE, V.P., and THOMPSON, M.E. (1984). Robust estimation through estimating equations. *Biometrika*, 71, 115-125.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.

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

LIANG, K.-Y., and WACLAWIW, M.A. (1990). Extension of the Stein Estimating Procedure through the use of estimating functions. *Journal of the American Statistical Association*, 85, 435-440.

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

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (Eds.) (1987). *Small Area Statistics An International Symposium*. New York: Wiley.

SÄRNDAL, C.-E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. *Bulletin of the International Statistical Institute*, 49, 494-513.

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

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

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

SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1991). Time series generalizations of Fay-Herriot estimation for small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 225–239
Statistics Canada

Maximum Likelihood Estimation from Complex Sample Surveys

ABBA M. KRIEGER and DANNY PFEFFERMANN[1]

ABSTRACT

Maximum likelihood estimation from complex sample data requires additional modeling due to the information in the sample selection. Alternatively, pseudo maximum likelihood methods that consist of maximizing estimates of the census score function can be applied. In this article we review some of the approaches considered in the literature and compare them with a new approach derived from the ideas of 'weighted distributions'. The focus of the comparisons is on situations where some or all of the design variables are unknown or misspecified. The results obtained for the new method are encouraging, but the study is limited so far to simple situations.

KEY WORDS: Design adjusted estimators; Ignorable and informative designs; Pseudo likelihood; Weighted distributions.

1. INTRODUCTION

Survey data are often used for analytic inference about model parameters such as means, regression coefficients, cell probabilities *etc.* The models pertain to the population data and are therefore referred to as the census models. The problem in applying 'classical' maximum likelihood methods to survey data is that the model holding for the sample can be very different from the model holding for the population due to sample selection effects.

In order to illustrate the problem and some of the solutions proposed in the literature, consider the following simple example. A population U is made up of N units labelled $\{1, \ldots, N\}$. Associated with unit i is a vector (Y_i, Z_i) of independent measurements drawn from a bivariate normal distribution with mean $\underset{\sim}{\mu}' = (\mu_Y, \mu_Z)$ and variance-covariance $(V - C)$ matrix

$$\Sigma = \begin{bmatrix} \sigma_Y^2, & \sigma_{YZ} \\ \sigma_{YZ}, & \sigma_Z^2 \end{bmatrix}.$$

The values (y_i, z_i) are observed for a sample s of $n < < N$ units selected by a probability sampling scheme. It is desirable to estimate μ_Y and σ_Y^2. We consider three cases distinguished by the selection process and data availability.

Case A – The sample is selected by simple random sampling with replacement and only the values $\{(y_i, z_i), i \in s\}$ are known. Denoting the sample labels as $\{1, \ldots, n\}$, we have that $Y_1, \ldots, Y_n \underset{ind}{\sim} N(\mu_Y, \sigma_Y^2)$ yielding

$$\hat{\mu}_Y = \bar{y}_s = \sum_{i=1}^{n} y_i/n;\ \hat{\sigma}_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y}_s)^2/n = s_y^2 \tag{1.1}$$

as the MLE of μ_Y and σ_Y^2. Clearly $E_M(\hat{\mu}_Y) = \mu_Y$ and $E_M\{[n/(n-1)]\hat{\sigma}_Y^2\} = \sigma_y^2$ where $E_M\{\cdot\}$ defines the expectation under the model, with the sample units held fixed.

[1] Abba M. Krieger, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104. Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905.

Case B - The sample is selected with probabilities proportional to z_i with replacement such that at each draw $k = 1, \ldots, n$, $P_i = P(i \epsilon s) = z_i / \sum_{j=1}^{N} z_j$. The data known to the analyst are $\{y_i, z_i, i \epsilon s\}$ and $\{z_{n+1}, \ldots, z_N\}$. Suppose that $\text{Corr}(Y,Z) > 0$. This implies that $P(Y_i > \mu_Y \mid i \epsilon s) > 1/2$ since the sampling scheme tends to select units with large values of Z and hence large values of Y. Clearly, the estimators defined in (1.1) are no longer MLE in this case.

The situation just described corresponds to the 'classical' example of missing data often analyzed in the literature (Anderson 1957). The MLE of μ_Y and σ_Y^2 are now

$$\hat{\mu}_Y = \bar{y}_s + b(\bar{Z} - \bar{z}_s); \hat{\sigma}_Y^2 = s_Y^2 + b^2(S_Z^2 - s_Z^2), \tag{1.2}$$

where $\bar{Z} = \sum_{i=1}^{N} z_i/N$, $\bar{z}_s = \sum_{i=1}^{n} z_i/n$, $b = \sum_{i=1}^{n} (y_i - \bar{y}_s)(z_i - \bar{z}_s) / \sum_{i=1}^{n} (z_i - \bar{z}_s)^2$, $S_Z^2 = \sum_{i=1}^{N} (z_i - \bar{Z})^2/N$ and $s_Z^2 = \sum_{i=1}^{n} (z_i - \bar{z}_s)^2/n$. Notice that the effect of the sample selection can be dealt with in this case by modeling the joint distribution of the response variable Y and the design variable Z. The sample selection process is then **ignorable** (see section 2.1).

Case C - Same as Case B but only the sample values $\{(y_i, z_i), i \epsilon s\}$ and the sample selection probabilities $\{P_i, i \epsilon s\}$ are known. Even though the values of z_i, $i = 1, \ldots, N$ are known at the sampling stage, it is often the case that information on the design variables or the inclusion probabilities for units outside the sample is not included in the files released to analysts performing secondary analysis.

The estimators defined by (1.2) are no longer operational in this case since the population mean and variance of Z are unknown. For large populations, however, such that $\bar{Z} \cong$ constant, an approximate MLE estimator of μ_Y is obtained as $\mu_Y^* = \bar{y}_s + b^*(1/N - \bar{P}_s)$ where $\bar{P}_s = \sum_{i=1}^{n} P_i/n$ and $b^* = \sum_{i=1}^{n} (y_i - \bar{y}_s)(P_i - \bar{P}_s) / \sum_{i=1}^{n} (P_i - \bar{P}_s)^2$. The rationale for μ_Y^* is that $P_i = Z_i/N\bar{Z}$ so that for $\bar{Z} =$ constant, (Y_i, P_i) is bivariate normal with $\bar{P} = \sum_{i=1}^{N} P_i/N = 1/N$. This estimator is an example of using the sample selection probabilities as surrogates for the design variables when information on the latter is incomplete, as recommended in Rubin (1985).

A possible way to obtain approximate MLE under Case C is to follow what is known in the literature as the pseudo likelihood approach. We describe the approach in more detail in section 2, but it basically consists of maximizing a design consistent estimator of the census score function, that is, the score function that would have been obtained in the case of a census. The latter is unaffected by the design. Application of this approach yields, under Case C the estimators

$$\tilde{\mu}_Y = \bar{y}_{ps} = \sum_{i=1}^{n} w_i^* y_i \Big/ \sum_{i=1}^{n} w_i^*; \tilde{\sigma}_Y^2 = s_p^2 = \sum_{i=1}^{n} w_i^* (y_i - \bar{y}_{ps})^2 \Big/ \sum_{i=1}^{n} w_i^*, \tag{1.3}$$

where $w_i^* = (1/nP_i)$. Since $\bar{y}_{ps}$ and s_p^2 are design consistent for $\bar{Y} = \sum_{i=1}^{N} y_i/N$ and $S_Y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2/N$ respectively, they are also consistent for μ_Y and σ_Y^2 in the sense that $\text{plim}_{n \to \infty, N \to \infty} (\bar{y}_{ps}, s_p^2) = (\mu_Y, \sigma_Y^2)$.

In this article we discuss a different approach for maximum likelihood estimation that is operational in principle even when the only information available to the analyst is the sample data. The method is derived from the theory of weighted distributions (Rao 1965, 1985, Patil and Rao 1978) and it utilizes the sample selection probabilities. The method is illustrated for the case of normal distributions with two different sampling designs and is shown to perform well in these cases. Another apparent advantage of the proposed approach emerging from the empirical study is that it is not very sensitive to misspecification of the design variables.

In section 2 we review the different approaches for MLE from survey data considered in the literature. Section 3 outlines the basic steps of the new approach. The empirical study is described and summarized in section 4. Section 5 contains concluding remarks.

2. REVIEW OF APPROACHES CONSIDERED IN THE LITERATURE

In this section we review briefly the approaches considered in the literature for MLE or approximate MLE from survey data. To better understand the complexity of the problem, we first discuss the notion of **ignorable sampling designs**. For a more detailed review of maximum likelihood and other approaches for analytic inferences from sample surveys see Pfeffermann (1993).

2.1 Ignorable and Informative Sampling Designs

Let $\underset{\sim}{Z}' = (Z_1, \ldots, Z_K)$ represent K design (auxiliary) variables used for designing the survey and denote by $Z = (\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_N)'$ the $N \times K$ matrix of measurements on $\underset{\sim}{Z}$ so that $\underset{\sim}{z}_i$ is the vector associated with unit i. The design variables may include strata indicator variables and quantitative measurements of cluster and unit characteristics. Let $\underset{\sim}{Y}' = (Y_1, \ldots, Y_p)$ represent the survey response variables. We assume for convenience that $\underset{\sim}{Y}$ is separate from $\underset{\sim}{Z}$ although as we mention below and consider in the empirical study, the sample selection probabilities may depend on the Y-values directly. The matrix $Y = (\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_N)$ of the response variables values can be decomposed as $Y = [Y_s, Y_{\bar{s}}]$ where $Y_s = \{\underset{\sim}{y}_i, i \in s\}$ and $Y_{\bar{s}} = \{\underset{\sim}{y}_i, i \notin s\}$. Let $\underset{\sim}{I} = (I_1, \ldots, I_N)'$ be a vector of sample inclusion indicators such that $I_i = 1$ for $i \in s$ and $I_i = 0$ otherwise.

The basic problem of MLE from complex survey data, as illustrated in the introduction, is that in general, $f(Y_s;\underset{\sim}{\lambda}^*) \neq \int f(Y;\underset{\sim}{\lambda}) dY_{\bar{s}}$ where the symbol $f(\cdot\,;\cdot)$ defines probability density functions (pdf). As further illustrated in the introduction, this problem can sometimes be resolved by modeling the joint distribution of Y and Z. Thus, suppose that the values of $\underset{\sim}{Z}$ are known for every unit in the population and that $\underset{\sim}{Y}$ is observed for only the sample units. The joint pdf of all the available data can be written as

$$f(Y_s,\underset{\sim}{I},Z;\underset{\sim}{\theta},\underset{\sim}{\phi},\underset{\sim}{\rho}) = \int f(Y_s,Y_{\bar{s}} \mid Z;\underset{\sim}{\theta}_1) P(\underset{\sim}{I} \mid Y,Z;\underset{\sim}{\rho}_1)\, g(Z;\underset{\sim}{\phi}) dY_{\bar{s}}. \tag{2.1}$$

Ignoring the sampling selection in the inference process implies that inference is based on the joint distribution of Y_s and Z, that is, the probability $P(\underset{\sim}{I} \mid Y, Z; \underset{\sim}{\rho}_1)$ on the right hand side of (2.1) is ignored. Hence the inference is based on

$$f(Y_s, Z; \underset{\sim}{\theta},\underset{\sim}{\phi}) = \int f(Y_s,Y_{\bar{s}} \mid Z;\underset{\sim}{\theta}_1)\, g(Z;\underset{\sim}{\phi}) dY_{\bar{s}}. \tag{2.2}$$

The sample selection is said to be ignorable when inference based on (2.1) is equivalent to inference based on (2.2). This is clearly the case for sampling designs that depend only on the design variables $\underset{\sim}{Z}$, since in this case $P(\underset{\sim}{I} \mid Y, Z;\underset{\sim}{\rho}_1) = P(\underset{\sim}{I} \mid Z;\underset{\sim}{\rho}_1)$. The exact conditions for the ignorability of the sample selection process are defined and illustrated in the articles by Rubin (1976), Little (1982) and Sugden and Smith (1984).

The complications of MLE from complex survey data based on (2.1) or (2.2) are now apparent. First and foremost, it requires that all the relevant design variables be identified and known at the population level. As often argued in the literature, (see Pfeffermann 1993 for references), this is not necessarily the case. Secondly, it requires that the sample selection is ignorable in the sense discussed above or alternatively that the probabilities $P(\underset{\sim}{I} \mid Y,Z;\underset{\sim}{\rho})$ be modeled and included in the likelihood. Finally, the use of MLE requires the specification of the joint pdf $f(Y,Z;\underset{\sim}{\theta},\underset{\sim}{\phi}) = f(Y \mid Z;\underset{\sim}{\theta}_1) g(Z;\underset{\sim}{\phi})$.

2.2 Exact MLE Based on Factorization of the Likelihood

Factoring the likelihood in the case of multivariate normal data was first suggested by Anderson (1957). The factorization is possible when the observed data have a nested pattern, that is, the set of survey variables $X_1, \ldots, X_p$ can be arranged such that X_j is observed for all units where X_{j+1} is observed, $j = 1, \ldots, (p-1)$. Extensions to other distributions and more general data patterns are given in Rubin (1974). Holt, Smith and Winter (1980) apply the ideas to MLE of regression coefficients from complex survey data.

Suppose that the sample selection is ignorable so that inference can be based on the joint distribution $f(Y_s,Z;\underset{\sim}{\theta},\underset{\sim}{\phi}) = f(Y_s \mid Z;\underset{\sim}{\theta}_1)\, g(Z;\underset{\sim}{\phi})$. The likelihood can be factored accordingly as

$$L(\underset{\sim}{\theta},\underset{\sim}{\phi};Y_s,Z) = L(\underset{\sim}{\theta}_1;Y_s \mid Z)\, L(\underset{\sim}{\phi};Z). \tag{2.3}$$

Assuming that the parameters $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\phi}$ are distinct in the sense of Rubin (1976), MLE of $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\phi}$ can be calculated independently from the two components.

Application of (2.3) to the case where $(\underset{\sim}{Y}_i',\underset{\sim}{Z}_i')$ are multivariate normal yields the following MLE for $\underset{\sim}{\mu}_Y = E(\underset{\sim}{Y})$ and $\sum_Y = V(\underset{\sim}{Y})$ (Anderson 1957).

$$\hat{\underset{\sim}{\mu}}_Y = \bar{\underset{\sim}{y}}_s + \hat{\underset{\sim}{\beta}}(\bar{\underset{\sim}{Z}} - \bar{\underset{\sim}{z}}_s); \quad \hat{\sum}_Y = s_{YY} + \hat{\underset{\sim}{B}}[S_{ZZ} - s_{ZZ}]\hat{\underset{\sim}{B}}', \tag{2.4}$$

where $(\bar{\underset{\sim}{y}}_s,\bar{\underset{\sim}{z}}_s) = \sum_{i=1}^{n}(\underset{\sim}{y}_i,\underset{\sim}{z}_i)/n$, $\bar{\underset{\sim}{Z}} = \sum_{i=1}^{N}\underset{\sim}{z}_i/N, S_{ZZ} = \sum_{i=1}^{N}(\underset{\sim}{z}_i - \bar{\underset{\sim}{Z}})(\underset{\sim}{z}_i - \bar{\underset{\sim}{Z}})'/N$, $s_{ZZ} = \sum_{i=1}^{n}(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}}_s)(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}}_s)'/n$ and $\hat{\underset{\sim}{B}} = \sum_{i=1}^{n}(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}_s)(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}}_s)' s_{ZZ}^{-1}/n$.

The MLE of the coefficient matrix B_{12} of the multivariate regression of $\underset{\sim}{Y}_1$ on $\underset{\sim}{Y}_2$ where $\underset{\sim}{Y}' = (\underset{\sim}{Y}_1',\underset{\sim}{Y}_2')$ is obtained straightforwardly from (2.4). Thus, if

$$\sum_Y = \begin{bmatrix} \sum_{11}, & \sum_{12} \\ \sum_{21}, & \sum_{22} \end{bmatrix},$$

where

$$\sum_{ij} = \mathrm{COV}[(\underset{\sim}{Y}_i',\underset{\sim}{Y}_j')'], \quad i,j = 1,2, \quad B_{12} = \sum_{12}\sum_{22}^{-1} \text{ and } \hat{B}_{12} = \hat{\sum}_{12}\hat{\sum}_{22}^{-1}.$$

For the explicit expression of $\hat{B}_{12}$ see Holt, Smith and Winter (1980).

2.3 Design Adjusted Estimators (DAE)

Assume that the sample selection mechanism is ignorable. Let $\ell_N(\underset{\sim}{\theta};Y)$ denote the log likelihood for θ that would be obtained in the case of a census. Denote by $h_N(Y \mid Z,Y_s;\underset{\sim}{\theta}_2)$ the conditional distribution of Y given Z and Y_s and let $E_{h_N}(\cdot \mid Z,Y_s)$ define the expectation operator under h_N. The DAE $\hat{\underset{\sim}{\theta}}_{ND}$ of $\underset{\sim}{\theta}$ as proposed by Chambers (1986) is defined as

$$E_{h_N}[-\ell_N(\hat{\underset{\sim}{\theta}}_{ND}) \mid Z,Y_s] = \min\{E_{h_N}[-\ell_N(\underset{\sim}{\theta}) \mid Z,Y_s];\underset{\sim}{\theta}\in\Theta\}. \tag{2.5}$$

Notice that the expectation $E_{ND}(\theta) = E_{h_N}[\ell_N(\theta) \mid Z,Y_s]$ depends on the vector parameter $\underset{\sim}{\theta}_1$ of the conditional distribution $f(Y \mid Z;\underset{\sim}{\theta}_1)$. The estimator $\hat{\underset{\sim}{\theta}}_{ND}$ of (2.5) is computed by substituting $\hat{\underset{\sim}{\theta}}_1$ for $\underset{\sim}{\theta}_1$ where $\hat{\underset{\sim}{\theta}}_1$ is the MLE of $\underset{\sim}{\theta}_1$ obtained from the data (Y_s,Z).

Simple algebra shows that for the multivariate normal model considered in section 2.2, the DAE of μ_Y and Σ_Y are the same as the MLE defined by (2.4). A possible advantage of this approach, however, is that it can be applied to other loss functions.

2.4 The Pseudo Likelihood Approach

The prominent feature of this approach is that it utilizes the sample selection probabilities to estimate the census likelihood equations. The estimated equations are then maximized with respect to the vector parameter of interest. No information on the values of the design variables is needed, although as illustrated in the empirical study, knowledge of these values at the population level can be used to improve the efficiency of the estimators.

Suppose that the population values $\underset{\sim}{Y}_i$ are independent draws from a common distribution $f(\underset{\sim}{Y};\underset{\sim}{\theta})$ and let $\ell_N(\underset{\sim}{\theta};\underset{\sim}{Y}) = \sum_{i=1}^{N} \log f(\underset{\sim}{Y}_i;\underset{\sim}{\theta})$ define the census log likelihood function. Under some regularity conditions, the MLE, $\hat{\underset{\sim}{\theta}}$, solves the equations

$$\underset{\sim}{U}(\underset{\sim}{\theta}) = d\ell_N(\underset{\sim}{\theta};\underset{\sim}{Y})/d\underset{\sim}{\theta} = \sum_{i=1}^{N} \underset{\sim}{u}(\underset{\sim}{\theta};\underset{\sim}{y}_i) = \underset{\sim}{0}, \tag{2.6}$$

where "d" defines the derivative operator and $\underset{\sim}{u}(\underset{\sim}{\theta},\underset{\sim}{y}_i) = d \log f(\underset{\sim}{Y}_i;\underset{\sim}{\theta})/d\underset{\sim}{\theta}$. The pseudo MLE of $\underset{\sim}{\theta}$ is defined as the solution of $\hat{\underset{\sim}{U}}(\underset{\sim}{\theta}) = \underset{\sim}{0}$ where $\hat{\underset{\sim}{U}}(\underset{\sim}{\theta})$ is a design consistent estimator of $\underset{\sim}{U}(\underset{\sim}{\theta})$ in the sense that $\text{plim}_{n\to\infty,N\to\infty}[\hat{\underset{\sim}{U}}(\underset{\sim}{\theta}) - \underset{\sim}{U}(\underset{\sim}{\theta})] = \underset{\sim}{0}$ for all $\underset{\sim}{\theta}\epsilon\Theta$. The commonly used estimator of $\underset{\sim}{U}(\underset{\sim}{\theta})$ is the Horvitz-Thompson (1952) estimator so that the pseudo MLE of $\underset{\sim}{\theta}$ is the solution of $\hat{\underset{\sim}{U}}(\underset{\sim}{\theta}) = \sum_{i=1}^{n} w_i^* \underset{\sim}{u}(\underset{\sim}{\theta};\underset{\sim}{y}_i) = \underset{\sim}{0}$ where for selection without replacement $w_i^* = [1/P(i\epsilon s)]$ and for selection with replacement $w_i^* = (1/nP_i)$.

For the multivariate normal model, the pseudo MLE of $\underset{\sim}{\mu}_Y$ and Σ_Y are

$$\tilde{\underset{\sim}{\mu}}_Y = \sum_{i=1}^{n} w_i^* \underset{\sim}{y}_i \Big/ \sum_{i=1}^{n} w_i^*; \quad \tilde{\Sigma}_Y = \sum_{i=1}^{n} w_i^* (\underset{\sim}{y}_i - \tilde{\underset{\sim}{\mu}}_Y)(\underset{\sim}{y}_i - \tilde{\underset{\sim}{\mu}}_Y)' \Big/ \sum_{i=1}^{n} w_i^*. \tag{2.7}$$

The pseudo MLE of the matrix coefficients B_{12} is obtained as $\tilde{B}_{12} = \tilde{\Sigma}_{12}\,\tilde{\Sigma}_{22}^{-1}$.

Various examples for the use of this approach under different models can be found in Skinner *et al.* (1989). See also Binder (1983), Chambless and Boyle (1985), Roberts, Rao and Kumar (1987) and Pfeffermann (1988).

Information on auxiliary design variables known at the population level can be used to improve the efficiency of the design estimators of $\underset{\sim}{U}(\underset{\sim}{\theta})$. The "probability weighted MLE" as proposed by Nathan and Holt (1980) and by Smith and Holmes (Skinner *et al.* 1989, Ch. 8) are examples of the use of the population values of the design variables. The estimators have the same structure as the exact MLE derived from (2.4) but with unweighted sample statistics replaced by weighted statistics. For example, $(\bar{y}_s,\bar{z}_s)$ in (2.4) are replaced by $\sum_{i=1}^{n} w_i^* (\underset{\sim}{y}_i,\underset{\sim}{z}_i) / \sum_{i=1}^{n} w_i^*$, with similar substitutions for the other expressions.

An important property of pseudo MLE is that they are in general design consistent for the population quantities that would be obtained by solving the corresponding census likelihood equations, irrespective of whether the model is correct and/or whether the sampling design is informative. See Pfeffermann (1993) for the implications of this property with references to other studies. Other theoretical properties of pseudo MLE are studied by Godambe and Thompson (1986).

