Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data For Social And Economic Research

ANALYSIS OF COMPLEX SURVEY DATA USING INVERSE SAMPLING

J. N. K Rao[1], A. J. Scott[2], and E. Benhin[3]

ABSTRACT

Application of classical statistical methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for micro data. Inverse sampling (Hinkins, Oh and Scheuren, 1997) provides an alternative approach by undoing the complex survey data structures so that standard methods can be applied. Repeated subsamples with simple random sampling structure are drawn and each subsample analysed by standard methods and then combined to increase the efficiency. This method has the potential to preserve confidentiality of micro data, although computer-intensive. We present some theory of inverse sampling and explore its limitations. An estimating equations approach is proposed for handling complex parameters such as ratios and "census" regression parameters.

KEY WORDS: Confidentiality; Estimating Equations; Repeated Subsampling.

1. INTRODUCTION

How do practitioners deal with the complexities of survey data structures such as unequal selection probabilities, clustering and stratification? Adapting a quote from Hinkins, Oh and Scheuren (1997) (abbreviated HOS hereafter): "If your only tool is a hammer, every problem looks like a nail!"; the hammer available to most people is one of the big statistical packages such as SAS. Most people still just push their data through a standard program and ignore the survey data structures. This is in spite of the fact that a great deal of effort over the last two decades has been spent on developing methods to analyze survey data that take account of design features (see e.g., Skinner, Holt and Smith, 1989), and specialized programs such as SUDAAN or WesVar are now available to implement some of these methods.

An alternative to developing complex new tools is to work backwards: instead of tailoring the methods to fit the data, tailor the data to fit the methods. HOS developed an approach along these lines. Their basic idea is to avoid the pain caused by a complicated sample by choosing a subsample that has a simple random sample structure unconditionally (or at least has a structure that is considerably simpler to handle than the original sample). Obviously this involves some loss in efficiency, especially if the subsample is very much smaller than the original sample, as often turns out to be necessary. However, we can increase the efficiency by repeating the process independently many times and averaging the results.

Is it possible to produce subsamples with the desired properties? The answer is often "yes", although the resulting subsample size, m, might have to be small. HOS give algorithms for producing simple random inverse samples for a number of standard designs. We summarize some inverse sampling schemes in Section 2 for ready reference. These schemes include both exact and approximate methods in terms of matching simple random sampling. In this paper we look at some of the properties of the repeated inverse sampling procedures given in Section 2. In particular, we develop some basic theory of inverse sampling in Section 3, and illustrate some of the