3. MLE DERIVED FROM WEIGHTED DISTRIBUTIONS

3.1 General Formulation

The weighted pdf of a random variable X^w is defined as

$$f^w(x) = w(x)f(x)/w, \tag{3.1}$$

where $f(x)$ is the unweighted pdf and $w = \int w(x)f(x)dx = E[w(X)]$ is the normalizing factor making the total probability equal to unity. Situations leading to weighted distributions occur when realizations x from $f(x)$ are observed and recorded with differential probabilities $w(x)$. The expectation w is then the probability of recording an observation and $f^w(x)$ is the pdf of the resulting random variable X^w.

The concept of weighted distributions was introduced by Rao (1965). Patil and Rao (1978) discuss various practical situations that give rise to pdf's of the form (3.1). One special case that occurs in many applications is when $w(x) = |x|$ where $|x|$ is some measure of the size of x. The pdf obtained in this case is called 'size biased' or 'length biased'. The properties of that distribution under a variety of densities $f(x)$ are examined in Cox (1969) and Patil and Rao (1978). Estimation of weighted distributions is considered by Vardi (1982).

How can the concept of weighted distributions be utilized for analytic inference from complex samples? Consider as before a finite population $U = \{1, \ldots, N\}$ with random measurements $X(i) = \underset{\sim}{x}_i' = (\underset{\sim}{y}_i', \underset{\sim}{z}_i')$ generated independently from a common pdf $h(\underset{\sim}{x};\underset{\sim}{\delta}) = f(\underset{\sim}{y}_i \mid \underset{\sim}{z}_i;\underset{\sim}{\theta}_1)\, g(\underset{\sim}{z}_i;\underset{\sim}{\phi})$. Suppose that unit i is sampled with probability $w(\underset{\sim}{x}_i;\underset{\sim}{\alpha})$ that depends on the measurements $\underset{\sim}{x}_i$ and possibly also on an unknown vector parameter $\underset{\sim}{\alpha}$. Denote by X_i^w the measurements recorded for unit $i \in s$. The pdf of X_i^w is then

$$\begin{aligned} h^w(\underset{\sim}{x}_i;\, \underset{\sim}{\alpha}, \underset{\sim}{\delta}) &= f(\underset{\sim}{x}_i \mid i \in s) = P[i \in s \mid X(i) = \underset{\sim}{x}_i]\, h(\underset{\sim}{x}_i;\underset{\sim}{\delta})/P(i \in s) \\ &= w(\underset{\sim}{x}_i;\, \underset{\sim}{\alpha})h(\underset{\sim}{x}_i;\underset{\sim}{\delta}) \Big/ \int w(\underset{\sim}{x}_i;\underset{\sim}{\alpha})h(\underset{\sim}{x}_i;\underset{\sim}{\delta})d\underset{\sim}{x}_i. \end{aligned} \tag{3.2}$$

Analytic inference focuses on the vector parameter $\underset{\sim}{\delta}$ or functions thereof as the target parameters. Let $s = \{1, \ldots, n\}$ define a sample of fixed size $n << N$ selected with replacement such that at each draw $k = 1, \ldots, n$, $P(j \in s) = w(\underset{\sim}{x}_j;\underset{\sim}{\alpha}), j = 1, \ldots, N$. The joint pdf of $\{X_i^w, i = 1, \ldots, n\}$ is then $\Pi_{i=1}^n h^w(\underset{\sim}{x}_i;\underset{\sim}{\alpha},\underset{\sim}{\delta})$ so that the likelihood is

$$L(\underset{\sim}{\delta};X_s,s) = \text{const} \times \Pi_{i=1}^n h(\underset{\sim}{x}_i;\underset{\sim}{\delta}) \Big/ \left[\int w(\underset{\sim}{x};\underset{\sim}{\alpha})h(\underset{\sim}{x};\underset{\sim}{\delta})d\underset{\sim}{x}\right]^n, \tag{3.3}$$

where $X_s' = [\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n]$. The likelihood (3.3) has the following properties:

(1) It is defined in terms of the vector parameter $\underset{\sim}{\delta}$. This has an advantage over the use of the factorized likelihood (2.3) where $\underset{\sim}{\delta}$ does not enter the likelihood directly.

(2) It is a function of the selection probabilities $w(\underset{\sim}{x}_i;\underset{\sim}{\alpha})$ that enter into the denominator.

(3) The likelihood relates to the conditional distribution of the sample data given the units in the sample. This is different from the likelihood derived from the pdf in (2.1) which is the joint pdf of the sample data and the vector $\underset{\sim}{I}$ of sample indicators. An example of the use of the latter pdf in conjunction with weighted distributions for MLE is given in Godambe and Rajarshi (1989).

(4) The use of the likelihood (3.3) requires a definition of the joint pdf $h(\underset{\sim}{x};\underset{\sim}{\delta})$ holding in the population and a specification of the relationship between the sample selection probabilities and the variables observed for the sample. The need to define the population pdf is common to all of the approaches for MLE proposed in the literature. The specification of the functions $\underset{\sim}{w}(\underset{\sim}{x})$ is unique to the present approach. This step can be carried out however by modeling the empirical relationship in the sample between the selection probabilities and the observed measurements. Having identified a suitable model, the probabilities $w(\underset{\sim}{x},\underset{\sim}{\alpha})$ can be estimated from the sample and the estimates can be substituted into the likelihood. In what follows we consider two examples which are analyzed empirically in section 4.

3.2 Examples

We assume the model considered in section 2 in which $\underset{\sim}{X}_i' = (\underset{\sim}{Y}_i', \underset{\sim}{Z}_i')$ are independent realizations from a multivariate normal distribution with mean $\underset{\sim}{\mu}_x' = (\underset{\sim}{\mu}_Y', \underset{\sim}{\mu}_Z')$ and $V - C$ matrix

$$\sum_{XX} = \begin{bmatrix} \sum_{YY}, & \sum_{YZ} \\ \sum_{ZY}, & \sum_{ZZ} \end{bmatrix}. \tag{3.4}$$

Consider the following sampling designs:

D1 – **PPS selection with replacement**: Let $T_i = \underset{\sim}{\alpha}_1' \underset{\sim}{Y}_i + \underset{\sim}{\alpha}_2' \underset{\sim}{Z}_i$ define a single design variable and suppose that the sample is selected with probabilities proportional to the T-values such that at each draw $k = 1, \ldots, n, P(i \epsilon s) = t_i/N\bar{T}, i = 1, \ldots, N$ where $\bar{T} = \sum_{j=1}^{N} t_j/N$. We assume that N is large enough so that the difference between $\bar{T}$ and $\mu_T = E(T)$ can be ignored. The coefficients $\underset{\sim}{\alpha} = (\underset{\sim}{\alpha}_1', \underset{\sim}{\alpha}_2')$ are fixed. In special cases $\underset{\sim}{\alpha}_1 = \underset{\sim}{0}$ hence T is a function of only the auxiliary design variables $\underset{\sim}{Z}$ or $\underset{\sim}{\alpha}_2 = 0$ in which case T is only a function of the response variables $\underset{\sim}{Y}$. Suppose as before that it is desirable to estimate the mean $\underset{\sim}{\mu}_Y$ and the $V - C$ matrix $\sum_{YY}$ or functions thereof.

When $\underset{\sim}{\alpha}_1 = \underset{\sim}{0}$ and T is known for every unit in the population, one can estimate the unknown parameters using the factorization (2.3). The corresponding MLE are given in (2.4) with $\underset{\sim}{Z}$ replaced by T. Suppose however that the only information available to the analyst is the sample values $\underset{\sim}{x}_i' = (\underset{\sim}{y}_i', z_i'), i = 1, \ldots, n$ and the sample selection probabilities $P_i = t_i/N\bar{T}$. Under the assumption $\bar{T} = \mu_T$, the likelihood for $[\underset{\sim}{\mu}_X, \sum_{XX}]$ can be written using (3.3) as

$$L(\underset{\sim}{\mu}_X, \sum_{XX}; X_s, s) = \prod_{i=1}^{n} (\underset{\sim}{\alpha}' \underset{\sim}{x}_i) \phi(\underset{\sim}{x}_i; \underset{\sim}{\mu}_X, \sum_{XX}) / (\underset{\sim}{\alpha}_1' \underset{\sim}{\mu}_Y + \underset{\sim}{\alpha}_2' \underset{\sim}{\mu}_Z)^n, \tag{3.5}$$

where $\phi(\underset{\sim}{x}; \underset{\sim}{\mu}_X, \sum_{XX})$ is the normal pdf with mean $\underset{\sim}{\mu}_X$ and $V - C$ matrix $\sum_{XX}$. The likelihood in (3.5) is a function also of the unknown vector coefficients $\underset{\sim}{\alpha}$. However, the values of $\underset{\sim}{\alpha}$ can actually be found up to a constant c (which cancels out in the likelihood) by regressing the sample selection probabilities P_i against $\underset{\sim}{\alpha}$.

In the simulation study described in section 4, we consider the case where not all the design variables are known even for the sample units. Thus, suppose that $\underset{\sim}{Z}_i' = (Z_{1i}, Z_{2i})$ and that the data available to the analyst consist of the selection probabilities $P_i, i = 1, \ldots, n$ and the observations $\{\underset{\sim}{x}_i^{*\prime} = (\underset{\sim}{y}_i', z_{1i}), i = 1, \ldots, n\}$. The likelihood (3.3) is now

$$L(\underset{\sim}{\mu}_X^*, \underset{\sim}{\Sigma}_{XX}^*; X_s^*, s) = \prod_{i=1}^{n} w(\underset{\sim}{x}_i^*)\phi(\underset{\sim}{x}_i^*; \underset{\sim}{\mu}_X^*, \underset{\sim}{\Sigma}_{XX}^*)/(w^*)^n, \tag{3.6}$$

where $w(\underset{\sim}{x}_i^*)$ are the selection probabilities expressed as functions of $\underset{\sim}{x}_i^*$. Clearly, the probabilities $w(\underset{\sim}{x}_i^*)$ are not fully determined by the values $\underset{\sim}{x}_i^*$ unless $\alpha_{22} = 0$. Assuming normality

$$w(\underset{\sim}{x}_i, \underset{\sim}{\alpha}) = \alpha_0^* + \underset{\sim}{\alpha}_1^{*\prime}\underset{\sim}{y}_i + \alpha_2^* z_{1i} + \epsilon_i, \tag{3.7}$$

where $\{\epsilon_i\}$ is white noise. Thus, the likelihood (3.6) can be approximated by substituting $w^*(\underset{\sim}{x}_i^*) = \alpha_0^* + \underset{\sim}{\alpha}_1^{*\prime}\underset{\sim}{y}_i + \alpha_2^* z_{1i}$ for $w(\underset{\sim}{x}_i^*)$. The values of $\underset{\sim}{\alpha}^* = (\alpha_0^*, \underset{\sim}{\alpha}_1^{*\prime}, \alpha_2^*)'$ can be estimated from the regression (3.7) and then substituted into the likelihood.

D2 – **Stratified sampling with T as the stratification variable**: Suppose that the population U is divided into L strata $U_1, \ldots, U_L$ of sizes $N_1, \ldots, N_L$, $\sum_{h=1}^{L} N_h = N$, based on the ascending values of T. Consider a simple random stratified sample of size $n = \sum_{h=1}^{L} n_h$ selected without replacement with fixed sample sizes $\{n_h\}$. The weighted pdf of X_i^w, the measurements recorded for unit $i \in s$ is in this case [compare with (3.2)]

$$h^w(\underset{\sim}{x}_i; \underset{\sim}{\alpha}, \underset{\sim}{\delta}) = f(\underset{\sim}{x}_i \mid i \in s) = \begin{cases} P_1 h(\underset{\sim}{x}_i; \underset{\sim}{\delta})/w & \text{if} \quad t_i \le t^{(1)} \\ P_2 h(\underset{\sim}{x}_i; \underset{\sim}{\delta})/w & \text{if} \quad t^{(1)} \le t_i \le t^{(2)} \\ \quad \vdots \qquad\qquad \vdots & \qquad\qquad \vdots \\ P_L h(\underset{\sim}{x}_i; \underset{\sim}{\delta})/w & \text{if} \quad t^{(L-1)} \le t_i \end{cases} \tag{3.8}$$

where $P_h = (n_h/N_h)$ and for $\{N_h\}$ sufficiently large, the probability $w = P(i \in S)$ can be closely approximated as

$$w = P(i \in s) \approx P_1 \int_{-\infty}^{t^{(1)}} \phi(t)dt + \sum_{h=2}^{L-1} P_h \int_{t^{(h-1)}}^{t^{(h)}} \phi(t)dt + P_L \int_{t^{(L-1)}}^{\infty} \phi(t)dt, \tag{3.9}$$

where $\phi(t)$ denotes the normal pdf of T.

Suppose that the strata are large enough so that selection within the strata can be considered as independent. Define $\mu_T = E(T) = \underset{\sim}{\alpha}'\underset{\sim}{\mu}_X$, $\sigma_T^2 = \mathrm{Var}(T) = \underset{\sim}{\alpha}'\underset{\sim}{\Sigma}_{XX}\underset{\sim}{\alpha}$ and let $\Phi_h = \int_{-\infty}^{t^{(h)}} \phi(t)dt$. For given boundaries $\{t^{(h)}\}$ and the vector coefficients $\underset{\sim}{\alpha}$, the likelihood for $\underset{\sim}{\delta}$ can be written as

$$L(\underset{\sim}{\delta}; X_s, s) = \text{const} \times \prod_{i=1}^{n} h(\underset{\sim}{x}_i; \underset{\sim}{\delta}) \prod_{h=1}^{L} P_h^{n_h} \Big/$$

$$\left\{P_1 \Phi_1 + \sum_{h=2}^{L-1} P_h[\Phi_h - \Phi_{h-1}] + P_L[1 - \Phi_{L-1}]\right\}^n. \tag{3.10}$$

Hausman and Wise (1981) use a variant of the likelihood (3.10) for estimating the vector of regression coefficients in a situation where the strata boundaries are determined by the values of the dependent variable. They assume that the strata boundaries are known, but allow the selection probabilities within the strata to be unknown in which case they are included in the set of unknown parameters with respect to which the likelihood is maximized.

In many practical situations, the strata boundaries are unknown and have to be estimated from the sample data. When the data include the values $\{t_i, i = 1, \ldots, n\}$, the vector $\underset{\sim}{\alpha}$ can be estimated from the regression of t_i on $\underset{\sim}{x}_i$, as in the PPS example discussed before. Furthermore, if $(t_{(1)} \leq \ldots \leq t_{(n)})$ are the ordered values of the t_i's, the strata boundaries can be estimated as, $t^{(1)} = \frac{1}{2}(t_{(n_1)} + t_{(n_1+1)}) \ldots t^{(L-1)} = \frac{1}{2}(t_{(n^*)} + t_{(n^*+1)})$ where $n^* = \sum_{h=1}^{L-1} n_h$. Substituting these estimates into (3.10) yields an approximation to the likelihood which can then be maximized as a function of $\underset{\sim}{\delta}$.

The situation is more complicated when the values t_i are unknown even for units in the sample. In the simulation study we attempt to deal with this problem by predicting t_i using Fisher's Linear Discriminant Function, that is, specifying the vector coefficients $\hat{\underset{\sim}{\alpha}}$ to be such that it maximizes the ratio of the between groups sum of squares to the within groups sum of squares of linear combinations $\underset{\sim}{\alpha}' \underset{\sim}{X}_i$. The groups are the strata. Once the predictors $\hat{t}_i = \hat{\underset{\sim}{\alpha}}' \underset{\sim}{x}_i$ are formed, the strata boundaries are estimated as in the previous case but with $\hat{t}_i$ instead of t_i. Also, $\hat{\mu}_T = \hat{\underset{\sim}{\alpha}}' \underset{\sim}{\mu}_X$ and $\hat{\sigma}_T^2 = \hat{\underset{\sim}{\alpha}}' \underset{\sim}{\Sigma}_{XX} \hat{\underset{\sim}{\alpha}}$. Substituting these estimators in (3.10) yields an approximation to the likelihood which can be maximized with respect to $\underset{\sim}{\delta}$.

As in the PPS example, the likelihood (3.10) can be modified to the case where only some of the design variables are known or observed. Maximization of the modified likelihood is carried out following the same steps as above.

4. SIMULATION RESULTS

4.1 General

In order to illustrate and compare the performance of the various MLE procedures described in this paper, we ran a small simulation study which consists of two stages. In the first stage we generated a single finite population of size $N = 8{,}000$ such that $\underset{\sim}{x}_i' = (y_{1i}, y_{2i}, z_{1i}, z_{2i})$, $i = 1, \ldots, 8{,}000$ are multivariate normal. In the second stage we selected independent samples of size $n = 800$ using the two sampling schemes described in section 3.2 with two different definitions for the design variable. The number of samples selected in each case was 300. We computed the various estimators for each of the samples based on the available sample data and then computed the empirical bias and root mean square error (RMSE) over the selected samples. In order to study and compare the conditional properties of the estimators considered, we classified the 300 samples selected in each case into 10 groups, based on the ascending values of the sample mean of the design variable and computed the bias and RMSE within each of the groups. In what follows we describe the various stages in some more detail.

4.2 Generation of the Population Values and Sample Selection Schemes

Values of z_{1i} and z_{2i} were generated independently from a normal $(20, 10^2)$ distribution. Values y_{1i} were generated as $y_{1i} = z_{1i} + z_{2i} + \epsilon_{1i}$; $\epsilon_{1i} \sim N(0, 10^2)$. Values y_{2i} were generated as $y_{2i} = y_{1i} + 0.5z_{1i} + 0.5z_{2i} + \epsilon_{2i}$; $\epsilon_{2i} \sim N(0, 20^2)$.

We employed the two sampling schemes described in section 3.2 using two different definitions for the design size variable. (i) $t_i = 0.5(z_{1i} + z_{2i})$ and (ii) $t_i = 0.25(y_{1i} + y_{2i} + z_{1i} + z_{2i})$. Thus, selection based on the first design variable satisfies the ignorability conditions defined in section 2.1, provided that the data for (Z_1,Z_2) are known for the entire population. When these data are only known for the sample, the sampling design is ignorable only with respect to the conditional distribution $f(y_1,y_2 \mid z_1,z_2)$. When selection is based on the second design variable, the sampling design is informative.

For the stratified selection D2, we generated eight equal sized strata defined by the ascending values of the size variable. The sample sizes within the strata were such that they increase with increasing values of the $t_i's$.

4.3 Estimators Considered

The parameters estimated in our study are the mean vector and the $V - C$ matrix of the marginal distribution of (Y_1,Y_2). We consider seven different estimators for the design D1 and nine estimators for the design D2. See section 3.2 for description of the computations involved in the derivation of the various estimators.

DESIGN D1

$ML(Z_1,Z_2)$ – The exact MLE for the case where the design is ignorable, (equation 2.4).

$WML(Z_1,Z_2)$ – The estimators obtained from $ML(Z_1,Z_2)$ by replacing the unweighted sample statistics by probability weighted statistics (see the discussion below equation 2.7).

$ML(Z_1)$ – Same as $ML(Z_1,Z_2)$ but with Z_1 as the only design variable so that $\underset{\sim}{Z} = Z_1$.

$WML(Z_1)$ – Same as $WML(Z_1,Z_2)$ but with Z_1 as the only design variable.

CPL – The classical pseudo likelihood estimators (equations 2.7).

$WDML(X^*)$ – The (weighted distribution) estimators obtained by maximization of the likelihood in (3.6).

$WDML(X^*,Z_1)$ – The estimators obtained by maximizing the likelihood in (3.6) but with the mean and variance of Z_1 fixed at their population values.

DESIGN D2

The first 5 estimators are the same as the estimators for the design D1. The other 4 estimators are defined as follows:

$WDML(X^*)$ – The estimators obtained by maximizing the likelihood (3.10) with the α^* – coefficients [(equation (3.7)] estimated by the linear discriminant function.

$WDML(X^*,Z_1)$ – Same as $WDML(X^*)$ but with the mean and variance of Z_1 fixed at their population values.

$WDML(X^*,\underset{\sim}{t}_s)$ – The estimators obtained by maximizing the likelihood (3.10) when the values $\underset{\sim}{t}_s = (t_1, \ldots, t_n)$ are known for units in the sample.

$WDML(X^*,\underset{\sim}{t}_s,Z_1)$ – Same as $WDML(X^*,\underset{\sim}{t}_s)$ but with the mean and variance of Z_1 fixed at their population values.

It should be emphasized that the estimators derived based on the weighted distributions are not really MLE because of the approximations involved in the maximization procedures as described in section 3.2 (see also comment 2 below).

Comments

(1) The estimators we consider can be classified according to the sample and population data they use and according to whether the design variables are correctly specified and the ignorability conditions are met. Thus, the estimators $ML(Z_1,Z_2)$ and $WML(Z_1,Z_2)$ use the population values of Z_1 and Z_2 and the sample values of Y_1 and Y_2. As mentioned in section 2.4 and further discussed in Pfeffermann (1993), the use of $WML(Z_1,Z_2)$ is to protect against possible model misspecifications or informative sampling schemes. The estimators $ML(Z_1)$, $WML(Z_1)$, $WDML(X^*,Z_1)$ and $WDML(X^*,\underline{t}_s,Z_1)$ use the known population data for Z_1 but not the data for Z_2 even for the sample units. The use of these estimators corresponds to situations where the design variables are misspecified or the values of some of them are unknown. The estimator $WDML(X^*)$ uses only the sample information for Y_1, Y_2 and Z_1 and the sample selection probabilities. The estimator $WDML(X^*,\underline{t}_s)$ uses in addition the sampling values of the design variable. The estimator CPL uses only the sample values of Y_1 and Y_2 and the sample selection probabilities.

(2) We maximized the likelihood derived from the weighted distributions using a quasi-Newton method in the subroutine library IMSL. The method employed requires partial derivatives of the likelihood with respect to each of the parameters as user supplied input. An issue that arose in the maximization is worth mentioning. It is easier to parameterize the likelihood in terms of $\sum^{-1}$ where $\sum$ is the covariance matrix among Y_1, Y_2 and Z_1. Furthermore, to insure that the six parameters that define $\sum^{-1}$ are unconstrained, we use the elements of the upper triangular matrix B so that $B'B = \sum^{-1}$. Any choice of the values for B leads to a matrix $\sum^{-1}$ that is positive semi-definite.

4.4 Results

We present the results obtained when estimating $\mu_1 = E(Y_1)$, $\sigma_1^2 = \text{Var}(Y_1)$ and B_{21} - the slope coefficient in the regression of Y_2 on Y_1, as representative of the results obtained when estimating the other parameters. Tables 1-3 contain the RMSE of the various estimators as obtained for the two sampling schemes and the two choices of the design variable. RMSE's dominated by large biases are indicated by an asterisk.

The main results emerging from the tables (and from estimating the other model parameters) can be summarized as follows:

(1) The estimator $ML(Z_1,Z_2)$ outperforms all of the other estimators when the ignorability conditions are met, but it is severely biased when the sampling design is informative. The estimator $WML(Z_1,Z_2)$ is essentially unbiased in all of the cases, but the use of the sampling weights increases the variance. Still, this estimator dominates in general the estimator CPL especially under the PPS design because of the use of the population values of (Z_1,Z_2).

(2) The estimator $ML(Z_1)$ is severely biased in almost all of the cases. Notice in particular the large biases in the case where $t_i = 0.5(z_{1i} + z_{2i})$, illustrating the sensitivity of the MLE's to the exact specification of the design variables. Like with $WML(Z_1,Z_2)$, the estimator $WML(Z_1)$ is unbiased, and for the PPS design it outperforms the estimator CPL.

(3) The estimator CPL is unbiased in all of the cases. An interesting result emerging from the tables is that relative to the other estimators considered, it performs better in estimating the mean than in estimating variances and covariances. An intuitive explanation for this outcome is that in the latter case the sampling weights are used twice, thereby increasing the variance.

Table 1

RMSE of Estimators of μ_1 for Different Sampling Schemes and Design Variables
(True Mean: $\mu_1 = 40$)

Estimators	D1 - PPS Sampling		D2 - Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1,Z_2)$	0.43	1.86*	0.47	3.43*
$WML(Z_1,Z_2)$	0.43	0.57	0.50	0.52
$ML(Z_1)$	2.67*	4.38*	6.39*	8.32*
$WML(Z_1)$	0.58	0.90	0.62	0.58
$WDML(X^*,Z_1)$	0.56	0.63	1.51*	0.59
$WDML(X^*)$	0.80	0.90	3.59*	0.49
CPL	0.77	1.19	0.56	0.47
$WDML(X^*,t_s)$	–	–	0.74	0.43
$WDML(X^*,t_s,Z_1)$	–	–	0.74	0.57

* RMSE dominated by bias.

Table 2

RMSE of Estimators of σ_1^2 for Different Sampling Schemes and Design Variables
(True Variance: $\sigma_1^2 = 300$)

Estimators	D1 - PPS Sampling		D2 - Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1,Z_2)$	12.33	18.35*	16.00	29.00*
$WML(Z_1,Z_2)$	14.00	18.72	20.87	19.83
$ML(Z_1)$	24.32*	33.66*	35.16*	53.66*
$WML(Z_1)$	18.61	26.61	24.22	20.35
$WDML(X^*,Z_1)$	14.36	17.41	26.94*	15.49
$WDML(X^*)$	16.37	19.68	41.08*	15.34
CPL	21.11	29.06	24.19	20.18
$WDML(X^*,t_s)$	–	–	26.18*	15.46
$WDML(X^*,t_s,Z_1)$	–	–	25.70*	15.72

* RMSE dominated by bias.

Table 3

RMSE of Estimators of B_{21} for Different Sampling Schemes and Design Variables
(True Coefficient: $B_{21} = 1.33$)

Estimators	D1 - PPS Sampling		D2 - Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1,Z_2)$	0.043	0.069*	0.048	0.120*
$WML(Z_1,Z_2)$	0.054	0.060	0.068	0.066
$ML(Z_1)$	0.045	0.078*	0.056	0.134*
$WML(Z_1)$	0.055	0.062	0.069	0.065
$WDML(X^*,Z_1)$	0.043	0.047	0.049	0.045
$WDML(X^*)$	0.044	0.049	0.050	0.046
CPL	0.055	0.063	0.069	0.065
$WDML(X^*,t_s)$	–	–	0.048	0.045
$WDML(X^*,t_s,Z_1)$	–	–	0.048	0.045

* RMSE dominated by bias.

(4) For the PPS design, the estimators $WDML(X^*)$ and $WDML(X^*,Z_1)$ perform very well with $WDML(X^*)$ clearly dominating CPL and $WDML(X^*,Z_1)$ dominating $WML(Z_1)$. Interestingly, the estimator $WDML(X^*)$ performs in general better than the estimator $WML(Z_1)$ despite the use of less information. The fact that $WDML(X^*)$ outperforms CPL could be explained by the fact that it is more "model dependent", although as discussed in section (2.4), one way of viewing CPL is as the estimator maximizing the design unbiased estimator of the likelihood equations holding in the population.

(5) Next consider the stratified design. In the case were $t_i = 0.25x_i$, the picture is very similar to the PPS case with $WDML(X^*)$ dominating again both CPL and $WML(Z_1)$. Actually, there is little to choose in this case among the four estimators derived from the weighted distribution likelihood despite the use of different sample and population data by each estimator. When $t_i = 0.5z_i$, all of the four estimators are inferior to $WML(Z_1)$ and CPL although interestingly enough, not with respect to the estimation of the regression coefficient where they all perform very similar to the optimal $ML(Z_1,Z_2)$. The particularly poor performance of $WDML(X^*)$ (and to a much lesser extent of $WDML(X^*,Z_1)$) in estimating the mean and variance is mainly the result of incorrect specification of the strata boundaries and hence incorrect specification of the denominator of the likelihood (3.10). This problem can possibly be resolved by either including the strata boundaries and the α^* - coefficients relating the values t_i to the observed data (equation 3.7) as part of the unknown parameters in the likelihood (3.10), or by replacing the linear discriminant function by some other (nonlinear) function such as logistic regression. The latter approach has the advantage of reducing the number of parameters over which the likelihood has to be maximized, which can be crucial when the number of strata is large.

We considered so far the unconditional bias and RMSE of the estimators. As mentioned in section 4.1, we studied also conditional properties by computing the bias and RMSE's over samples with similar sample means of the design variable. The conclusions reached from that study are very similar to the conclusions stated before. Thus, estimators which are approximately unbiased unconditionally are also approximately conditionally unbiased and vice versa.

This result is somewhat surprising because it has often been illustrated in the literature that the CPL estimator, for example, has poor conditional properties. Possible explanations in our case are that the sample size considered is large or that the division of the sample into the ten groups was not sharp enough. Because of space limitations we omit the results illustrating conditional properties of the estimators.

5. CONCLUDING REMARKS

The results of the simulation study show that estimators obtained by maximizing the likelihood derived from weighted distributions are a favorable alternative to the pseudo likelihood estimators obtained by maximizing design consistent estimators of the census likelihood equations. The estimators perform particularly well in our study when using an informative sampling scheme for which the "classical" MLE can become severely biased. The use of these estimators requires, however, the modeling of the relationship between the sample selection probabilities and the observed sample data. As illustrated in the simulation study, failure to model or estimate the relationship correctly may introduce large biases.

The key question to the practical use of these estimators is therefore whether the model relating the sample selection probabilities to the observed response and design variables can be successfully identified from the sample data. It would seem that this question can only be answered by considering actual surveys that use common sampling designs. Other important questions related to the use of these estimators are the availability of reliable variance estimators so that accurate confidence intervals can be set and the protection against misspecification of the parent distribution of the response variables in the population. These two questions are common to other MLE procedures. We hope that the initial results of our study will encourage further research on these and other related questions.

ACKNOWLEDGMENT

The work of Danny Pfeffermann was supported by the Statistics Canada Research Fellowship Program.

REFERENCES

ANDERSON, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52, 200-203.

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

CHAMBERS, R.L. (1986). Design adjusted parameter estimation. *Journal of the Royal Statistical Society* A, 149, 161-173.

CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.

COX, D.R. (1969). Some sampling problems in technology. In: *New Developments in Survey Sampling*, (Eds. N. Johnson and H. Smith Jr.). New York: Wiley, 506-527.

GODAMBE, V.P., and RAJARSHI, M.B. (1989). Optimal estimation for weighted distributions: semiparametric models. In *Statistical Data Analysis and Inference*, (Ed. Y. Dodge). Amsterdam: Elsevier Science, 199-208.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.

HAUSMAN, J.A., and WISE, D.A. (1981). Stratification on endogenous variables and estimation; the Gary Income Maintenance Experiment. In *Structure Analysis of Discrete Data with Econometric Applications*, (Eds. C.F. Mansky and D. McFadden). Cambridge, Mass.: MIT Press, 366-391.

HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society* A, 143, 474-487.

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

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

NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society* B, 42, 377-386.

PATIL, G.P., and RAO, C.R. (1978). Weighted distributions and size biased sampling with application to wildlife populations and human families. *Biometrics*, 34, 179-189.

PFEFFERMANN, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association*, 83, 824-833.

PFEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review* (Forthcoming).

RAO, C.R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions*, (Ed. G.P. Patil). Calcutta: Statistical Publishing Society, 320-332.

RAO, C.R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? In *A Celebration in Statistics* (Eds. A.C. Atkinson and S.E. Fienberg). New York: Springer-Verlag, 543-569.

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

RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 469-474.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.

RUBIN, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics* 2, (Eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith). Amsterdam: Elsevier Science, 463-472.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.

SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10, 616-620.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 241–252
Statistics Canada

Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used

CARL-ERIK SÄRNDAL[1]

ABSTRACT

In almost all large surveys, some form of imputation is used. This paper develops a method for variance estimation when single (as opposed to multiple) imputation is used to create a completed data set. Imputation will never reproduce the true values (except in truly exceptional cases). The total error of the survey estimate is viewed in this paper as the sum of sampling error and imputation error. Consequently, an overall variance is derived as the sum of a sampling variance and an imputation variance. The principal theme is the estimation of these two components, using the data after imputation, that is, the actually observed values and the imputed values. The approach is model assisted in the sense that the model implied by the imputation method and the randomization distribution used for sample selection will together determine the appearance of the variance estimators. The theoretical findings are confirmed by a Monte Carlo simulation.

KEY WORDS: Single value imputation; Variance estimation; Imputation model; Model assisted inference.

1. DIFFERENT TYPES OF IMPUTATION

This paper reports work carried out in connection with the development of Statistics Canada's Generalized Estimation System (GES). Variance estimates are to be routinely calculated in the different estimation modules that define the GES. There was a need to develop suitable methods for variance estimation when the data set contains imputed values, which is the case in practically all surveys.

Two principal approaches to estimation with missing data are weighting and imputation. In the recent literature, the weights used to compensate for nonresponse are usually viewed as the inverse of the response probabilities associated with an assumed response mechanism. Since the response probabilities are ordinarily unknown, they need to be estimated from the available data. Imputation, on the other hand, has the advantage that it yields a complete data matrix. Such a matrix simplifies data handling, but it does not imply that "standard estimation methods" can be used directly. The imputed values are sample-based, thus they have their own statistical properties, such as a mean and a variance.

In our age, imputation is an extensively used tool. It is interesting to note what Pritzker, Ogus and Hansen (1965) say about imputation policy at the US Bureau of the Census: "Basically our philosophy in connection with the problem of ... imputation is that we should get information by direct measurement on a very high proportion of the aggregates to be tabulated, with sufficient control on quality that almost any reasonable rule for ... imputation will yield substantially the same results ... With respect to imputation in censuses and sample surveys we have adopted a standard that says we have a low level of imputation, of the order of 1 or 2 percent, as a goal."

[1] Carl-Erik Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec) H3C 3J7.

Ideally, we should still strive for the goal of only one to two percent imputation. But in our time most surveys carried out by large survey organizations show a rate of imputation that is much higher. Clearly, if 30% of the values are imputed, the effects of imputation can not be ignored. Imputation can create systematic error (bias) in the point estimate; this is perhaps the most serious concern. But even if an imputation method can be found such that there is no appreciable systematic error, one must not ignore the often considerable effect that imputation has on the precision (the variance) of the point estimate. There is a need for simple yet valid variance estimation methods for survey data containing imputations, so that the coefficients of variation of the survey estimates can be properly reported.

A variety of imputation methods have been proposed. These can be classified in different ways. One way to classify is by the number of imputations carried out. In **single imputation** methods, a single value is imputed for a missing value. A complete data matrix is obtained, in which the imputed values are flagged. Estimates are calculated with the aid of the completed set. In **multiple imputation**, two or more values are imputed for each missing value. Several completed data sets are thus obtained. Estimates are calculated with the aid of the completed data sets.

Imputation methods also differ with respect to the modeling underlying the imputation. Some imputation methods use an **explicit** model, as when the imputed value is obtained by a regression fit, a ratio or mean imputation. In other methods, the model is only **implicit**, as in hot deck imputation and nearest neighbour donor imputation. The distinctions just made are important for this paper.

Statistics Canada currently uses imputation methods such as nearest neighbour donor, current ratio, current mean, previous value, previous mean, auxiliary trend. All of these are single imputation methods. The imputed values originate in the Generalized Edit and Imputation System (GEIS), from where they enter into the Generalized Estimation System (GES), where the point estimates and the variance estimates are calculated in a number of different estimation modules. This paper deals in particular with current ratio imputation, which represents a case of explicit modeling.

2. SOME THOUGHTS ON MULTIPLE IMPUTATION

Multiple imputation was suggested by D.B. Rubin around 1977. His ideas are explained in a number of papers, of which Herzog and Rubin (1983) and Rubin (1986) are expository, and in a book, Rubin (1987). Multiple imputation has advantages as well as disadvantages; the same is true for single imputation.

Rubin (1986) sees as a disadvantage of single imputation that "... the one imputed value cannot in itself represent uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reason for nonresponse are known."

Multiple imputation is attractive because it communicates the idea that imputation has variability. It is precisely this variability – the variability within and between the several completed data sets – that is exploited in the variance estimation methods proposed under multiple imputation. These methods make powerful use of basic statistical concepts. (On the other hand, one can argue that sample selection also has variability, but most surveys cannot afford more than a single sample, and estimation must be carried out with this unique sample.)

Simple examples show that treating imputed values just like observed values can lead to severe underestimation of the true uncertainty; survey samplers have long been aware of this. And

it is a fact that users sometimes treat imputed values just like observed values, with wrong statement of precision as a result. With modern computers, it is easy to impute by some rule or another, but not so easy to obtain valid variance estimates.

The citation above seems to conclude that because a single imputed value does not display variation, we cannot obtain reasonable variance estimates; we are necessarily led to underestimation. I do not share this opinion. The methods that I discuss show that valid variance estimation is indeed possible with single imputation.

A method for variance estimation in the presence of imputed values should have the following properties: (a) a sound theoretical backing; (b) robustness to the assumptions underlying the imputation; (c) it must be practical, easy to carry out, and readily accepted by users.

While multiple imputation has the ingredients (a) and (b), it is clear that, in some applications at least, it does not have the property (c). In the development of the GES we must depend on procedures that are easy to administer and easy to accept by the user. The user of a data set (someone who is not primarily a statistician) can easily understand that the statistician imputes once, with the objective to fill in the best possible value for one that is missing. While it is true that for some purposes, such as secondary analyses, it might be interesting to have several completed data matrices, the costs of storage of multiple data sets will often rule out this option.

Multiple imputation may well be useful in other contexts and for other reasons than those that are essential to the development of the GES. The multiple imputation method has indicated one way of handling the problem of understatement of the variance, at least for some situations. The method has recently come under criticism by Fay (1991) and is not the only answer. Let us see what can be done with single imputation methods. The method described below is based on Särndal (1990).

3. IMPUTATION VARIANCE AND SAMPLING VARIANCE

An imputation rule corresponds to an (explicit or implicit) model for the relationship among variables of interest to the survey. That is, when the analyst has fixed an imputation rule, he or she has in fact chosen a model. The principle for the developments that follow is that if this rule is considered good enough for the point estimates (no systematic error), the rule is also good enough for the corresponding estimates of variance. In other words, the model maker should take responsibility for control of the bias as well as for the appropriateness of the variance estimate.

Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population; let y denote one of the study variables in the survey. The objective is to estimate the population total of y, $t = \sum_U y_k$. (If C is any set of population units, where $C \subseteq U$, $\sum_C$ is used as shorthand for $\sum_{k \in C}$, for example, $t = \sum_U y_k$ means $\sum_{k \in U} y_k$.) A probability samples s is selected with a given sampling design. The inclusion probabilities are known, and ordinary design-based variance estimates would be obtained if all units $k \in s$ are observed. However, there are missing data. Let r be the subset s for which the values y_k are actually observed. For the complement, $s - r$, imputations are calculated. The **data after imputation** consist of the values denoted $y_{\bullet k}$, $k \in s$, such that

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ y_{\text{imp},k} & \text{if } k \in s-r \end{cases},$$

where y_k is an actually observed value, and $y_{\text{imp},k}$ denotes the imputed value for the unit k. The case $r = s$ implies no imputation; all data are actual observations.

Let us write the estimator of t that would be used in the case of 100% response (that is, $r = s$) as $\hat{t} = \sum_{k\in s} w_k y_k = \sum_s w_k y_k$, where w_k is the weight given to the observation y_k. For example, in simple random sampling without replacement (SRSWOR) of n units from N, $w_k = N/n$ for all $k\in s$ when the expanded sample mean is used to estimate t, and $w_k = (\bar{z}_U/\bar{z}_s)(N/n) = (\sum_U z_k)/(\sum_s z_k)$ for all $k\in s$ when the ratio estimator is used with z as an auxiliary variable.

When the data contain imputations, the estimator of t is $\hat{t}_\blacklozenge = \sum_s w_k y_{\blacklozenge k}$. That is, we assume that the weights w_k are identical to those used when all data are actual observations. This principle is used in the estimation modules of the GES. It embodies an assumption that imputation by the chosen rule causes little or no systematic error in the estimates.

The variance of an estimated total is increased by imputation, because imputation does not (except in truly exceptional circumstances) reproduce the true value y_k. Concrete evidence of this is the fact that if the imputation rule is applied to the actually observed sample units, there will always be error. If the rule is not without error for the responding units, it is not without error for the nonresponding units either. In Section 4 we express the variance of $\hat{t}_\blacklozenge$ as a sum of two components, a sampling variance, and a variance due to imputation,

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}.$$

The imputation variance V_{imp} is zero if all data are actually observed values, or if the imputation procedure is capable of exactly reproducing the true value y_k for every unit requiring imputation. (Neither case is likely in practice.) The procedure given in Section 4 uses the data after imputation, $y_{\blacklozenge k}$, $k\in s$, to obtain estimates of each of the two components, leading to

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}.$$

The component $\hat{V}_{\text{sam}}$ is calculated in two steps:

(1) Compute the standard design-based variance estimate using the data after imputation. (For example, if SRSWOR is used, and $r = s$, the standard unbiased variance estimate of $N\bar{y}_s$ is $N^2(1/n - 1/N)\sum_s(y_k - \bar{y}_s)^2/(n-1)$. This formula, calculated on the data after imputation, yields $N^2(1/n - 1/N)\sum_s(y_{\blacklozenge k} - \bar{y}_{\blacklozenge s})^2/(n-1)$, where $\bar{y}_{\blacklozenge s}$ is the mean of the n values $y_{\blacklozenge k}$.)

(2) Add a term to correct for the fact that many imputation rules give data with "less than natural" variability, which would lead to understatement of the sampling variance unless corrective action is taken. Finally, the component $\hat{V}_{\text{imp}}$ is readily computed from the data after imputation. The user will easily accept the argument that the variance obtained by the standard formula is not sufficient in itself; something must be added because the imputation rule is less than perfect.

The method has the good property that if no imputation is required, that is, $r = s$, then $\hat{V}_{\text{imp}} = 0$ and $\hat{V}_{\text{sam}}$ equals the "standard variance estimator" that one would have used with 100% actually observed values.

4. THEORETICAL DEVELOPMENTS

The total error of $\hat{t}_\blacklozenge$ is decomposed as

$$\hat{t}_\blacklozenge - t = (\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t}) = \text{sampling error} + \text{imputation error}.$$

The imputation error is the difference between the unknown estimate that would have been calculated if the data had consisted entirely of actual observations and the estimate that can be calculated on the data after imputation. The imputation error is

$$\hat{t}_\blacklozenge - \hat{t} = -\sum_{s-r} w_k e_k,$$

where

$$e_k = y_k - y_{\text{imp},k}$$

is an **imputation residual** which can not be observed for a unit $k \in s-r$. The magnitude of e_k depends on how well the imputation model fits. The residuals are small if the imputation method gives nearly perfect substitute values. To pursue the argument, different directions may be taken. Here, we use a **model assisted** approach in which three different probability distributions are considered. The corresponding expectation symbols are written as E_ξ, E_s, and E_r. Here, ξ indicates "with respect to the imputation model"; s indicates "with respect to the sampling design", and r indicates "with respect to the response mechanism, given s". The model is implied by the imputation rule, so it is known; the sampling design is the given probability sampling distribution, so it is also known; the response mechanism is an ordinarily unknown distribution governing the response, given the sample s.

The estimator $\hat{t}_\blacklozenge$ is overall unbiased in the sense that $E_\xi E_s E_r(\hat{t}_\blacklozenge - t) = 0$ if two conditions hold:

(a) the order of the expectation operators can be changed so that $E_\xi E_s E_r(\cdot)$ can be evaluated as $E_s E_r\{E_\xi(\cdot \mid s, r)\}$, and

(b) the imputation residual $e_k = y_k - y_{\text{imp},k}$ has zero model expectation for every $k \in r$, that is, $E_\xi(e_k) = 0$, which implies that $E_\xi(\hat{t}_\blacklozenge - \hat{t}) = 0$.

Condition (a) is satisfied if the response mechanism is one that may depend on s and on auxiliary data, but not on the y-values, y_k, $k \in s$. That is, the probability $q(r)$ of realizing the response set r is of the form $q(r) = q(r \mid s, \{x_k : k \in s\})$, where $\{x_k : k \in s\}$ denote the auxiliary data. The response mechanism can then be said to be ignorable.

We now examine the overall variance given by

$$V_{\text{tot}} = E_\xi E_s E_r\{(\hat{t}_\blacklozenge - t)^2\},$$

which may also be called the anticipated variance under the imputation model ξ. We obtain

$$\begin{aligned} V_{\text{tot}} &= E_{\xi sr}(\hat{t}_\blacklozenge) = E_\xi E_s E_r\{(\hat{t}_\blacklozenge - t)^2\} \\ &= E_\xi E_s E_r\{(\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t})\}^2 \\ &= E_\xi V_p + E_s E_r V_{\xi c}, \end{aligned} \tag{4.1}$$

where $V_p = E_s\{(\hat{t} - t)\}^2$ is the design-based variance of $\hat{t}$, supposing $\hat{t}$ is design unbiased for the total t. (For an estimator with some slight design bias, V_p is the design-based mean square error of $\hat{t}$.) Note that $(\hat{t} - t)$ depends on s only, and not on r. Moreover,

$$V_{\xi c} = E_\xi\{(\hat{t}_\blacklozenge - \hat{t})^2 \mid s, r\}$$

is the model variance of the imputation error, conditionally on s and r. The subscript c stands for "conditional". The derivation of (4.1) assumes that condition (a) holds so that the expectation E_ξ can be moved inside $E_s E_r$, and that the mixed term

$$2E_\xi E_s[(\hat{t} - t)\{E_r(\hat{t}_\blacklozenge - \hat{t}) \mid s\}] \tag{4.2}$$

vanishes or is sufficiently close to zero that we can ignore it. This would be the case if the expected imputation error is zero or negligible under the response mechanism, conditionally on the realized probability sample s. Even if (4.2) is not exactly zero for the mechanism that determines the response, we can in many cases approximate (4.2) by zero and still use the method below to obtain a variance estimate that is much better than pretending naively that imputed data are as good as actually observed data. For ratio imputation and SRSWOR, which is an application considered in Section 5, the term (4.2) is exactly zero.

If we denote $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_s E_r V_{\xi c}$ in (4.1), then

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$$

or

$$\text{overall variance} = \text{sampling variance} + \text{imputation variance}.$$

The objective is to estimate the overall variance, so that a valid confidence interval for the unknown t can be calculated. Our approach is to obtain separate estimates, $\hat{V}_{\text{sam}}$ and $\hat{V}_{\text{imp}}$, of the two components $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_s E_r V_{\xi c}$. The data available for this estimation are $y_{\blacklozenge k}$, $k \in s$. The argument for obtaining $\hat{V}_{\text{sam}}$ and $\hat{V}_{\text{imp}}$ is as follows:

(i) Estimation of the sampling variance component. Let $\hat{V}_p$ be the standard (design-unbiased or nearly design-unbiased) estimator of the design variance V_s. Denote by $\hat{V}_{\blacklozenge p}$ the quantity obtained by calculating $\hat{V}_p$ from the data after imputation, $y_{\blacklozenge k}$, $k \in s$. For many imputation rules, $\hat{V}_{\blacklozenge p}$ underestimates V_{sam}. The underestimation is compensated in the following way. Evaluate the conditional expectation

$$E_\xi(\hat{V}_p - \hat{V}_{\blacklozenge p} \mid s, r) = V_{\text{dif}}.$$

Then for given s and r, find a model unbiased estimator, denoted $\hat{V}_{\text{dif}}$, of V_{dif}. This will usually require the estimation of certain parameters of the model ξ. Consequently,

$$E_\xi(\hat{V}_{\text{dif}} \mid s, r) = E_\xi(\hat{V}_p - \hat{V}_{\blacklozenge p} \mid s, r).$$

Then

$$\hat{V}_{\text{sam}} = \hat{V}_{\blacklozenge p} + \hat{V}_{\text{dif}}$$

is overall unbiased for the component $V_{\text{sam}} = E_\xi V_p$, as the following derivation shows:

$$\begin{aligned} E_\xi E_s E_r(\hat{V}_{\text{sam}}) &= E_s E_r\{E_\xi(\hat{V}_{\blacklozenge p}) + E_\xi(\hat{V}_{\text{dif}})\} \\ &= E_s E_r\{E_\xi(\hat{V}_p)\} = E_\xi E_s(\hat{V}_p) \\ &= E_\xi V_p = V_{\text{sam}}. \end{aligned}$$

(ii) Estimation of the imputation variance component. Simply find an estimator, $\hat{V}_{\xi c}$, that is model unbiased for $V_{\xi c}$. That is, $E_\xi(\hat{V}_{\xi c}) = V_{\xi c}$. Again, this may require the estimation of unknown parameters of the model ξ. Then $\hat{V}_{\xi c}$ is overall unbiased for the imputation variance component V_{imp}, since

$$E_s E_r E_\xi(\hat{V}_{\xi c}) = E_s E_r V_{\xi c} = V_{\text{imp}}.$$

Finally, an overall unbiased estimator of V_{tot} is given by

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}},$$

where $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}}$ and $\hat{V}_{\text{imp}} = \hat{V}_{\xi c}$. Note that the role of $\hat{V}_{\text{dif}}$ is to correct for the fact that the data after imputation may display "less than natural" variation. This often happens when $y_{\text{imp},k}$ equals the predicted value from a fitted regression, that is, "the value on the line". The variation around the line is not reflected in the predicted value.

To be overall unbiased, the estimator $\hat{V}_{\text{tot}}$ constructed above requires that condition (a) holds, that (4.2) is zero, and that the imputation model is correct, so that $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$ are model unbiased for V_{dif} and $V_{\xi c}$, respectively. Mild departures from the assumed imputation model may not have serious consequences, but if the imputation model is grossly misspecified it is clear that $\hat{V}_{\text{tot}}$ may be considerably biased because of the model bias of $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$. Monte Carlo simulations reported in Lee, Rancourt and Särndal (1992) show that the variance estimator $\hat{V}_{\text{tot}}$ is fairly robust to imputation model breakdown. To add the terms $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$ is in any case a vast improvement on simply using the naive uncorrected variance estimator $\hat{V}_{\bullet p}$.

Note that if the imputation model holds, an unbiased variance estimate is obtained with the method even if the response probabilities differ among units, as long as they depend on the x_k-values only. That is, we can allow a systematic response pattern such that large x_k-value units are less likely to respond than small x_k-value units. If the response probabilities depend explicitly the y_k-values, then the situation is different; the response mechanism is nonignorable and condition (a) does not hold. There will now be bias in $\hat{V}_{\text{tot}}$ due to nonignorability; the simulations in Lee, Rancourt and Särndal (1992) throw some light on the magnitude of this bias.

Example. The sample s is drawn with SRSWOR; n units from N. Let m denote the size of the response set r. Suppose the respondent mean is imputed for units requiring imputation. The corresponding imputation model ξ states that $y_k = \beta + \epsilon_k$, where the ϵ_k are uncorrelated errors terms with $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2$. That is, $y_{\bullet k} = y_k$ if $k \in r$ and $y_{\bullet k} = \hat{\beta} = \bar{y}_r$ if $k \in s - r$, and we obtain the estimator $\hat{t}_\bullet = (N/n)\sum_s y_{\bullet k} = N\bar{y}_r$. Here the standard design-based variance estimator for 100% response is $\hat{V}_p = N^2(1/n - 1/N)\sum_s(y_k - \bar{y}_s)^2/(n-1)$; when this formula is computed on data after imputation we get $\hat{V}_{\bullet p} = N^2(1/n - 1/N)\{(m-1)/(n-1)\}S^2_{yr}$, where $S^2_{yr} = \sum_r(y_k - \bar{y}_r)^2/(m-1)$. Other derivations give $\hat{V}_{\text{dif}} = N^2(1/n - 1/N)\{(n-m)/(n-1)\}S^2_{yr}$ and $\hat{V}_{\text{imp}} = N^2(1/m - 1/n)S^2_{yr}$. Thus, $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}} = N^2(1/n - 1/N)S^2_{yr}$, and $\hat{V}_{\text{tot}} = N^2(1/m - 1/N)S^2_{yr}$, which is easy to accept as a "good" variance estimator for this simple imputation rule. The following table shows the contribution of each of the three terms to the total variance estimator $\hat{V}_{\text{tot}}$, for different rates of imputation, assuming that N is large compared to m and n, and $(m-1)/m \approx (n-1)/n \approx 1$.

Imputation rate in %	% contribution to $\hat{V}_{tot}$		
$100\ (1 - m/n)$	$\hat{V}_{\bullet p}$	$\hat{V}_{dif}$	$\hat{V}_{imp}$
10	81	9	10
20	64	16	20
30	49	21	30

The table illustrates the dangers of acting as if imputations are real data: with 30% imputed values, the standard formula variance estimator $\hat{V}_{\bullet p}$ in this example covers less than half of the correctly estimated total variance. Imputation by the respondent mean is useful as an example; the results are particularly simple. But usually in practice, respondent mean imputation is neither justified nor efficient. The underlying model is not sophisticated enough to avoid systematic error in the point estimates, and the residuals $e_k = y_k - \bar{y}_r$ can vary considerably.

5. APPLICATION TO IMPUTATION BY THE CURRENT RATIO METHOD

The method assumes that a positive auxiliary value x_k is known for every unit $k \in s$. If $k \in s-r$, we impute $y_{imp,k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$. The data after imputation are

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{B}x_k & \text{if } k \in s-r. \end{cases}$$

The model behind current ratio imputation is

$$y_k = \beta x_k + \epsilon_k, \tag{5.1}$$

where the ϵ_k are uncorrelated model errors such that

$$E_\xi(\epsilon_k) = 0, \quad V_\xi(\epsilon_k) = \sigma^2 x_k. \tag{5.2}$$

Suppose that the sample s is selected by SRSWOR. Let the respective sizes of s, r, and $s - r$ be n, m, and $n - m$. If no imputation was needed, the estimator of $t = \sum_U y_k$ would be $\hat{t} = N\bar{y}_s$. Using the data after imputation, we get

$$\hat{t}_\bullet = (N/n) \sum_s y_{\bullet k} = N\bar{x}_s\bar{y}_r/\bar{x}_r. \tag{5.3}$$

(Overbar and subscript s, r, or $s - r$ indicates "straight mean", for example, $\bar{y}_r = \sum_r y_k/m$, $\bar{x}_{s-r} = \sum_{s-r} x_k/(n - m)$, *etc.*) Using the results of the preceding section, we have $V_{tot} = V_{sam} + V_{imp}$ with $V_{sam} = E_\xi\{N^2(1/n - 1/N)S_{yU}^2\}$ and $V_{imp} = E_s E_r\{N^2(1/m - 1/n)C_1\sigma^2\}$, where $S_{yU}^2 = \sum_U (y_k - \bar{y}_U)^2/(N - 1)$ and $C_1 = \bar{x}_s\bar{x}_{s-r}/\bar{x}_r$, a known constant. The mixed term (4.2) is exactly zero in this case. Our method of variance estimation gives $\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp}$, where

$$\hat{V}_{sam} = N^2(1/n - 1/N)\{S_{y\bullet s}^2 + C_0\hat{\sigma}^2\}, \tag{5.4}$$

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)C_1\hat{\sigma}^2, \tag{5.5}$$

where $S^2_{y\bullet s} = \sum_s(y_{\bullet k} - \bar{y}_{\bullet s})^2/(n-1)$ is the variance calculated on data after imputation, and we have chosen to estimate σ^2 by the model unbiased formula

$$\sigma^2 = \frac{1}{\bar{x}_r\{1 - (1/m)(cv_{xr})^2\}} \frac{\sum_r(y_k - \hat{B}x_k)^2}{m-1},$$

where $cv_{xr} = S_{xr}/\bar{x}_r$ is the coefficient of variation of x in the response set r. The constant C_0 is obtained as

$$C_0 = \frac{1}{\sigma^2} E_\xi(S^2_{ys} - S^2_{y\bullet s}),$$

where

$$S^2_{ys} = \frac{1}{n-1}\sum_s (y_k - \bar{y}_s)^2$$

is the (unknown) sample variance based on data with 100% actual observations. After evaluation,

$$C_0 = \frac{1}{n-1}\left\{\sum_{s-r} x_k - \frac{\sum_{s-r}x_k^2}{\sum_r x_k} + \frac{1}{n}\frac{\sum_{s-r}x_k \sum_s x_k}{\sum_r x_k}\right\}.$$

If m is not too small, the approximations $\hat{\sigma}^2 \approx (\sum_r e_k^2)/(\sum_r x_k)$ with $e_k = y_k - \hat{B}x_k$ and $C_0 \approx (1 - m/n)\bar{x}_{s-r}$ are sufficiently good for most applications.

We can write the imputation variance component as

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)A\bar{x}_s\hat{\sigma}^2,$$

where $A = \bar{x}_{s-r}/\bar{x}_r$. The constant A reflects the selection effect due to nonresponse. If large units are less inclined to respond than small units, then A may be considerably greater than unity, and, for a given a sample s and a given number m of respondents, the component $\hat{V}_{\text{imp}}$ tends to be large, relative to a case where, say, all units are equally likely to respond. This tendency makes good sense intuitively.

Two special cases are noted: (1) If all $x_k = 1$, the estimated total variance becomes simply

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}} = N^2(1/m - 1/N)S^2_{yr},$$

where S^2_{yr} is the variance of the m actual observations y_k. This agrees with the variance obtained under a two-phase sampling design with SRSWOR in each phase. (2) If no imputation is required, that is, if $s = r$, then $\hat{V}_{\text{imp}} = 0$, and

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} = N^2(1/n - 1/N)S^2_{ys}.$$

That is, our method yields the well known variance estimator for SRSWOR.

A Monte Carlo study with 100,000 repeated response sets r was carried out to confirm the above results for current ratio imputation. A finite population of size $N = 100$ was generated according to the model consisting of (5.1) and (5.2). The typical response set r was obtained

as follows: Draw a SRSWOR sample s of size $n = 30$; given s, generate r by a response mechanism in the form of independent Bernoulli trials, one for each $k \in s$, with probability θ_k for the outcome "response". Three different response mechanisms were used: Mechanism 1: θ_k increases with y_k in such a way that $\theta_k = 1 - \exp(-a_1 y_k)$; Mechanism 2: θ_k increases as y_k decreases in such a way that $\theta_k = \exp(-a_2 y_k)$; Mechanism 3: θ_k is constant at 0.7, that is, a uniform response mechanism. The constants a_1 and a_2 in the first two response mechanisms (which can be described as non-ignorable) were fixed to obtain an average response probability of 0.7. The sizes of the realized response sets r thus varied around a mean of 21 for all three mechanisms. For each r, the point estimate $\hat{t}_\blacklozenge$ given by (5.3) was calculated as well as three different variance estimators, $\hat{V} = \hat{V}(\hat{t}_\blacklozenge)$. These were: (1) the **model assisted** variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ equal to the total of (5.4) and (5.5); (2) the **two-phase** sampling variance estimator $N^2(1/n - 1/N)S^2_{yr} + N^2(1/m - 1/n)\sum_r e_k^2/(m-1)$, an estimator which follows from standard two-phase sampling theory with an assumption of SRSWOR subsampling of m respondents from the n units in the initial sample (Rao 1990); and (3) the **standard unadjusted** variance estimator $N^2(1/n - 1/N)S^2_{y\blacklozenge s}$ obtained by acting as if imputations are as good as actual data. The results are shown in the following table.

Estimator $\hat{V}$	Relative bias of $\hat{V}$ in %		
	Mechanism 1	Mechanism 2	Mechanism 3
Model assisted	−0.20	−4.64	−3.99
Two-phase	9.95	−12.49	−1.11
Standard unadjusted	−25.73	−37.90	−33.21

The relative bias of an estimator $\hat{V}$ was calculated as $\{\text{mean}(\hat{V}) - \text{var}(\hat{t}_\blacklozenge)\}/\text{var}(\hat{t}_\blacklozenge)$, where mean $(\hat{V})$ is the mean of the 100,000 values of $\hat{V}$, and $\text{var}(\hat{t}_\blacklozenge)$ is the variance of the 100,000 values of $\hat{t}_\blacklozenge$. The simulation shows that the model assisted variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ is nearly unbiased for all three response mechanisms. In a way, this is not surprising because the population was generated to agree with the ratio imputation model. Mechanisms 1 and 2 are of the nonignorable kind and do not verify condition (a) of Section 4 required for unbiasedness of $\hat{V}_{\text{tot}}$. Interestingly, though, in this example the bias of $\hat{V}_{\text{tot}}$ remains small despite this. The two-phase estimator works well for the uniform response mechanism 3, the case for which it was conceived; otherwise it is biased. Finally, to act as if imputed data are as good as actual data leads, as expected, to a dramatic understatement of the true variance for all three mechanisms. A more extensive Monte Carlo study of ratio estimation is reported in Lee, Rancourt and Särndal (1992). This paper gives an idea of the effect of imputation model misspecification, which is also discussed in Rao (1992).

6. IMPUTED VALUES THAT HAVE AN ADDED RESIDUAL

We can distinguish two types of imputed values: (1) the imputed value $y_{\text{imp},k}$ consists of a predicted value only, $y_{\text{pred},k}$, as when the value on a fitted regression line or surface is used. For example in the current ratio imputation method as used above, $y_{\text{imp},k} = y_{\text{pred},k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$; (2) the imputed value $y_{\text{imp},k}$ consists of a predicted value and a

residual, so that $y_{\text{imp},k} = y_{\text{pred},k} + e_k^*$. The residual term, whose purpose is to make imputed values more like actual observations, may be obtained by sampling the residuals $e_k = y_k - y_{\text{pred},k}$ calculated for the responding units $k \in r$. A scheme for this is given below. This type of imputation is sometimes recommended in the literature as a means of preserving the distributions of the imputed data; see, for example, the discussion in Little (1988). The imputation process then requires more effort to complete, and for the purposes of the GES (whose principal aim is valid estimation of the precision of survey estimates), it is not clear that the advantages gained are worth the extra effort.

Let us, however, indicate one scheme for imputation by "predicted value plus residual" in the case where the current ratio imputation model is taken as the point of departure: For $k \in r$, calculate $e_k = y_k - \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$, then $\tilde{e}_k = e_k/\sqrt{x_k}$. This gives a supply of m "standardized residuals" $\tilde{e}_k$. Then for a unit $k \in s-r$, calculate $e_k^0 = \sqrt{x_k}\tilde{e}_k$, where $\tilde{e}_k$ is drawn by SRSWR from the supply, and x_k belongs to the unit requiring imputation. Then large x-value units tend to obtain larger residuals e_k^0, which is consistent with the model. Then set $e_k^* = e_k^0 - (\sum_{s-r} e_k^0)/(n-m)$. For $k \in s-r$, impute $y_{\text{imp},k} = \hat{B}x_k + e_k^*$, $k \in s-r$; for $k \in r$, we have actual observations, y_k. Since the e_k^* were made to sum to zero over $s-r$, the point estimator is given by $\hat{t}_\blacklozenge = (N/n)\sum_s y_{\blacklozenge k} = N\bar{x}_s\bar{y}_r/\bar{x}_r$ as in Section 5, but its variance is different. It can be shown that $E_\xi E_s E_r E_\# (S^2_{y\blacklozenge s} - S^2_{ys}) \approx 0$, where $E_\#$ denotes average with respect to the random selection of a standardized residual. That is, the difference between the variance calculated on data after imputation, $S^2_{y\blacklozenge s}$, and the unknown variance of a sample consisting entirely of actual observations, S^2_{ys}, is approximately zero on the average. We can use $\hat{V}_{\text{sam}} = N^2(1/n - 1/N)S^2_{y\blacklozenge s}$ as an approximately overall unbiased estimator of the sampling variance component. There is no need now to add a correction $\hat{V}_{\text{dif}}$. However, an estimator of the imputation variance $V_{\text{imp}} = N^2(1/m - 1/n)C_1\sigma^2$ must still be calculated and added to $\hat{V}_{\text{sam}}$.

7. CONCLUDING REMARKS

The continued work on the variance estimation techniques outlined in this paper has the following objectives: (1) extensions to imputation procedures based on models that are implicit only, in particular the nearest neighbour donor method; (2) extensions to the case where there is a mixture of several imputation procedures in the same survey.

Deville and Särndal (1992) present results for an extension in which the Horwitz-Thompson estimator, $\hat{t} = \sum_s y_k/\pi_k$, serves as the prototype. The estimator using data after imputation is then

$$\hat{t}_\blacklozenge = \sum_r y_k/\pi_k + \left(\sum_{s-r} x_k/\pi_k\right)'\hat{B} = \sum_s y_k/\pi_k - \sum_{s-r} e_k/\pi_k,$$

where $e_k = y_k - x_k'B$ is the imputation residual for unit k obtained by multiple regression.

ACKNOWLEDGEMENTS

I am indebted to M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee, and G. Reinhardt of Statistics Canada for their collaboration in the work that led to this paper. The comments of two referees led to improvements in the original manuscript and are gratefully acknowledged.

REFERENCES

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Variance estimation for survey data with regression imputation. Technical report.

HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in surveys. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 209-245.

FAY, R.E. (1991). A design-based perspective on missing data variance. Proceedings, 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1992). Experiments with variance estimation from survey data with imputed values. Report, Business Survey Methods Division, Statistics Canada, submitted for publication.

LITTLE, R.J.A. (1988). Missing-data adjustments in large surveys (with discussion). *Journal of Business and Economic Statistics*, 6, 287-301.

PRITZKER, L., OGUS, J., and HANSEN, M.H. (1965). Computer editing methods: some applications and results. *Bulletin of the International Statistical Institute*, 41, 442-466.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Manuscript seen by courtesy of the author.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Manuscript seen by courtesy of the author.

RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.

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

SÄRNDAL, C.-E. (1990). Estimation of precision in the generalized estimation system when imputation is used. Report, Informatics and Methodology Field, Statistics Canada, March 31, 1990.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 253–262
Statistics Canada

A Sample Allocation Method for Two-Phase Survey Designs

J.B. ARMSTRONG and C.F.J. WU[1]

ABSTRACT

Motivated by a business survey design at Statistics Canada, we formulate the problem of sample allocation for a general two-phase survey design as a constrained nonlinear programming problem. By exploiting its mathematical structure, we propose a solution method that consists of iterations between two subproblems that are computationally much simpler. Using an approximate solution as a starting value, the proposed method works very well in an empirical study.

KEY WORDS: Optimal allocation; Convex programming.

1. INTRODUCTION

The purpose of this paper is to propose a method of sample allocation for two-phase survey designs. Suppose it is necessary to stratify a population of size N into L strata according to an auxiliary variable, z, whose information is not known before sampling. Values of a second auxiliary (size) variable, x, that is correlated with the variable of interest, y, are known for all units in the population. At the first phase of sampling, the population is divided into G strata according to x. An initial sample is drawn from size stratum $g (g = 1, 2, \ldots, G)$, using simple random sampling with sampling fraction v_g, and the z-value for each sampled unit is observed. At the second phase, units in the sample from size stratum g with z-value in class $h (h = 1, 2, \ldots, L)$, are subsampled using sampling fraction v_{gh}. The value of y is observed for units in the second-phase sample.

In the case of no size stratification ($G = 1$) Cochran (1977) gives the allocation that minimizes the variance of the estimate $\hat{Y} = \sum_h \sum_{i \in s2 \cap h} y_i/(v \cdot v_h)$ of the population total $Y = \sum_h N_h \cdot \bar{Y}_h$, subject to a fixed survey cost, C, where N_h and $\bar{Y}_h$ are the population size and population mean, respectively, for stratum h and $\sum_{i \in s2 \cap h} y_i$ denotes the sum of y-values for units in the second phase sample, $s2$, with z-value in class h. If survey estimates are used for analytical purposes, the variance of the estimated total for z class h, $\hat{Y}_h = \sum_{i \in s2 \cap h} y_i/(v \cdot v_h)$, is also of interest. Sedransk (1965), Booth and Sedransk (1969), Rao (1973) and Smith (1989) have studied allocation problems involving the minimization of a function of variances of estimated class totals, subject to a cost constraint.

The method described in this paper can be used to solve the allocation problem for general G when there is a constraint on the variance of the estimated total for each z class. The method was motivated by an application in a business survey conducted by Statistics Canada. The survey involves the sampling of tax records for businesses.

Information about the population of taxfilers is made available to Statistics Canada by Revenue Canada. There is a requirement to produce estimates of financial variables for domains defined by a cross-classification of four-digit Standard Industrial Classification (SIC4) and province. Only two digits of SIC are coded by Revenue Canada with sufficient accuracy. In

[1] J.B. Armstrong, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6 and C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1.

order to standardize the precision of estimates for SIC4 domains within each province, a two-phase sample design was implemented. The first-phase sample of taxfilers is selected at Revenue Canada using strata defined using SIC2 and gross business income (size). Before the second phase sample is selected, an SIC4 code, considered more accurate than codes available from Revenue Canada, is assigned to each sampled unit by Statistics Canada. Strata defined using SIC4 and size are employed during selection of the second-phase sample. The same size boundaries are used for both phases of sampling. A detailed description of the sample design can be found in Choudhry, Lavallée and Hidiroglou (1989b).

First-phase sample selection is done using Bernoulli sampling (also called Poisson sampling). Suppose that taxfiler i falls in first-phase stratum g within a particular province $\times$ SIC2 cell. To determine whether taxfiler i is included in the first-phase sample, a pseudo-random number in the interval (0,1), say R_i, is generated using the taxfiler's unique identification number. The taxfiler is included in the first-phase sample if $R_i \in (0, v_g)$. Bernoulli sampling based on a different set of pseudo-random numbers is used to select the second-phase sample. Using Bernoulli sampling, selection and processing can begin before complete information about the taxfiler universe is available. This advantage of Bernoulli sampling is important, since taxfiler universe information is accumulated over a two-year period. Sample sizes obtained using Bernoulli sampling are random. Choudhry, Lavallée and Hidiroglou (1989b) derive the variance of $\hat{Y}_{h-\mathrm{STRAT}} = \sum_g \sum_{i \in s2 \cap g \cap h} y_i/(v_g \cdot v_{gh})$ using simple random sampling as an approximation to Bernoulli sampling as discussed in Sunter (1986). Under the approximation, a simple random sample of fixed size $n'_g = v_g \cdot N_g$ is selected in size stratum g at the first phase. Let n'_{gh} denote the number of units with SIC4 h in the first-phase sample for size stratum g. At the second phase, a simple random sample of size $n_{gh} = v_{gh} \cdot n'_{gh}$ is selected for SIC4 h and size stratum g, with v_{gh} considered fixed. The variance of $\hat{Y}_{h-\mathrm{STRAT}}$ is given by

$$V_h = \sum_g \left(\frac{1}{v_g \cdot v_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{v_g} - 1 \right) \cdot B_{gh},$$

where

$$A_{gh} = N_{gh} \cdot S^2_{gh},$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y^2_{gh}}{N_{gh}} - S^2_{gh} \right),$$

and S^2_{gh} is the population variance in the second-phase SIC4 $\times$ size stratum gh.

The plan of the paper is as follows. In Section 2, the optimal allocation problem is formulated in the context of the two-phase tax sample. An iterative solution procedure, called the exact method, is proposed. Section 3 includes a description of an approximation to the optimal allocation that can be used to obtain starting values for the exact method. The results of an empirical study involving comparison of various starting values for the exact method are reported in Section 4. Section 5 concludes the paper.

2. EXACT METHOD

In this section the optimal allocation problem is described and an iterative solution method, called the exact method, is proposed. To formulate the problem in the context of two-phase tax sampling, it is sufficient to consider one SIC2 cell in a particular province containing N

units. The cost of selecting a unit in the first-phase sample is K_1, regardless of the stratum in which the unit falls, while the cost of selecting a unit in the second-phase sample is K_2, regardless of stratum. Under Bernoulli sampling, the cost function is

$$F^* = K_1 \cdot \sum_g n'_g + K_2 \cdot \sum_g \sum_h n_{gh}.$$

Since sample sizes n'_g and n_{gh} are random, we use the expected cost

$$F = K_1 \cdot \sum_g \upsilon_g \cdot N_g + K_2 \cdot \sum_g \sum_h \upsilon_g \cdot \upsilon_{gh} \cdot N_{gh}. \tag{1}$$

Rao (1973) and Smith (1989) also solve allocation problems for two-phase sample designs using expected values of random cost functions. In the tax sampling context, the total cost for a province is the sum of the costs for all SIC2 cells within the province. The estimated coefficient of variation of the cost of two-phase tax sampling for the province of Quebec, calculated using 1988 data, was about 1.85%. Coefficients of variation for overall (national) costs were smaller.

It is necessary to minimize (1) with respect to υ_g, $g = 1, 2 \ldots, G$, and υ_{gh}, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$ under the constraints

$$\sum_g \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1\right) \cdot A_{gh} + \sum_g \left(\frac{1}{\upsilon_g} - 1\right) \cdot B_{gh} \leq C_h^2 \cdot Y_h^2, \quad h = 1, 2, \ldots, H, \tag{2}$$

$$0 < \upsilon_g \leq 1, \quad g = 1, 2, \ldots, G,$$

$$0 < \upsilon_{gh} \leq 1, \quad g = 1, 2, \ldots, G, \quad h = 1, 2, \ldots, H,$$

where C_h denotes the target coefficient of variation for SIC4 domain h.

Attempts at direct solution of this problem using the IMSL (1987) implementation of the successive quadratic programming algorithm of Schittkowski (1985) produced mixed results. The algorithm worked well for problems with small numbers of variables and constraints. However, satisfactory solutions for problems including more than approximately 35 variables or more than approximately 50 constraints could not be obtained.

Some costs obtained using direct application of Schittkowski's algorithm in the tax sampling context are given in Table 1. The algorithm was applied to the allocation problems for some SIC2 cells in the province of Quebec involving large numbers of variables and/or constraints using data for tax year 1988. All first-phase and second-phase sampling fractions were started at one when the direct approach was used. The lowest cost obtained using the method that we call the exact method, which will be described later in this section, is also given. The information in the table indicates that direct use of the IMSL implementation of Schittkowski's algorithm is an inappropriate strategy for SIC2 cells with large numbers of variables and constraints.

The exact method is based on a substantial simplification of the problem defined by (1) and (2) that can be achieved by exploiting its structure. In particular, we divide the problem into two main steps that can be solved iteratively. At the first step, (1) is minimized with respect to υ_g, $g = 1, 2, \ldots, G$, conditional on values for all second-phase sampling fractions. This

Table 1

Results for Direct and Exact Methods

SIC 2	No. of variables	No. of constraints	Cost ($) - direct	Cost ($) - exact
30	62	86	5155**	1897
35	37	51	551	512
39	38	50	1667	1450
427*	39	48	27528**	3383

* Three digits of SIC are used for first-phase stratification for construction industries.
** The IMSL routine terminated with an internal error that could not be rectified after consulting published documentation.

step requires the use of nonlinear optimization techniques. The second step involves minimizing (1) with respect to the second-phase sampling fractions, conditional on the values of the first-phase sampling fractions obtained in the first step. No iterations are required for this minimization, since it has a closed form solution. Furthermore, it can be done independently for each $h = 1, 2, \ldots, H$. After completion of the second step, the first step is repeated and the iterative process continued. Convergence is declared when changes in the cost function between consecutive iterations are small.

Let $v_g^{(i)}$ and $v_{gh}^{(i)}$ denote the estimates of the optimal values of v_g and v_{gh} obtained after i iterations (each iteration including one repetition of the two steps described above). At the beginning of iteration $i + 1$, the transformation of variables given by $X_g^{(i+1)} = 1/v_g^{(i+1)} - 1$ is required. This transformation redefines the optimization problem involved in the first step of the iteration as a problem with linear constraints and a convex objective function. Such a convex programming problem is easier to solve.

More precisely, each iteration involves:

(i) Minimization of

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h v_{gh}^{(i-1)} \cdot N_{gh} \right) \Big/ (X_g^{(i)} + 1)$$

with respect to $X_g^{(i)}$, $g = 1, 2, \ldots, G$, subject to the constraints

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(i)} + 1}{v_{gh}^{(i-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(i)} \cdot B_{gh} \geq 0, \quad h = 1, 2, \ldots, H$$

$$X_g^{(i)} \geq 0, \quad g = 1, 2, \ldots, G.$$

(ii) Calculation of $v_g^{(i)} = 1/(X_g^{(i)} + 1)$, $g = 1, 2, \ldots, G$. Minimization, independently for each $h = 1, 2, \ldots, H$, of

$$F_h = \sum_g v_g^{(i)} \cdot v_{gh}^{(i)} \cdot N_{gh}$$

with respect to $v_{gh}^{(i)}$, $g = 1, 2, \ldots, G$, subject to the constraints

$$C_h^2 \cdot \hat{Y}_h^2 - \sum_g \left(\frac{1}{v_g^{(i)} \cdot v_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^{(i)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < v_{gh}^{(i)} \leq 1, \quad g = 1, 2, \ldots, G,$$

where h is considered fixed.

It will be shown in Section 3 that solution of step (ii) does not require use of numerical methods. Therefore, the exact method only requires the solution of a series of convex programming problems, each involving only G variables. A convex programming problem is much easier to solve than a general nonlinear programming problem. A local solution of a convex programming problem is also a global solution.

Let $F^{(i)}$ denote the value of the cost function, (1), obtained using $v_g^{(i)}$ and $v_{gh}^{(i)}$. The $F^{(i)}$ values form a monotonically decreasing sequence and therefore converge to a limit. Whether this limit value and the corresponding sampling fractions give the global minimum depends on the starting value. This problem is caused by the geometry of the constraints in (2). In practice one should try several starting values to get the best solution. One starting value is given by the approximate method, which is described in the next section and does not require iterations.

3. APPROXIMATE METHOD

In this section, an allocation method that gives an approximation to the optimal allocation is described. The method was first suggested by Choudhry, Lavallée and Hidiroglou (1989a). Assuming that all the second-phase sampling fractions are equal to one, an approximation to the optimal allocation of the first-phase sample is calculated. Then the second-phase sample is allocated, conditional on the first-phase sampling fractions. Since the cost of sampling a unit in both phases of sampling does not depend on the stratum in which the unit falls, minimizing cost is equivalent to minimizing sample size at each step of this method.

At the first step of the method, an approximate solution to the optimal allocation problem for a one-phase sample design is calculated. This step involves finding the minimum, independently for each h, of

$$F^{(h)} = \sum_g v_{g|h} \cdot N_g \tag{3}$$

with respect to $v_{g|h}$, $g = 1, 2, \ldots, G$. The notation $v_{g|h}$ is used to denote the fact that a sampling fraction for size stratum g is determined subject to only one precision constraint, namely the constraint for SIC4 domain h, where h is fixed. In particular, the minimization must be done subject to the constraints

$$\sum_g \left(\frac{1}{v_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \tag{4}$$

$$0 < v_{g|h} \leq 1, \quad g = 1, 2, \ldots, G. \tag{5}$$

One can show that the minimum of (3) is obtained when (4) holds with equality, so that the problem defined by (3), (4), and (5) is equivalent to finding the critical point of the lagrangian

$$L = \sum_g v_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \right].$$

Setting the derivatives with respect to $v_{g|h}$ equal to zero yields

$$v_{g|h} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot (-\lambda)^{1/2}, \quad g = 1, 2, \ldots, G. \tag{6}$$

Setting $\partial L/\partial \lambda = 0$ we obtain

$$(-\lambda)^{1/2} = \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \Big/ \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{7}$$

After substitution of (7) into (6), we obtain the optimal sampling fraction for size stratum g given only one precision constraint, for SIC4 domain h,

$$v_{g|h}^* = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot$$

$$\sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \Big/ \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{8}$$

If one or more of the sampling fractions given by (8) are greater than one, one can set them equal to one and solve a modified allocation problem with a reduced number of strata. This approach corresponds to the overallocation procedure discussed by Cochran (1977). It is necessary to calculate (8) for $h = 1, 2, \ldots, H$. The approximate first-phase sampling fraction for size stratum g, v_g^*, is set equal to the largest value in the set $\{v_{g|h}^*, h = 1, 2, \ldots, H\}$ for $g = 1, 2, \ldots, G$, an approach that ensures that the precision constraint for each SIC4 domain will be satisified.

Given first-phase sampling fractions, optimal second-phase sampling fractions can be easily determined. Assume that, for the SIC2 × province cell h, the size strata included in the allocation problem correspond to a set of integers, Γ. We set the second-phase sampling fractions equal to one for those size strata that are not included in the allocation problem. Normally, one would have $\Gamma = \{1, 2, \ldots, G\}$ but because of overallocation during allocation of the second-phase sample, for example, Γ may not include all integers between 1 and G. The problem of allocating the second-phase sample is equivalent to the problem of finding the minimum of

$$F_h = \sum_{g \in \Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} \tag{9}$$

with respect to v_{gh}, $g \in \Gamma$, subject to the constraints

$$\sum_{g \in \Gamma} \left(\frac{1}{v_{gh}} - 1 \right) \cdot \frac{A_{gh}}{v_g^*} \leq M_h, \tag{10}$$

$$0 < v_{gh} \leq 1, \quad g \in \Gamma, \tag{11}$$

where

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_g^*} - 1\right) \cdot (A_{gh} + B_{gh}).$$

Note that the expected number of units with SIC4 h in the second-phase sample for size stratum g, $v_g^* \cdot N_{gh}$, is employed in (9). It is easy to show that (9) attains a minimum when the constraint (10) holds with equality. Consequently, the minimization problem is equivalent to finding the critical point of the lagrangian

$$L_h = \sum_{g\in\Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} + \lambda \cdot \left(M_h - \sum_{g\in\Gamma} \cdot \left(\frac{1}{v_{gh}} - 1\right) \cdot \frac{A_{gh}}{v_g^*}\right),$$

with respect to and v_{gh}, $g\in\Gamma$, and λ, subject to the constraints

$$0 < v_{gh} \leq 1, \quad g\in\Gamma.$$

Setting the first derivatives of L_h equal to zero and simplifying, one obtains

$$v_{gh} = (-\lambda \cdot A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*), \quad g\in\Gamma, \tag{12}$$

$$(-\lambda)^{1/2} = \sum_g (N_{gh} \cdot A_{gh})^{1/2}/D_{\Gamma h}, \tag{13}$$

where

$$D_{\Gamma h} = C_h^2 \cdot Y_h^2 \sum_{g\in\Gamma} \left(\frac{1}{v_g^*}\right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^*} - 1\right) \cdot (A_{gh} + B_{gh}).$$

Note that there is no solution to the allocation problem unless $D_{\Gamma h}$ is positive. Substituting (13) into (12) yields

$$v_{gh}^* = (A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*) \cdot \sum_{g\in\Gamma} (N_{gh} \cdot A_{gh})^{1/2}/D_{\Gamma h}. \tag{14}$$

If v_{gh}^* is greater than one for certain gh, the overallocation procedure described above can obviously be employed. Note that (14) also provides the solution for step (ii) of each exact method iteration.

4. EMPIRICAL STUDY

The approximate method serves two purposes. First, it provides a good starting value for the exact method. Second, it may be easier to implement in practice. In this section, we report the results of an empirical comparison using data from the province of Quebec for tax year 1988. Results obtained using the exact method with various starting points, as well as the approximate method, are reported. Since the quantities N_{gh}, Y_h and S_{gh}^2 required by both methods were unknown, estimates based on the data were used.

The size stratification used by the survey, including four take-some strata and one take-all stratum, was employed. Allocations were computed for 64 SIC2 cells (all of the Quebec data excluding a few small SIC2s). The number of sampling fractions determined in these allocations ranged from 8 to 92 with a median of 24. The number of constraints ranged from 9 to 115 with a median of 31. There were 20 SIC2 cells involving more than 35 variables and 18 of these cells also involved more than 50 constraints. A total of 1850 second-phase strata including about 230,000 population units were involved.

The first-phase sampling cost, corresponding to the cost of microfilming or photocopying a tax return at Revenue Canada, sending the information to Statistics Canada and determining an SIC4 code, was set at \$1.40 per unit. The second-phase sampling cost, corresponding to the cost of transcribing values for financial variables, was set at \$7.00. These costs are comparable to those incurred during operation of the actual survey.

Allocations were computed using the exact method with three starting values: I – solution of the approximate method; II – all first-phase sampling fractions set to one with the corresponding conditionally optimal second-phase fractions; and III – a randomly chosen set of feasible first-phase sampling fractions, with the corresponding conditionally optimal second-phase fractions. In addition, the exact method was started at a perturbation of each of these starting values. The perturbed value for the first-phase sampling fraction for size stratum g for starting value I was $v_g^{(0)} = 0.1 + 0.9 \cdot v_g^*$, where v_g^* is the solution of the approximate method. Second-phase sampling fractions were started at values that are optimal, conditional on the perturbed first-phase fractions. Starting value III was perturbed analogously. The perturbed value corresponding to starting value II was $v_{gh}^{(0)} = 0.1 + 0.9 \cdot v_{gh}^{**}$, where v_{gh}^{**} is optimal, conditional on a census at the first phase of sampling. For each starting value, the best result obtained using either the value itself or the corresponding perturbed value was retained. Convergence was declared if the absolute relative change in the cost function between consecutive iterations was less than 10^{-4}. The IMSL implementation of Schittkowski's successive quadratic programming algorithm was used to solve nonlinear programming problems.

Results are reported in Table 2. Total costs for four alternatives are given. In addition, the number of SIC2 cells for which each starting value for the exact method produced better results than alternative starting values is shown. Computing costs are not reported, since they were small enough to be inconsequential.

The results indicate that the approximate solution provided the best starting values for the exact method. Although starting value II produced better results than starting value I for 17 SIC2 cells, the total cost associated with starting value II was higher than the total cost for the approximate method. The exact method performed poorly when starting values were determined by random selection of a feasible set of first-phase sampling fractions.

Table 2

Results for Exact and Approximate Methods

Method	Exact – Starting value			Approximate
	I	II	III	
Total cost (\$)	122779	139347	200998	130228
No. cells with best result*	48	17	1	

* For two cells starting values I and II produced the same result, which had lower cost than the result obtained using starting value III. Consequently, the numbers reported in this row of the table add to 66 rather than 64.

Although the total cost using the exact method with starting value I was only 5.7% lower than the cost of the approximate method, it should be noted that the exact method with starting value I can do no worse than the approximate method. The exact method with starting value I produced better results than the approximate method for 42 cells.

5. CONCLUSION

A sample allocation problem for two-phase survey designs is formulated as a constrained optimization problem in Sections 1 and 2. If the numbers of variables and constraints involved in the problem are small, the solution can be obtained through direct application of numerical methods. However, the direct approach does not work well for large numbers of variables and constraints.

By exploiting the mathematical structure of the problem, it can be divided into two subproblems: the first is a convex programming problem with linear constraints that involves a much smaller number of variables, and the second can be solved without the use of numerical methods. The algorithm proposed in Section 2 consists of iterations between the two subproblems. It is computationally simpler and more effective in practice than the direct approach for problems involving large numbers of variables and constraints. An approximate solution to the sample allocation problem that does not require use of numerical methods is proposed in Section 3. The empirical study in Section 4 shows that it works especially well as a starting value for the algorithm proposed in Section 2.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of Pierre Lavallée, who was the first to derive the expressions for conditionally optimal second-phase sampling fractions involved in the approximate method. Thanks are due to a referee and an associate editor for useful comments. C.F.J. Wu is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

ARMSTRONG, J.B., BLOCK, C., and SRINATH, K.P. (1991). Two-phase sampling of tax records for business surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 228-233.

BOOTH, G., and SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.

CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989a). Two-phase sample design for tax data. Unpublished document, Business Survey Methods Division, Statistics Canada.

CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989b). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

IMSL (1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

SCHITTKOWSKI, K. (1985). NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition? *Journal of the American Statistical Association*, 84, 916-921.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 263–277
Statistics Canada

The Role of the Interviewer in Survey Participation

MICK P. COUPER and ROBERT M. GROVES[1]

ABSTRACT

Using data from a survey of U.S. Census Bureau interviewers, this paper examines whether experienced interviewers achieve higher response rates than inexperienced interviewers, controlling for differences in survey design and attributes of the populations assigned to them. After demonstrating that the relationship is positive and curvilinear, it attempts to explain the mechanisms by which experienced interviewers achieve these rates and elaborate the nature of the relationship. It examines what behaviors and attitudes underlie the higher success, with the hope that they might be instilled in trainees.

KEY WORDS: Interviewers; Nonresponse; Response rates; Survey participation.

1. INTRODUCTION

Survey methodologists have long suspected the interviewer to be an important source of variation in response rates. Indicators of this include observed differences among trainees in the ability to absorb and put into practice the interviewing guidelines, interviewer variation in item missing data rates, individual interviewers' response rates, and the ability of some interviewers to convert the initial refusals of others. However, several of these indicators are affected by the fact that interviewers often do their work in different subpopulations, and thus face different challenges to complete their assignments.

Much of what we believe about the impact of the interviewer on survey participation remains untested or inconclusive. In an oft-cited study, Durbin and Stuart (1951) found experienced interviewers to be "decidedly superior" to student volunteers in terms of response rates. Groves and Fultz (1985) found that novice interviewers (1 to 6 months of tenure) had the highest refusal rates in a telephone survey. In a study cited by Inderfurth (1972), nonresponse rates for Census Bureau interviewers trained in 1962 and 1963 declined steadily over the first months of service, reaching the level of experienced interviewers after 22 months. In contrast, Singer, Frankel and Glassman (1983, p. 74) found the effect of experience on response rates in a telephone survey to be counter-intuitive, that is, more experienced interviewers did **not** achieve higher response rates. They do note, however, that this result is based on only six interviewers. In a study of 16 field interviewers in Sweden, Schyberger (1967) found nonresponse rates to be **higher** for experienced than for newly recruited interviewers. In short, the common belief of experienced interviewers being more successful is not uniformly supported empirically.

This paper examines the role of various interviewer characteristics, particularly experience, in achieving respondent cooperation. It should be noted that the interviewer represents only one part of a large set of factors that can affect survey participation. Such factors include respondent characteristics, the respondent-interviewer interaction, survey design features, and contextual and situational factors. For a review of these factors, see Groves, Cialdini and Couper (1992).

[1] Mick P. Couper and Robert M. Groves, U.S. Bureau of the Census and University of Michigan. Room 2315-3, Bureau of the Census, Washington, DC 20233.

We should also note that different models may be more suitable for different components of nonresponse. For instance, interviewer motivation, tenacity and effort expended may be more important in reducing noncontacts, while persuasion skills play a greater part in the refusal component of nonresponse. The data analyzed here do not permit us to distinguish between these components of nonresponse. This may weaken the explanatory power of the models tested.

In this paper we will address two questions: (a) do experienced interviewers achieve higher response rates? (b) if so, what are the mechanisms underlying the relationship between experience and rates? These questions are important to the survey research community. If the behaviors used by successful experienced interviewers can be taught to inexperienced interviewers, then their success might be transferred to the new recruits. If not, then the value of reducing turnover among experienced interviewers remains high for survey organizations.

2. TOWARD A MODEL OF SURVEY PARTICIPATION

A number of interviewer characteristics can be identified that have a potential impact on survey participation. These are illustrated in Figure 1. The effects of interviewer experience, expectations and behavior on response rates, controlling for assignment area and survey design features, will be explored. Each of the sets of variables will be discussed in turn.

2.1 Interviewer experience

First, interviewers' experience is expected to have a positive effect on the response rates they obtain. This stems from lessons learned through trial and error application of alternative techniques over time, and from alternative training guidelines and experiences on different surveys. Experience thus has two components: length and breadth. Length of experience might be indicated by the number of years a person has worked as an interviewer. One indicator of breadth of experience is the number of different organizations an interviewer has worked for, or the number of different kinds of studies an interviewer has worked on. It is argued that length and breadth of experience both serve to increase the variety of different interviewing situations to which an interviewer is exposed.

We expect the relationship between length of experience (as measured by tenure) and response rates to be curvilinear. Experience in the first few years of interviewing will have a greater impact on response rates than in later years. After a certain point, the number of new situations faced by interviewers declines, and interviewers become comfortable dealing with the wide variety of sample persons and assignment areas they may face. After this, additional years of experience may not produce further gains in response rates.

An alternative hypothesis is that self-selection rather than experience produces higher response rates among interviewers with longer tenure. In other words, it is not that individual interviewers get better over time, but that better interviewers tend to stay, while weaker interviewers leave the job. We believe that a combination of these two factors explains variations in interviewer performance. However, the self-selection hypothesis cannot be tested in a cross-sectional study such as this, and caution must be exercised in drawing inferences from these analyses.

If experienced interviewers achieve higher response rates, we hypothesize that this takes place through the intervening effects of interviewer expectations (*e.g.* confidence) and behavior (*e.g.* effective oral presentation). Note that we posit no direct effect of experience on response rates. In other words, is it possible to identify interviewer attitudes and behaviors that may account for possible differences in response rates?

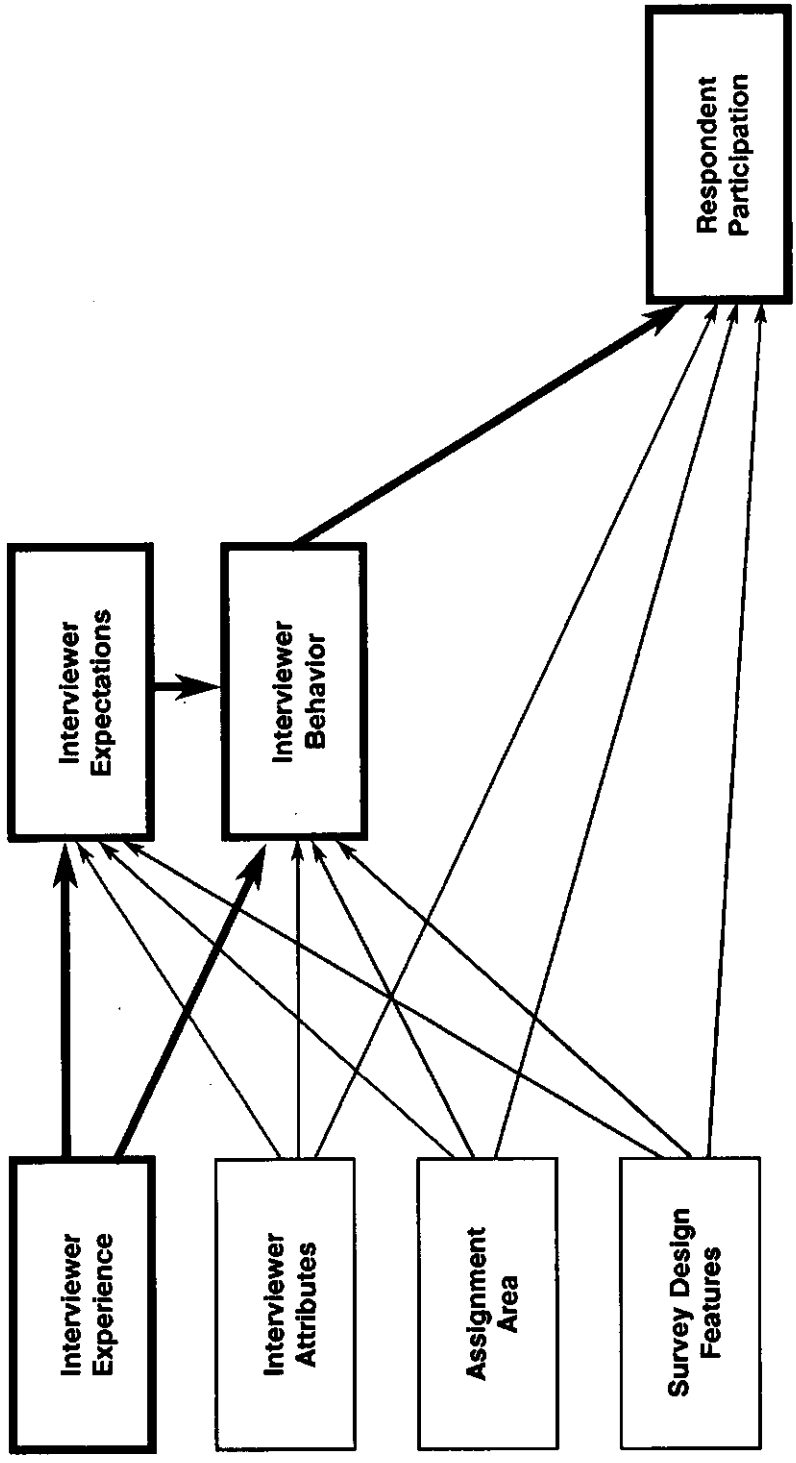


Figure 1. Model of Survey Participation Role of the Interviewer.

2.2 Interviewer expectations

It is hypothesized that positive interviewer expectations lead to higher response rates. Interviewers who have a greater belief in their ability to persuade sample persons to participate, who believe in the legitimacy of the work they are doing, and who are confident that most people agree to participate in surveys, are likely to get higher response rates than those who believe otherwise. This argument has some empirical support in the study by Singer, Frankel and Glassman (1983), in which it was found that interviewers who anticipated prior to the survey that the task of persuading respondents was "moderately easy", achieved higher response rates than those who believed the task to be "moderately difficult".

2.3 Interviewer behavior

With regard to interviewer behaviors, we seek to identify the mechanisms by which greater experience and positive expectations translate into higher response rates. The behavior of interviewers in gaining cooperation from sample persons may be likened to that of other "compliance professionals" (such as salespersons, fundraisers, *etc.*). Based on an extensive review of experimental and observational evidence, Cialdini (1984, 1990) identifies six compliance principles used to decide whether to accede to a request. Briefly, these principles are as follows:

(a) Reciprocation: One should be more willing to comply with a request to the extent that the compliance constitutes the repayment of a perceived gift, favor, or concession.

(b) Consistency: After committing oneself to a position, one should be more willing to comply with requests for behaviors that are consistent with that position.

(c) Social validation: One should be more willing to comply with a request to the degree that one believes that similar others would comply with it.

(d) Authority: One should be more willing to yield to the requests of someone who one perceives as a legitimate authority.

(e) Scarcity: One should be more willing to comply with requests to secure opportunities that are scarce.

(f) Liking: One should be more willing to comply with requests of liked others.

We are interested in the extent to which interviewers make use of these principles to persuade sample persons to participate in a survey.

It is argued that interviewers who make appropriate use of each of these strategies are likely to have greater success in persuading reluctant sample persons to participate. However, the use of such techniques indiscriminately in all situations may backfire. For example, the invocation of the authority principle in areas where suspicion of government is high may well have a negative effect on cooperation. The use of these compliance principles may not be universally effective in all situations or for all sample persons.

Thus, it is not just **whether** these techniques are used by interviewers, but also **how** they are used. Two concepts are of interest here. One is the number of different techniques that an interviewer has at his/her disposal, and the second is how appropriately such techniques are applied. The first we will refer to as the "repertoire of techniques" available to the interviewer. A novice interviewer may learn one or two "canned" introductions during training, and use them on all sample persons he/she encounters. In contrast, the experienced interviewer has a wide repertoire of approaches upon which to draw, and can apply them as the situation warrants.

The second concept is that of appropriate application of the skills or techniques at the interviewer's disposal. We refer to this as "tailoring". An interviewer is expected to be an "astute psychological diagnostician" (Cannell 1964), to be able to size up a situation quickly, and apply the appropriate persuasive messages. These skills are gained through experience, either on the job or in life in general. The novice interviewer, with fewer skills and less confidence, may rigidly adhere to a small number of "tried and trusted" approaches. The experienced interviewer is better able to tailor his/her approach to each potential respondent.

It may be that adaptability and appropriate application of persuasive techniques are more critical than the actual behaviors or techniques themselves. If so, it should be possible to develop a more parsimonious model using only the latter concepts and dropping the specific behaviors measured.

2.4 Assignment area

To examine the effect of interviewers on survey participation, we need to take into account the fact that they are assigned different areas to interview. Ideally, the research design would have randomly assigned interviewers to sample areas, removing any statistical confounding between interviewer and population characteristics. Without such randomization, we attempt to specify those population characteristics important to response rate and statistically control for them.

First, the problem of obtaining cooperation from sample persons in inner-city areas is well known (see Steeh 1981, Smith 1983). House and Wolf (1978) found that rising crime rates, particularly in high density urban areas, have been a major deterrent to survey participation, and to trusting and helping behavior in general (Korte and Kerr 1975). We expect this arises both because of residents' reluctance to interact with strangers, and unease among interviewers on entering these neighborhoods.

Turning to characteristics of sample households, household size has been found to correlate positively with response rates (see Gower 1979; Paul and Lawes 1982; Rauta 1985). Single-person households tend to have relatively high refusal rates (see Brown and Bishop 1982; Wilcox 1977). This may be due in part to the large proportion of elderly persons living alone. Families with dependent children, on the other hand, tend to have higher response rates. Lievesley (1988) notes that higher response rates in certain areas of the U.K. may be explained by the high probability of finding someone at home arising from high proportions of children aged 0-4.

The findings on sample person characteristics are somewhat more mixed. A number of researchers (see Brown and Bishop 1982; Hawkins 1975; Herzog and Rogers 1988; Weaver 1975) have found age to be associated with nonresponse. The impact of other sample person characteristics such as race, education, socio-economic status, gender, *etc.* are somewhat inconsistent (see Groves (1989) and Goyder (1987) for reviews of these factors).

2.5 Survey design features

Finally, survey design features (topic, burden, respondent selection rules, *etc.*) are likely to influence a sample person's decision to participate, both directly and in terms of constraints on interviewer expectations and behavior.

2.6 Interaction effects on response rate

We suspect that there may be a number of statistical interaction effects of influences on nonresponse. One question is whether there are some areas (such as high density central city areas) in which interviewer experience is more important than other areas. For example, high density urban areas may be more diverse, requiring greater experience to deal with a greater

variety of different situations. Behavior in areas where the situations presented to interviewers are all very similar could be more easily learned, as fewer persuasion strategies would be needed.

We also suspect that different surveys may obtain varying response rates for different subpopulations as a result of the differential salience of the survey topic to such groups. For example, it may be expected that the National Crime Survey (which focuses on criminal victimization) may get higher response rates in high crime areas than in low crime areas. Similarly, the National Health Interview Survey (which measures health-related activities) may obtain higher response rates in areas with an older than average population. Similar interactions may be expected between the Consumer Expenditure Survey and such variables as average household size and income level.

3. METHOD

3.1 Data collection strategies

The results in this paper are part of a larger study of survey participation in face-to-face surveys in the United States. The first part of the work involved a series of focus groups with interviewers working on a variety of different surveys around the country. The insights gained from these groups led to the development of a structured questionnaire to test some of these hypotheses on a larger audience of interviewers.

The interviewer surveys had the goal of measuring behavioral, experiential and attitudinal influences on levels of cooperation obtained by interviewers. The questionnaire was developed and tested by staff at the Survey Research Center in collaboration with staff from the U.S. Census Bureau.

This questionnaire was administered to U.S. Census Bureau interviewers working on the following three personal visit surveys:

(a) the Consumer Expenditure Quarterly Survey (CE), sponsored by the Bureau of Labor Statistics;

(b) the National Health Interview Survey (HIS), sponsored by the National Center for Health Statistics; and

(c) the National Crime Survey (NCS), sponsored by the Bureau of Justice Statistics.

The questionnaire was mailed in February, 1990, to Census Bureau interviewers working on these three surveys. All interviewers were paid their normal salary rate for completing the questionnaire (most were paid for an hour of their time). In an effort to seek candid responses and eliminate the threat of supervisory intervention, interviewers were assured that their individual responses would not be seen by or discussed with any of their supervisors, and that the results would be reported only as statistical totals.

Questionnaires were mailed back to the central office. Reminder letters and telephone calls were used to increase the response rate. A total of 1,013 completed questionnaires were received, representing a response rate of 97.1%. A number of questionnaires were excluded from the analyses reported here. All supervisory interviewers (256) were excluded. These people often have no regular assignments of their own, and typically work on a number of different surveys. They are often used for refusal conversion, or to "clean up" otherwise incomplete assignments. With supervisory interviewers excluded, transfer of assignments from one interviewer to another on these surveys is rare. For purposes of calculating interviewer-level response rates, each nonresponse case was counted against the original interviewer, regardless of whether it was later converted by another. In addition, those interviewers who started work during the period

in which the interviewer survey was administered, and for whom no historical response rate information was available, were also excluded (46 interviewers). This left a total of 711 interviewers, 207 from CE, 139 from HIS and 365 from NCS. The numbers of cases included in the analyses may be further reduced due to missing data on certain variables.

3.2 Data structure

In addition to the questionnaire responses, other variables were added to the data file. These included a set of variables to represent each interviewer's assignment area. Typically, the primary sampling unit (PSU) in which an interviewer works consists of one or more coterminous counties. County-level data were extracted from the County and City Data Book (Bureau of the Census 1988), aggregated to the PSU level, and attached to the interviewer records. Note that these variables can only reflect gross differences in assignment area and cannot, for example, distinguish between central city and suburban areas.

The date each interviewer was hired by the Census Bureau was obtained from administrative records to create a variable to serve as a measure of tenure. Although it does not indicate length of experience on a particular survey, it does reflect the length of time an interviewer was employed by the Census Bureau.

A major drawback of this study is that it was not possible to obtain measures of race, age, gender, or other demographic attributes of the interviewer. Confidentiality restrictions prevented access of personnel records for this information, nor could these be asked in the interviewer questionnaire.

3.3 Analytic plan

Three different surveys are represented in the data set. Instead of introducing control variables measuring key design features of the surveys, dummy variable indicators of the survey were used to control on important design differences among them.

The dependent variable is aggregate response rate for the six month period, October 1989, through March 1990. It was not possible to obtain interviewer-level data on the components of nonresponse (particularly refusals) for this period. These rates thus do not distinguish between noncontact and refusal components of nonresponse. Hence, it should be noted that the analyses reported here are based on interviewer-level **response** rates rather than **refusal** rates.

The nonresponse rates for the three surveys for 1990 (based on national sample totals) are presented in Table 1.

Refusals as a proportion of total nonresponse varies from 87% for CE to 52% for NCS. We suspect that different sets of factors operate to affect these two components of nonresponse. Ideally, separate models would be fitted for each component, but this was not possible given the current data. To the extent that factors affecting refusals are different from those affecting other components of nonresponse (such as noncontacts), the results will be confounded (see Lievesley 1988). It can also be seen that nonresponse rates for these three surveys are low to begin with. This may further restrict the ability of these models to explain differences among interviewers.

Given that the size of the interviewer assignments vary (and hence affect the variance of the measured individual response rates), we used weighted least squares (WLS) with assignment size as the weight. Comparisons of the WLS results with those using ordinary least squares (OLS) solutions were made, and it was found that WLS reduces the size of the coefficients marginally, but does not affect the sign or relative strength of the coefficients. All the analyses reported here are based on the WLS solutions.

Table 1
1990 Nonresponse Rates for Three Surveys

Survey	Nonresponse rate	Refusal rate
	%	%
Consumer Expenditure Survey	13.4	11.6
Health Interview Survey	4.5	2.8
National Crime Survey	3.1	1.6

A series of tests were performed to determine the appropriateness of the models specified. A number of outliers in the dependent variable were detected. However, removal of these outliers had little or no effect on the results obtained, and they were therefore retained in all analyses. Tests of the normality assumption were also conducted. The normal probability plots show that the residuals from these models do not differ markedly from a normal distribution.

It is hypothesized that the effect of tenure on response rate is greater in the first few years. The tenure variable is transformed (the natural log is used) to reflect this. The transformed variable indeed produced an improvement in fit over the linear tenure variable.

A more detailed description of the variables used in these analyses can be found in Appendix A.

4. LIMITATIONS

Before describing the analyses, it is important to note some of the limitations of these data. First, these findings refer only to interviewers working on three ongoing national surveys at the Census Bureau at the time at which the interviewer survey was conducted. It is not possible to generalize to other face-to-face or telephone surveys conducted by academic or private sector organizations.

Furthermore, the data are cross-sectional in nature. Cohort and period effects are confounded with the effects of experience. That is, any observed response rate differences by interviewer experience may be due to changes in the quality of interviewers hired over time, in the effectiveness of interviewer training over time, or in differential turnover by interviewer quality. Hypotheses can be constructed to support both positive and negative effects of these factors on response rates. Hence, the measured impact of interviewer experience on response rates is a complex combination of these factors. Longitudinal measurement of interviewers is needed to disentangle these effects.

Interviewers are not randomly assigned to areas. Although we have attempted to control for a number of characteristics of assignment area that may impact on response rates, there may be many other factors that could explain differences in response rates across assignment area. Further, we are limited to weak controls, on attributes of counties and groups of counties, not on attributes of specific assignment areas within counties given to interviewers. A hierarchical analysis containing data on individual respondents and interviewers assigned to them would improve these control factors.

Finally, the dependent variable was measured for a time period up to and including the administration of the interviewer questionnaire. More recent response rate data were not available at the time. Given that behaviors and expectations were not measured before the response rates were obtained, caution should be exercised in attributing causality.

Despite these limitations, these data provide us with the opportunity to test prevailing beliefs about the role of interviewer experience in response rates, and to explore the role of interviewer expectations and behavior in face-to-face surveys.

5. RESULTS

First, we measured the impact of experience, controlling for characteristics of assignment areas and dummy variables for the surveys (Model 1 in Table 2). Let us first examine the coefficients of the control variables. With few exceptions, most of the assignment area variables have a significant impact on response rates. Both population density and crime rate act as expected, with lower response rates being obtained in high crime, high density areas. The negative effect of household size is contrary to expectation. This may be explained in part by the fact that these surveys all collect information from or about **all** adult household members, thereby increasing the reporting burden for large households. This is contrary to many surveys where a single adult is selected from each household. The effect of age is as hypothesized, with response rates tending to be lower (but not significantly so) in areas with larger proportions of persons over 65, but higher in areas with many households who have young children.

The large effects for the two survey variables (relative to the omitted category of the Consumer Expenditure Survey) reflect differences in the mean response rates for these three surveys. Such differences can be attributed to a host of survey design differences (length of the interview, respondent selection rules, panel versus cross-sectional designs, content of the questionnaires, *etc.*) that are beyond the scope of this paper. Nevertheless, it is clearly necessary to control for these differences.

Now, let us examine the measured effect of experience, given these control variables. It can be seen that tenure has a strong positive effect on response rates, even when controlling for the nature of the area to which an interviewer is assigned. This appears to confirm prevailing beliefs about the role of interviewer experience. Interviewer differences in response rates appear to be more than simply artifacts of differences in the areas to which they are assigned, and experience plays a key role in such interviewer differences.

The inclusion of an indicator for breadth of experience was also tested, but found to have no significant effect in the presence of the remaining variables. It thus appears that, for Census Bureau interviewers at least, experience working for other survey organizations does not appear to have any marginal impact on response rates over and above that of tenure.

Does tenure have a differential impact on response rates in different assignment areas? Model 2 in Table 2 includes an interaction term between the log of tenure and population density. An additional interaction term between tenure and crime rate was also tested, but this coefficient was found to be insignificant, and the interaction had little impact on remaining elements of the model. The interaction term in Model 2 is statistically significant, but the sign is opposite to that expected. We hypothesized that experience would have a greater impact in high density areas, but this does not appear to be the case. An alternative explanation may be a "burnout effect". More experienced interviewers in high density urban areas may be losing their enthusiasm sooner than experienced interviewers in less stressful rural areas, and this contributes to lower response rates. Interviewer burnout may be one factor contributing to higher turnover rates in the large metropolitan areas.

Table 2

Results of WLS Regression Analyses of NCS, HIS, CE Interviewer-Level Response Rates

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Intercept	96.94	(3.19)	96.21	(5.39)	94.95	(3.25)	93.44	(3.35)
Assignment area:								
Population density	−0.00017**	(0.000023)	−0.000078*	(0.000038)	−0.000084*	(0.000038)	−0.000071	(0.000038)
Crime rate	−0.00024**	(0.000055)	−0.00021**	(0.000055)	−0.00023**	(0.000056)	−0.00022**	(0.000056)
Percent 65 or older	−0.057	(0.051)	−0.054	(0.050)	−0.061	(0.051)	−0.061	(0.052)
Percent under 5	0.41*	(0.16)	0.37*	(0.16)	0.29	(0.17)	0.35*	(0.17)
Household size	−3.20*	(1.70)	−2.92*	(1.24)	−2.88*	(1.26)	−3.09*	(1.27)
Survey Indicators:								
NCS[1]	6.72**	(0.40)	6.67**	(0.40)	6.68**	(0.41)	6.55**	(0.42)
HIS[1]	5.65**	(0.46)	5.63**	(0.46)	5.64**	(0.47)	5.65**	(0.48)
Interviewer experience:								
Log (tenure)	0.62**	(0.14)	0.74**	(0.14)	0.69**	(0.15)	0.72**	(0.15)
Log (tenure) × density			−0.00010**	(0.000032)	−0.00011**	(0.000032)	−0.00011**	(0.000032)
Interviewer expectations:								
Confidentiality					0.61	(0.37)	0.59	(0.37)
Rate/quality					0.046	(0.40)	−0.00073	(0.41)
Efficacy					0.55**	(0.15)	0.53**	(0.15)
Interview behaviors:								
Authority							0.14**	(0.055)
Reciprocation							0.67*	(0.29)
Social proof							0.18	(0.32)
Saliency							−0.19	(0.33)
Scarcity							−0.66*	(0.29)
Consistency							−0.21	(0.29)
Repertoire							−0.0068	(0.065)
Tailoring							−0.042	(0.054)
Adjusted R^2	0.3553		0.3640		0.3784		0.3873	
(n)	(679)		(679)		(645)		(639)	

** $p < .01$
* $p < .05$
[1] CE Interviewers are the omitted category.

Interactions between the three surveys and various assignment characteristics were also tested. None of these appear to have any noticeable effect in these models, and are not discussed further. As a further test for the presence of additional interactions involving the survey variables, separate models were fitted for each of the three surveys. The models obtained are essentially the same for each of the three surveys examined. Thus, although the level of response differs across the three surveys, the **relative** impact of tenure on response rates appears to be the same.

Given that it appears that experienced interviewers achieve higher response rates regardless of the areas to which they are assigned, we can proceed to address the question of **how** experience impacts on levels of cooperation. What makes a more experienced interviewer better at gaining cooperation from respondents?

The first step involves the addition of interviewer expectation variables to Model 2. The results are presented as Model 3 in Table 2. All three expectation variables act in the expected direction, although only one achieves statistical significance at traditional levels. It appears that those interviewers who have a greater belief in their ability to convince reluctant respondents to participate, actually achieve higher response rates.

It should be cautioned that the causal link between expectations and response rates cannot be established in a cross-sectional study such as this. It may be that greater success leads to greater expectations of future success, rather than the other way around. This interpretation opposes the hope that instilling a greater sense of self-efficacy in interviewers will produce higher levels of response. Nevertheless, this finding is an intriguing one that demands further attention.

The next step was to add the set of interviewer behaviors into the model. The results can be seen in Model 4 in Table 2. Two things can be noted about these results. First, the inclusion of this set of interviewer behaviors failed to explain away the effect of tenure. In fact, the coefficient for tenure is hardly affected by the addition of either the expectation variables or the behavior variables.

Second, the results for the specific behaviors are somewhat mixed. It was expected that the coefficients for all the behavior variables would be positive. This is not the case. The results for authority and reciprocation indicate that interviewers who use these techniques achieve higher response rates. In contrast, use of the scarcity principle appears to have the opposite effect. Pressure on a respondent to meet certain deadlines may well backfire. The remainder of the behavior variables do not appear to have a significant effect on the response rates attained by Census Bureau interviewers.

It was suggested earlier that a reduced model, using only repertoire and tailoring, should be considered. In Table 2 it was seen that these two variables do not have significant effects in the presence of the other behavior variables. Even after removing the other behavior variables from the model, repertoire and tailoring still have little impact on response rates. Thus, the argument that the way interviewers use various compliance techniques are more important than the actual behaviors themselves gains little empirical support from these data. However, the measures of these two concepts may be weak, and a better test of their role should be done at the contact-level of analysis.

6. DISCUSSION

This paper set out to measure whether experienced interviewers achieve higher response rates than inexperienced interviewers. It found they do. It then tried to explain why they do. It largely failed. One reason may be that the model is incorrect. However, continued discussions with interviewers and supervisory staff lead us to believe that this theoretical formulation has some merit.

Four explanations can be posited. First, the model is being tested at the wrong level of aggregation. Although the questionnaire focused on what interviewers usually or typically do, we are more interested in how they act in specific situations. A more appropriate test of these ideas should be conducted at the contact or household level. Second, the measurement of various concepts may be inadequate. Improvements in the translation of concepts from the compliance literature into specific interviewer behaviors may be made. Third, it should again be noted that these models deal with response rates not refusal rates. It may be that certain behaviors are more appropriately directed at persuading sample persons to participate (aimed at reducing refusals), while others may serve more to gain access to sample persons (the non-contact portion of nonresponse). Separate models for these two processes could not be developed here. Finally, other unmeasured characteristics of interviewers (appearance, voice quality, dress, *etc.*) may also play a role in influencing the respondent's decision.

These possible shortcomings do not negate the role of these behaviors in affecting response rates. Rather, the findings suggest further research and analysis to explore the relationships between specific behaviors and their application on the one hand, and interviewer-level response rates on the other. We feel that this line of inquiry has merit, and are working toward a fuller understanding of the role of interviewer experience, expectations and behavior in survey participation.

ACKNOWLEDGEMENTS

This work was supported by the Bureau of the Census, Bureau of Labor Statistics, Bureau of Justice Statistics, and the National Center for Health Statistics. Views expressed are those of the authors and do not necessarily reflect those of the Bureau of the Census or any other organization. The authors wish to thank Lorraine McCall for her assistance with this research. The reviewers are also thanked for their valuable suggestions.

APPENDIX A

VARIABLES USED IN ANALYSES

The creation of the variables used in the analyses are summarized here. Copies of the questionnaire can be obtained from the authors.

Dependent variable

Response rate: This is the response rate obtained by each interviewer for the six-month period in question, expressed as a percentage.

Assignment area

Population density: Population density (persons per square mile).

Crime rate: Crime rate (crimes per 100,000 population).

Percent 65 or older: Percentage of population 65 years of age and older.

Percent under 5: Percentage of population under 5 years of age.

Household size: Average household size.

Survey

Set of dummies to indicate which survey each interviewer works on:

HIS: Does interviewer work on the Health Interview Survey.

1 = Yes
0 = No

NCS: Does interviewer work on the National Crime Survey.

1 = Yes
0 = No

CE: (the Consumer Expenditure Survey) is thus the omitted category.

Interviewer experience

Tenure: Measured in days of service employed at the Census Bureau as an interviewer, rescaled to fractional years.

Breadth of experience: A count of the number of different survey organizations for which an interviewer has worked.

Interviewer expectations

Confidentiality: Interviewers were asked whether they thought there were any situation under which the Census Bureau would give individual survey response to any of a number of agencies (FBI, CIA, INS, IRS, state and local government agencies).

1 = High confidentiality belief (Census Bureau would not give responses to any of these agencies).
0 = Low confidentiality belief (Census Bureau would give responses to one or more of the agencies).

Rate/quality: Trade-off between response rate and data quality. Which one of the following statements comes closest to how you feel as an interviewer:

1 = It's better to persuade a reluctant respondent to participate than to accept a refusal.
0 = It's better to accept a refusal from a reluctant respondent.

Efficacy: Interviewers were asked the extent to which they agreed or disagreed with the following statement: With enough effort, I can convince even the most reluctant respondent to participate.

Four-point ordinal scale, 1 = strongly disagree, 4 = strongly agree. High score indicates greater belief in self-efficacy.

Interviewer behaviors

Authority: Interviewers were asked how often they left various materials (request for appointment, copy of the advance letter, *etc.*) at respondents' home when they found no-one at home. The responses to these questions were combined to form a scale of frequency of use of these authority-enhancing materials. High score indicates greater use of authority.

Reciprocation: How often do you make a point of complimenting something about respondent's home or personal appearance?
1 = Always, sometimes
0 = Rarely, never

Social proof: How often do you say "Most people enjoy doing the interview"?
1 = Always, sometimes
0 = Rarely, never

Saliency: How often do you explain to respondents how the survey results could affect them personally?
1 = Always, sometimes
0 = Rarely, never

Scarcity: How often do you tell a respondent that the interview must be completed by a certain date?
1 = Always, sometimes
0 = Rarely, never

Consistency: Before a respondent has shown any sign of cooperating, how often do you begin asking the survey questions?
1 = Always, sometimes
0 = Rarely, never

Repertoire: In an open-ended question, interviewers were asked to list all things they usually do to persuade reluctant respondent to participate. A count of the number of distinct things mentioned serves as an indicator of the repertoire of techniques available.

Tailoring: In a series of 15 behavior items, interviewers responded whether they always, sometimes, rarely or never performed such behavior. An indicator of tailoring in the application of various persuasion techniques is obtained by counting the number of times an interviewer used the middle categories (sometimes or rarely) to these questions. A high score indicates greater use of tailoring.

REFERENCES

BROWN, P.R., and BISHOP, G.F. (1982). Who refuses and resists in telephone surveys? Some new evidence. Paper presented at the MAPOR Annual Conference.

CANNELL, C.F. (1964). Factors affecting the refusal rate in interviewing. Ann Arbor: Survey Research Center (unpublished working paper).

CIALDINI, R.B. (1984). *Influence; The New Psychology of Modern Persuasion*. New York: Quill.

CIALDINI, R.B. (1990). Deriving psychological concepts relevant to survey participation from the literatures on compliance, helping, and persuasion. Paper presented at the Workshop on Household Survey Nonresponse, Stockholm.

DURBIN, J., and STUART, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *Journal of the Royal Statistical Society, Series A*, 114, 163-206.

GOWER, A.R. (1979). Non-response in the Canadian Labour Force Survey. *Survey Methodology*, 5, 29-58.

GOYDER, J. (1987). *The Silent Minority; Nonrespondents on Sample Surveys*. Boulder, CO: Westview Press.

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

GROVES, R.M., CIALDINI, R.B., and COUPER, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly* (forthcoming).

GROVES, R.M., and FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.

GROVES, R.M., and KAHN, R.L. (1979). *Surveys by Telephone*. New York: Academic Press.

HAWKINS, D.F. (1975). Estimation of nonresponse bias. *Sociological Methods and Research*, 3, 461-488.

HERZOG, A.R., and RODGERS, W.L. (1988). Age and response rates to interview sample surveys. *Journals of Gerontology*, 43, S200-S205.

HOUSE, J.S., and WOLF, S. (1978). Effects of urban residence on interpersonal trust and helping behavior. *Journal of Personality and Social Psychology*, 36, 1029-1043.

INDERFURTH, G.P. (1972). Investigation of Census Bureau interviewer characteristics, performance and attitudes: A summary. U.S. Bureau of the Census: Working Paper 34.

KORTE, C., and KERR, N. (1975). Responses to altruistic opportunities in urban and nonurban settings. *Journal of Social Psychology*, 95, 183-184.

LIEVESLEY, D. (1988). Unit non-response in interview surveys. London: Social and Community Planning Research (unpublished working paper).

PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.

RAUTA, I. (1985). A comparison of the census characteristics of respondents and nonrespondents to the 1981 General Household Survey (GHS). *Statistical News*, 71, 12-15.

SCHYBERGER, B.W. (1967). A study of interviewer behavior. *Journal of Marketing Research*, 4, 32-35.

SINGER, E., FRANKEL, M.R., and GLASSMAN, M.B. (1983). The effect of interviewer characteristics and expectation on response. *Public Opinion Quarterly*, 47, 68-83.

SMITH, T.W. (1983). The hidden 25 percent: An analysis of nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly*, 47, 386-404.

STEEH, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.

U.S. BUREAU OF THE CENSUS (1988). *County and City Data Book, 1988*. U.S. Government Printing Office.

WEAVER, C.N., HOLMES, S.L., and GLENN, N.D. (1975). Some characteristics of inaccessible respondents in a telephone survey. *Journal of Applied Psychology*, 60, 260-262.

WILCOX, J.B. (1977). The interaction of refusal and not-at-home sources of nonresponse bias. *Journal of Marketing Research*, 14, 592-597.

Survey Methodology, December 1992
Vol. 18, No. 2, pp. 279–292
Statistics Canada

A Multivariate Procedure Towards Composite Estimation of Consumer Expenditure for the U.S. Consumer Price Index Numbers

P. LAHIRI and WENYU WANG[1]

ABSTRACT

We consider the problem of estimating the "cost weights" and "relative importances" of different item strata for the local market basket areas. The estimation of these parameters is needed to construct the U.S. Consumer Price Index Numbers. We use multivariate models to construct composite estimators which combine information from relevant sources. The mean squared errors (MSE) of the proposed and the existing estimators are estimated using the repeated half samples available from the survey. Based on our numerical results, the proposed estimators seem to be superior to the existing estimators.

KEY WORDS: Consumer expenditure; Composite estimation; Consumer Price Index; Cost weight; Diary survey; Half sample; Laspeyres Index; Mean squared error; Synthetic estimation.

1. INTRODUCTION

The U.S. Consumer Price Index (CPI) is an indicator of price changes for a set of items, goods and services, whose quantity and quality are fixed over a period of time. The U.S. Bureau of Labor Statistics (BLS) computes a number of consumer price indices each month for various geographical areas, consumer units and item classification (*vide* BLS Handbook of Methods 1988).

The smallest group of item classification for which the BLS computes the CPI is known as an "item stratum". It is a prespecified set of consumer goods and services, *e.g.*, fresh whole milk, which can be purchased in the retail market during a "base period" by a specified set of consumer units. A consumer unit may consist of all members of a particular household related by blood, marriage, adoption, or other legal arrangements. A number of item strata constitutes an expenditure class (*e.g.*, dairy products).

The U.S. is divided into eight major areas for sampling purposes. A major area may be either "self-representing" or "non-self-representing" and belongs to one of the four regions (Northeast, Midwest, South and West). A self-representing area consists of all large cities within a region. A non-self-representing area generally consists of a county or a group of contiguous counties. For publication purposes, a major area is further divided into a number of "market basket areas" or "publication areas".

The Laspeyres formula used by the BLS to compute the CPI for a given area and an expenditure class (say, E) is defined below. Let

P_{it} = the average price of all items in the ith item stratum at time t ($t = 0,T$),

Q_{i0} = the quantity of all items in the ith item stratum purchased at time $t = 0$ (base period).

[1] P. Lahiri, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588 0323, USA. Wenyu Wang, SUNY Health Science Center at Brooklyn, Box 1203, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.

Then the Laspeyres index at time $t = T$ is given by

$$I_T = \sum_{i \in E} Q_{i0} P_{iT} \Big/ \sum_{i \in E} Q_{i0} P_{i0}$$

$$= \frac{\sum_{i \in E} C_i (P_{iT}/P_{i0})}{\sum_{i \in E} C_i}$$

$$= \sum\nolimits_{i \in E} R_i (P_{iT}/P_{i0}),$$

where

$C_i = Q_{i0} P_{i0}$ = total expenditure for all items in the ith item stratum at $t = 0$,

$R_i = C_i / \sum_{i \in E} C_i$ = proportion of total expenditure spent on the ith item stratum at $t = 0$.

The quantities C_i and R_i are referred to as the "cost weight" and "relative importance" of the ith item stratum within the expenditure class, E.

The Bureau of Labor Statistics computes the consumer price indices using data from the U.S. Consumer Expenditure Survey (CES). The survey has two different components - Diary survey and Interview survey, each having separate sampling schemes and questionnaires. In this paper we consider data from the Diary survey only. The sampling design selects all the primary stage units (PSU's) within a particular self-representing area with certainty. But only a sample of PSU's is selected for a particular non-self-representing area according to a probability sampling scheme. From each selected PSU, a sample of consumer units (CU's) is selected again using some probability sampling design. Each respondent keeps a diary of expenditures on various items for two consecutive 1-week periods. For a detailed account on the CPI and CES, the reader is referred to the BLS Handbook of Methods (1988).

The efficiency of the traditional sample survey estimators of the cost weight and relative importance of an item stratum at the publication area level is generally very low compared to their efficiency at a larger area (*e.g.*, major area) level. This is due to the fact that only a few consumer units are available from a given publication area. Thus, there is a need to improve the traditional estimator by borrowing strength from related resources. Marks (1978) and Cohen and Sommers (1984) considered certain composite estimators which pool information from related areas. Ghosh and Sohn (1990) obtained composite estimators of the cost weight and relative importance using an empirical Bayes approach.

The current procedure used by the Bureau of Labor Statistics consists of several steps. First composite estimators of the relative importances are obtained using a method suggested by Cohen and Sommers (1984). The estimators of the cost weights are then obtained from these estimators of the relative importances using an iterated "raking" procedure. The final estimates of the cost weights for the entire expenditure class and for the major area are identical to the corresponding preliminary estimates. One reason for ensuring this "data consistency" by raking may be due to the fact that the performances of the preliminary estimators are generally satisfactory at a higher level of aggregation compared to their performances at a lower level. At the last step, the final estimators of the relative importances are obtained directly from the final cost weight estimators by division.

Unlike earlier authors, we use the correlations between the item strata in proposing our composite estimators in Section 2. The shrinkage factor of the composite estimator obtained by minimizing the mean squared error within an appropriate class of estimators involves some unknown parameters. These unknown parameters are estimated using the balanced repeated replications available from the survey. The estimator proposed by Cohen and Sommers (1984) turns out to be a special case of our estimator if one assumes that the preliminary estimators are all uncorrelated.

In Section 2 we concentrate our attention to the estimation of the cost weight of an item stratum for a publication area. However, we can obtain estimators of the cost weights at a higher level of aggregation (*e.g.*, expenditure class for a publication area, *etc.*) by appropriate summation. From our study, it turns out that in terms of the mean squared error criterion these estimators always perform better than the corresponding preliminary estimators and hence better than the BLS estimators (note that due to the raking procedure the BLS estimators are identical to the preliminary estimators at higher levels of aggregation).

In Section 3 we propose a composite estimator of relative importance of an item stratum at the publication area level. Instead of using the preliminary estimators of the cost weights we use the preliminary estimators of the relative importances for all the item strata belonging to the expenditure class under consideration. The preliminary estimators of relative importances of all the item strata within an expenditure class add up to unity. Thus, the variance covariance matrix of the preliminary estimators is singular and this makes the problem different from the problem of estimation of the cost weights. Our procedure deletes one item stratum in an optimal manner and thus avoids the problem of singularity of the variance covariance matrix of the preliminary estimators. Our numerical results show that in terms of the mean squared error criterion the proposed estimator is always the best among all the rival estimators considered.

In Section 4, we present all the numerical results. We have evaluated different estimators of the cost weight and relative importance based on estimated mean squared error obtained by using the balanced repeated half samples (see McCarthy 1969, Ghosh and Sohn 1990). Based on our results, the proposed estimators seem to be superior to all the rival estimators considered in the paper.

2. ESTIMATION OF THE COST WEIGHT

Let X_{ijl} be the average of two consecutive weeks of expenditure for all the items in the ith item stratum by the lth consumer unit belonging to the jth publication area within a particular major area ($i = 1, \ldots, I; j = 1, \ldots, m; l = 1, \ldots, n_j$). Let W_{jl} be the sampling weight attached to the lth consumer unit in the jth publication area ($j = 1, \ldots, m; l = 1, \ldots, n_j$). This represents a number of consumer units in the population and is obtained by the Census Bureau using a complex procedure which takes into account various factors such as inclusion probabilities, nonresponse, *etc.* In this section, we consider estimation of θ_{ij}, the true average weekly expenditure per consumer unit for the ith item stratum and jth publication area. The cost weight is simply defined as $N_j\theta_{ij}$, where N_j denotes the total number of consumer units in the jth publication area. The preliminary estimator of θ_{ij} is given by

$$Y_{ij} = \sum_{l=1}^{n_j} W_{jl}X_{ijl} \Big/ \sum_{l=1}^{n_j} W_{jl}, \ (i = 1, \ldots, I; \ \ j = 1, \ldots, m). \tag{2.1}$$

Similarly, the corresponding estimator for the major area is given by

$$Y_{i\cdot} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl} \Big/ \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}. \tag{2.2}$$

The variability of $Y_{i\cdot}$ is much lower than that of Y_{ij}. Thus, a composite estimator of θ_{ij} which increases the precision is needed. Let $Y_j = (Y_{1j}, \ldots, Y_{Ij})'$ and $\theta_j = (\theta_{ij}, \ldots, \theta_{Ij})'$, $j = 1, \ldots, m$. Let V_j be the true variance covariance matrix of Y_j, $(j = 1, \ldots, m)$. Under a synthetic assumption, *i.e.*, $\theta_j = \mu$, a $I \times 1$ column vector, $(j = 1, \ldots, m)$, the best estimator of θ_j is given by

$$\tilde{\mu} = \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \sum_{j=1}^{m} V_j^{-1} Y_j, \tag{2.3}$$

which is obtained by minimizing $\sum_{j=1}^{m} (Y_j - \mu)' V_j^{-1} (Y_j - \mu)$ with respect to μ. The synthetic assumption, however, is hardly satisfied. In the other extreme when there is absolutely no similarity between the θ_j's, it is appropriate to take Y_j as an estimator of θ_j. When the real situation is in between these two extremes one may take a composite estimator given by

$$\hat{\theta}_{ij}(a_{ij}) = (1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu}, \tag{2.4}$$

where a_{ij}'s are constants $(0 \leq a_{ij} \leq 1)$, e_i is a $I \times 1$ column vector having 1 for the ith elements and 0 for the others.

We obtain a_{ij} by minimizing the mean squared error

$$E[\{(1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu} - \theta_{ij}\}^2 \mid \theta_{ij}] \tag{2.5}$$

with respect to a_{ij}. The optimal choice is given by

$$\tilde{a}_{ij} = \frac{e_i' \left[V_j - \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \right] e_i}{E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]}. \tag{2.6}$$

Thus, the optimal estimator of θ_{ij} in the class described by (2.4) is given by

$$\tilde{\theta}_{ij} = (1 - \tilde{a}_{ij}) Y_{ij} + \tilde{a}_{ij} e_i' \tilde{\mu}. \tag{2.7}$$

Remark 1: In the derivation of the optimal estimator $\tilde{\theta}_{ij}$, the quantities V_j, $(j = 1, \ldots, m)$ and $E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]$ are assumed to be fixed and known.

Remark 2: The estimator proposed by Cohen and Sommers (1984) can be obtained from $\tilde{\theta}_{ij}$ as a special case when

$$V_j = \left(\sum_{l=1}^{n_j} W_{jl} \right)^{-1} \text{Diag}(\sigma_1^2, \ldots, \sigma_I^2).$$

Note that according to their assumption the correlation between any two item strata is zero which appears to be very restrictive from our study.

Remark 3: Note that using a familiar matrix inversion result (see Rao 1973),

$$V_j - \left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1} = V_j\left[V_j + \left(\sum_{s\neq j} V_s^{-1}\right)^{-1}\right]^{-1} V_j$$

which is positive definite. Also,

$$E[(Y_{ij} - e_i'\tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m] = e_i'V_j\left[V_j + \left(\sum_{s\neq j} V_s^{-1}\right)^{-1}\right]^{-1} V_j e_i + \left[\theta_{ij} - e_i'\left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1}\left(\sum_{j=1}^{m} V_j^{-1}\theta_j\right)\right]^2.$$

Also, when $\theta_j = \mu$, one gets $\tilde{a}_{ij} = 1$ and thus $\tilde{\theta}_{ij} = e_i'\tilde{\mu}$. Otherwise the size of the shrinkage factor depends on the size of

$$\left[\theta_{ij} - e_i'\left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1}\left(\sum_{j=1}^{m} V_j^{-1}\theta_j\right)\right]^2.$$

The larger the distance of θ_{ij} from $e_i'(\sum_{j=1}^{m} V_j^{-1})^{-1}(\sum_{j=1}^{m} V_j^{-1}\theta_j)$ the smaller is the size of $\tilde{a}_{ij}$. This means that if a particular area is very different from the general nature of all the areas then our procedure will give less weight on the synthetic part of the estimator. This explains the great deal of variation of the shrinkage factors in Table 1.

We shall estimate $\tilde{a}_{ij}$ using the 20 balanced repeated half samples available from the survey. Let $w_{jl}^{(k)}$ denote the weight assigned to the lth consumer unit of the jth area for the kth replication ($j = 1, \ldots, m$; $l = 1, \ldots, n_j$; $k = 1, \ldots, 20$). These replicated weights are constructed by the Census Bureau using a complex procedure. For any replication, approximately half the consumer units receive zero weights and the remaining consumer units receive positive weights.

Table 1

Shrinkage Factors $\hat{a}_{ij}$ in West Non-Self-Representing Area

i \ j	1	2	2
1	0.8479225	0.7057626	0.9214804
2	0.8434894	0.5692695	0.8092725
3	0.0969009	0.0786758	0.6953904
4	0.4446537	0.5444809	1
5	0.6999551	0.3460123	0.5487382
6	0.0318442	0.4981756	0.2598752

Define

$$\hat{a}_{ij}^{*} = \frac{e_i'\left[\hat{V}_j - \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\right]e_i}{\frac{1}{20}\sum_{k=1}^{20}[Y_{ij}^{(k)} - e_i'\hat{\mu}^{(k)}]^2},$$

$$\hat{\mu} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1} Y_j\right],$$

$$\hat{\mu}^{(k)} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1} Y_j^{(k)}\right],$$

$$Y_{ij}^{(k)} = \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} \Big/ \sum_{l=1}^{n_j} W_{jl}^{(k)},$$

$$Y_j^{(k)} = [Y_{1j}^{(k)}, \ldots, Y_{Ij}^{(k)}]',$$

$$\hat{V}_j = 1/20 \sum_{k=1}^{20} [Y_j^{(k)} - Y_j][Y_j^{(k)} - Y_j]'.$$

Then we propose the following estimator of θ_{ij}:

$$\hat{\theta}_{ij}^{*} = (1 - \hat{a}_{ij}^{*}) Y_{ij} + \hat{a}_{ij}^{*} e_i'\hat{\mu}. \tag{2.8}$$

Remark 4: Using argument given in Remark 3, $\hat{a}_{ij}^{*} \geq 0$. But it is possible that sometimes $\hat{a}_{ij}^{*}$ may exceed unity. Thus, we consider the following estimator:

$$\hat{\theta}_{ij} = (1 - \hat{a}_{ij}) Y_{ij} + \hat{a}_{ij} e_i'\hat{\mu}, \tag{2.9}$$

where $\hat{a}_{ij} = \min[1, \hat{a}_{ij}^{*}]$.

In Table 1, we give values of $\hat{a}_{ij}$ for the West non-self-representing area.

3. ESTIMATION OF THE RELATIVE IMPORTANCE

Let $R_{ij} = Y_{ij} / \sum_{i=1}^{I} Y_{ij}$ be the preliminary estimator of the relative importance $r_{ij} = \theta_{ij} / \sum_{i=1}^{I} \theta_{ij}$, $(i = 1, \ldots, I; j = 1, \ldots, m)$. Let $R_j = (R_{1j}, \ldots, R_{Ij})'$, $(j = 1, \ldots, m)$. Since $\sum_{i=1}^{I} R_{ij} = 1$, $(j = 1, \ldots, m)$, the variance covariance matrix of R_j is singular. Thus, the method described in Section 2 is not directly applicable to this situation. In order to avoid this singularity problem, we delete one item stratum from the expenditure class under consideration. Without any loss of generality, let the Ith item stratum be deleted. Then apply the procedure described in Section 2 to obtain the following estimator for r_{ij}, $(i = 1, \ldots, I - 1; j = 1, \ldots, m)$

$$\hat{r}_{ij}^{*} = (1 - \hat{d}_{ij}) R_{ij} + \hat{d}_{ij} e_i'\hat{\xi}, \tag{3.1}$$

where

$$\hat{d}_{ij} = \min[1, \hat{d}^*_{ij}],$$

$$\hat{d}^*_{ij} = \frac{e_i'\left[\hat{D}_j - \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\right]e_i}{\frac{1}{20}\sum_{k=1}^{20}[R_{ij}^{(k)} - e_i'\hat{\xi}^{(k)}]^2},$$

$$R_{ij}^{(k)} = Y_{ij}^{(k)} \Big/ \sum_{i=1}^{I} Y_{ij}^{(k)},$$

$$R_j^{(k)} = (R_{1j}^{(k)}, \ldots, R_{I-1,j}^{(k)})',$$

$$\hat{D}_j = \frac{1}{20}\sum_{k=1}^{20}(R_j^{(k)} - R_j)(R_j^{(k)} - R_j)',$$

$$\hat{\xi}^{(k)} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{D}_j^{-1}R_j^{(k)}\right],$$

$$\hat{\xi} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{D}_j^{-1}R_j\right].$$

For $i = I$,

$$\hat{D}_{II}^{(j)} = \frac{1}{20}\sum_{k=1}^{20}(\hat{R}_{Ij}^{(k)} - R_{Ij})^2,$$

$$R_{I\cdot} = \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}R_{Ij}\right] \Big/ \sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1},$$

$$\hat{d}_{Ij} = \min[1, \hat{d}^*_{Ij}],$$

$$\hat{d}^*_{Ij} = \frac{\hat{D}_{II}^{(j)} - \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}\right]^{-1}}{\frac{1}{20}\sum_{k=1}^{20}[R_{Ij}^{(k)} - R_{I\cdot}^{(k)}]^2},$$

$$R_{I\cdot}^{(k)} = \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}R_{Ij}^{(k)}\right] \Big/ \sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}.$$

We estimate r_{Ij} by a univariate procedure which yields the following estimator of r_{Ij}, $(j = 1, \ldots, m)$:

$$\hat{r}^*_{Ij} = (1 - \hat{d}_{Ij})R_{Ij} + \hat{d}_{Ij}R_I.$$

We obtain the final estimator of r_j as $\hat{r}_j = (\hat{r}_{1j}, \ldots, \hat{r}_{Ij})'$, where $\hat{r}_{ij} = \hat{r}^*_{ij} / \sum_{i=1}^{I} \hat{r}^*_{ij}$. There are I possible choices of deleting one item stratum. We choose the combination which yields the smallest average (over item strata) estimated MSE. One may obtain an alternative estimator of r_{Ij} by subtracting $\sum_{i=1}^{I-1} r_{ij}$ from unity. However, according to the procedure, there is a positive probability that r_{Ij} estimate is negative.

4. NUMERICAL RESULTS

In this section, we evaluate various estimators of the cost weight and relative importance based on estimated mean squared error. We consider four rival estimators: the preliminary estimator, estimator proposed by Cohen and Sommers (1984), the estimator currently used by the BLS and the empirical Bayes estimator considered recently by Ghosh and Sohn (1990). The Cohen-Sommers estimator of the cost weight (before raking) is given by

$$\begin{aligned}
\hat{\theta}^{CS}_{ij} &= \hat{\theta}^{CS*}_{ij} \quad \text{if} \quad |\hat{\theta}^{CS*}_{ij} - Y_{ij}| < c \cdot \mathrm{sd}(Y_{ij}) \\
&= Y_{ij} + c \cdot \mathrm{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}^{CS*}_{ij} \geq Y_{ij} + c \cdot \mathrm{sd}(Y_{ij}) \\
&= Y_{ij} - c \cdot \mathrm{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}^{CS*}_{ij} \leq Y_{ij} - c \cdot \mathrm{sd}(Y_{ij})
\end{aligned}$$

where

$$\hat{\theta}^{CS*}_{ij} = (1 - \hat{a}^{CS}_{ij}) Y_{ij} + \hat{a}^{CS}_{ij} Y_{i\cdot},$$

$$\hat{a}^{CS}_{ij} = \min\left[1, (1 - N_j/N)\left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y_{ij})^2\right] \Big/ \left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y^{(k)}_{i\cdot})^2\right]\right],$$

$$Y^{(k)}_{i\cdot} = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W^{(k)}_{jl} X_{ijl} \Big/ \sum_{j=1}^{m}\sum_{l=1}^{n_j} W^{(k)}_{jl},$$

N_j = total number of consumer units in the population for the jth publication area,

$$N = \sum_{j=1}^{m} N_j,$$

$$\mathrm{sd}(Y_{ij}) = \sqrt{\left\{\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y_{ij})^2\right\}},$$

c = a safety factor determined by the BLS (see Table 2).

Table 2

Values of the Safety Factor c for the Major Areas

Major Area	NCNS	NCSR	NENS	NESR	SSNS	SSSR	WWNS	WWSR
	1	2	3	4	5	6	7	8
c	1.0	.5	1.0	.5	3.0	.25	1.0	.5

NCNS: North Central (Midwest) non-self-representing.
NCSR: North Central self-representing.
NENS: North East non-self-representing.
SSNS: South non-self-representing.
SSSR: South self-representing.
WWNS: West non-self-representing.
WWSR: West self-representing.

Their estimator for the relative importance is given by

$$\hat{r}_{ij}^{CS} = \hat{r}_{ij}^{CS*} \quad \text{if} \quad |\hat{r}_{ij}^{CS*} - R_{ij}| \leq c \cdot \text{sd}(R_{ij})$$

$$= R_{ij} + c \cdot \text{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{CS*} \geq R_{ij} + c \cdot \text{sd}(R_{ij})$$

$$= R_{ij} + c \cdot \text{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{CS*} \leq R_{ij} - c \cdot \text{sd}(R_{ij}),$$

where

$$\hat{r}_{ij}^{CS*} = (1 - \hat{d}_{ij}^{CS})R_{ij} + \hat{d}_{ij}^{CS}R_{i\cdot}^{CS},$$

$$R_{i\cdot}^{CS} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}X_{ijl} \Big/ \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}X_{ijl},$$

$$\hat{d}_{ij}^{CS} = \hat{d}_{ij}^{CS*} \quad \text{if} \quad 0 < \hat{d}_{ij}^{CS*} < 1,$$

$$= 0 \quad \text{if} \quad \hat{d}_{ij}^{CS*} \leq 0,$$

$$= 1 \quad \text{if} \quad \hat{d}_{ij}^{CS*} \geq 1,$$

$$\hat{d}_{ij}^{CS*} = \frac{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})^2 - \frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})(R_{i\cdot}^{CS(k)} - R_{i\cdot}^{CS})}{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{i\cdot}^{CS(k)})^2},$$

$$R_{i\cdot}^{CS(k)} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)}X_{ijl} \Big/ \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)}X_{ijl},$$

$$\mathrm{sd}(R_{ij}) = \sqrt{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})^2}.$$

Since $\sum_{i=1}^{I} \hat{r}_{ij}^{CS} \neq 1$, for our comparison purpose, we have divided $\hat{r}_{ij}^{CS}$ by $\sum_{i=1}^{I} \hat{r}_{ij}^{CS}$.

The procedure currently used by the Bureau of Labor Statistics (see United States Department of Labor 1988) consists of a number of steps.

Step 1: Obtain an estimator of the cost weight as follows:

$$\hat{\theta}_{ij}^{CS(1)} = \hat{r}_{ij}^{CS} \sum_{i=1}^{I} Y_{ij}.$$

Step 2: Final estimator of θ_{ij} is obtained from $\hat{\theta}_{ij}^{CS(1)}$ using a "raking" procedure. The final estimator, denoted by $\hat{\theta}_{ij}^{BLS}$, satisfies the following two conditions:

$$\sum_{i=1}^{I} \hat{\theta}_{ij}^{BLS} = \sum_{i=1}^{I} Y_{ij},$$

$$\sum_{j=1}^{m} N_j \hat{\theta}_{ij}^{BLS} = \sum_{j=1}^{m} N_j Y_{ij}.$$

Step 3: Finally an estimator for the relative importance is obtained as follows:

$$\hat{r}_{ij}^{BLS} = \hat{\theta}_{ij}^{BLS} \Big/ \sum_{i=1}^{I} \hat{\theta}_{ij}^{BLS}.$$

In our numerical work, we have estimated N_j by $\sum_{l=1}^{n_j} W_{jl}$.

The MSE of an estimator e_{ij} of θ_{ij} is given by:

$$\begin{aligned}\mathrm{MSE} &= E(e_{ij} - \theta_{ij})^2 \\ &= E(e_{ij} - Y_{ij})^2 - V(Y_{ij}) + 2\,\mathrm{Cov}(e_{ij}, Y_{ij}),\end{aligned}$$

where it is assumed $E(Y_{ij} \mid \theta_{ij}) = \theta_{ij}$. The above formula is given in Cohen and Sommers (1984). As in the Ghosh and Sohn (1990) we estimate the three terms by the balanced repeated half samples available from the survey. For example,

$$E(e_{ij} - Y_{ij})^2 \doteq \frac{1}{20}\sum_{k=1}^{20}(e_{ij}^{(k)} - Y_{ij}^{(k)})^2,$$

$$V(Y_{ij}) \doteq \frac{1}{20}\sum_{k=1}^{20}(Y_{ij}^{(k)} - Y_{ij})^2,$$

Table 3

Average Estimated MSE's for Different Estimators of θ_{ij}

Major Area	Average Estimated MSE of				
	Y_{ij}	$\hat{\theta}_{ij}^{GS}$	$\hat{\theta}_{ij}^{CS}$	$\hat{\theta}_{ij}^{BLS}$	$\hat{\theta}_{ij}$
NCNS	.020047	.011549 (22)	.009342 (53)	.014885 (25)	.009428 (52)
NCSR	.036620	.024783 (32)	.016017 (56)	.023627 (35)	.016155 (55)
NENS	.018162	.013299 (26)	.007327 (59)	.013046 (28)	.005504 (69)
NESR	.052883	.051100 (3)	.038911 (26)	.045610 (13)	.028958 (45)
SSNS	.021757	.013146 (39)	.009954 (54)	.014415 (33)	.006418 (70)
SSSR	.047500	.028984 (38)	.031743 (33)	.044238 (6)	.009270 (80)
WWNS	.052387	.029938 (42)	.017433 (66)	.030069 (42)	.010849 (79)
WWSR	.018223	.033529 (−83)	.009925 (45)	.014898 (18)	.005761 (68)

Note: The figures in the parenthesis represents percent improvement over the preliminary estimator, Y_{ij}.

$$\text{Cov}(e_{ij}, Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - e_{ij})(Y_{ij}^{(k)} - Y_{ij}).$$

In the above $e_{ij}^{(k)}$ is the estimator e_{ij} based on the kth half sample ($k = 1, \ldots, 20$). For example,

$$\hat{\theta}_{ij}^{CS(k)} = (1 - \hat{a}_{ij}^{CS}) Y_{ij}^{(k)} + \hat{a}_{ij}^{CS} Y_{i\cdot}^{(k)},$$

$$\hat{\theta}_{ij}^{(k)} = (1 - \hat{a}_{ij}) Y_{ij}^{(k)} + \hat{a}_{ij} e_i' \hat{\mu}^{(k)}.$$

We obtain $\hat{\theta}_{ij}^{BLS(k)}$ by the multistep procedure used to obtain $\hat{\theta}_{ij}^{BLS}$ where we replace Y_{ij}, R_{ij}, $\hat{r}_{ij}^{CS}$ by $Y_{ij}^{(k)}$, $R_{ij}^{(k)}$ and $\hat{r}_{ij}^{CS(k)}$ respectively. Note that the above procedure does not take into account the variation due to the estimation of the coefficients (*i.e.*,a_{ij}'s) in the composite estimators. Cohen and Sommers (1984) recommended the use of half samples of half samples, or quarter samples to capture this additional variability. We could not use their procedure since our dataset did not contain these quarter samples.

The data we analyze arise out of 1982-83 Consumer Expenditure Survey (Diary survey). The expenditure class we consider is dairy products. There are in all six item strata in this class. They are (1) fresh whole milk, (2) other fresh milk and cream, (3) butter, (4) cheese, (5) ice cream and related products, and (6) other dairy products.

The MSE's of all the estimators considered are estimated for each publication area and item stratum. In Table 3 we report the average estimated MSE's of the estimators of θ_{ij}, the average being taken over all the item strata and all the publication areas within a major area. Notice that all the composite estimators except the one proposed by Ghosh and Sohn (1990) are better than the preliminary estimator for all the major areas in the average MSE sense. Both θ_{ij}^{CS} and $\hat{\theta}_{ij}$ are better than $\hat{\theta}_{ij}^{BLS}$. Our proposed estimator $\hat{\theta}_{ij}$ is better than $\hat{\theta}_{ij}^{CS}$ in six out of eight major areas. In two major areas (NCNS and NCSR), $\hat{\theta}_{ij}^{CS}$ is better than $\hat{\theta}_{ij}$, but the difference is very negligible.

In Tables 4 and 5, we try to demonstrate that the raking procedure may not be necessary. In Table 4, the parameter of interest is $\sum_{i=1}^{I}\theta_{ij}$, the true cost weight for the expenditure class. Here, due to the "raking" procedure, $\sum_{i=1}^{I}\hat{\theta}_{ij}^{BLS} = \sum_{i=1}^{I} Y_{ij}$. We propose an alternative estimator as $\sum_{i=1}^{I}\hat{\theta}_{ij}$ and compare the average estimated MSE (over publication areas in a major area) with that of $\sum_{i=1}^{I} Y_{ij}$. In all the cases, we gain considerably.

Table 4

Average Estimated MSE's of Two Estimators of Average Consumer Expenditure for the Expenditure Class

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.12384	0.07969	36
NCSR	0.29819	0.13040	56
NENS	0.21658	0.07602	65
NESR	0.67486	0.20119	70
SSNS	0.21506	0.08303	61
SSSR	0.68415	0.06462	90
WWNS	0.35446	0.05175	85
WWSR	0.19292	0.05524	71

Table 5

Average Estimated MSE's of Two Estimators of Average Consumer Expenditure for the Major Area

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.008181	0.0045468	44
NCSR	0.003672	0.0031047	15
NENS	0.006174	0.0029128	53
NESR	0.011680	0.0056922	51
SSNS	0.007501	0.0036401	51
SSSR	0.004434	0.0013751	69
WWNS	0.008203	0.0022560	72
WWSR	0.002786	0.0007882	72

In Table 5, the parameter of interest is the cost weight of an item stratum for the major area. The preliminary estimator (identical to the BLS estimator due to the raking procedure) is $(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl} Y_{ij})/(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. Our estimation procedure can also generate estimators at the major area level. We propose the estimator as $\hat{\theta}_i. = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl}\hat{\theta}_{ij} / (\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. The average estimated MSE's for these two estimators are reported in Table 5. Here also our estimator is superior to the preliminary (BLS) estimator.

The results of Table 4 and 5 suggest that the data consistency step followed by the BLS may not be necessary. Indeed, it may be possible to improve the traditional estimators at higher levels of aggregation also.

Table 6 provides the average estimated MSE's (over all the item strata and publication areas in a major area) of various estimators of relative importance. Notice that as in Table 3, all the estimators other than $\hat{r}_{ij}^{GS}$ are better than the preliminary estimator $\hat{R}_{ij}$ for all the major areas. Our proposed estimator $\hat{r}_{ij}$ is the best among all the estimators considered.

Recently, Swanson (1992) has compared different methods of estimating cost weights for 12 of the approximately 70 expenditure classes in the CPI. His investigation shows that overall our proposed method is superior to all the rival methods.

Table 6

Average Estimated MSE's for Different Estimators of Relative Importance

Major Area	Average Estimated MSE of				
	R_{ij}	$\hat{r}_{ij}^{GS}$	$\hat{r}_{ij}^{CS}$	$\hat{r}_{ij}^{BLS}$	$\hat{r}_{ij}$
NCNS	.0006342	.00046480 (27)	.00033143 (48)	.00042130 (34)	.00018592 (71)
NCSR	.0009125	.00071967 (21)	.00040226 (56)	.00044815 (51)	.00021309 (77)
NENS	.0003588	.00026894 (25)	.00014146 (61)	.0001620 (55)	.00011105 (69)
NESR	.0004264	.00072001 (−69)	.00028862 (32)	.00030555 (28)	.00016744 (61)
SSNS	.0005071	.00033736 (33)	.00019352 (62)	.00021385 (58)	.00011925 (76)
SSSR	.0006564	.00048569 (26)	.00053173 (19)	.00053603 (18)	.00030979 (53)
WWNS	.0013709	.00086849 (37)	.00051474 (62)	.00061901 (55)	.00028519 (79)
WWSR	.0003540	.00070770 (−100)	.00021384 (40)	.00023255 (34)	.00013750 (61)

Note: The figure given in the parenthesis represents percent improvement over R_{ij}.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation (NSF) under grant SES-9001399, "On-Site Research to Improve the Quality of Labor Statistics." This research was conducted at the U.S. Bureau of Labor Statistics while the authors were participants in the American Statistical Association/Bureau of Labor Statistics Research Program, which is supported by the Bureau of Labor Statistics and through the NSF grant. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of Labor Statistics. We wish to thank Sylvia Leaver, Stuart Scott, Malay Ghosh, Richard Valliant, Stephen Miller, So Young Sohn, Paul Hsen, Adriana Silberstein for many valuable discussions. We also thank two referees and associate editor for helpful comments on an earlier version of this paper.

REFERENCES

COHEN, M.P., and SOMMERS, J.P. (1984). Evaluation of methods of composite estimation of cost weights for the CPI. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 466-471.

GHOSH, M. and SOHN, S.Y. (1990). An Empirical Bayes Approach Towards Composite Estimation of Consumer Expenditure. Technical Report, U.S. Bureau of Labor Statistics.

MARKS , H. (1978). Composite estimation techniques used for the CPIR weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 311-315.

McCARTHY P.J. (1969). Pseudoreplication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

SWANSON, D. (1992). An evaluation of 4 cost weight composite estimation methods for the CPI. Memorandum for Janet Williams, Chief, CPI Survey Research Branch, Statistical Methods Division, U.S. Bureau of Labor Statistics.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd Edition). New York: J. Wiley & Sons.

UNITED STATES DEPARTMENT OF LABOR (1988). *Handbook of Methods*. Bureau of Labour Statistics. Washington DC: U.S. Government Printing Office.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1992:

P. Ardilly, *INSEE*
M.G. Arellano, *Advanced Linkage Technologies of America*
K.G. Basavarajappa, *Statistics Canada*
T.R. Belin, *UCLA*
D.R. Bellhouse, *University of Western Ontario*
P. Biemer, *Research Triangle Institute*
D. Binder, *Statistics Canada*
P.D. Bourke, *University College, Cork, Ireland*
J.M. Brick, *Westat*
N.J. Carter, *California State University*
R.G. Carter, *Statistics Canada*
G.H. Choudhry, *Statistics Canada*
P.A. Cholette, *Statistics Canada*
M.P. Cohen, *U.S. National Center for Education Statistics*
B. Cox, *Research Triangle Institute*
E.B. Dagum, *Statistics Canada*
J.-C. Deville, *INSEE*
C. Dippo, *U.S. Bureau of Labor Statistics*
D. Drew, *Statistics Canada*
R.E. Fay, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
M. Frankel, *Baruch College, CUNY*
J. Gentleman, *Statistics Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R. Groves, *U.S. Bureau of the Census*
K.P. Hapuarachchi, *Statistics Canada*
M.A. Hidiroglou, *Statistics Canada*
D. Holt, *University of Southampton*
P. Jagers, *Chalmers and Gothenburg Universities*
G. Kalton, *University of Michigan*
P.S. Kott, *NASS/U.S. Department of Agriculture*
R.A. Kulka, *University of Chicago*
P. Lahiri, *University of Nebraska*
G. Lagrange, *Statistics Canada*
K.C. Land, *Duke University*
J.M. Landwehr, *AT & T Bell Laboratories*
N. Laniel, *Statistics Canada*
P. Lavallée, *Statistics Canada*
H. Lee, *Statistics Canada*
F. Maranda, *Statistics Canada*
M. March, *Statistics Canada*
G.D. Meeden, *University of Minnesota*
W.J. Mitofsky, *Voter Research and Surveys*
H.B. Newcombe, *Consultant*
W.L. Nicholls II, *U.S. Bureau of the Census*
D. Norris, *Statistics Canada*
D. Northrup, *Coda Inc.*
C. O'Muircheartaigh, *London School of Economics and Political Science*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistics Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
D. Royce, *Statistics Canada*
D. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
N.C. Schaiffer, *University of Wisconsin - Madison*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York - Albany*
G.M. Shapiro, *U.S. Bureau of the Census*
C.J. Skinner, *University of Southampton*
B.D. Spencer, *Northwestern University*
N.L. Spruill, *U.S. Office of the Secretary of Defence*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
H. Tamura, *University of Washington*
A. Théberge, *Statistics Canada*
M.E. Thompson, *University of Waterloo*
R. Valliant, *U.S. Bureau of Labor Statistics*
K. Wachter, *University of California - Berkeley*
J. Waksberg, *Westat*
T. Wellens, *Zentrum für Umfragen, Zuma*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*

Acknowledgements are also due to those who assisted during the production of the 1992 issues: J. Beauseigle, S. Beauchamp and S. Lineger (Photocomposition), and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Special Issue on Measurement Errors in Surveys: Part I
Contents JOS 1992, Volume 8, Number 1

All inquires about submissions and subscriptions should be directed to the Chief Editor:

Lars Lyberg, U/SFI, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 41, No. 3, 1992

Printed in Great Britain at the Alden Press, Oxford

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.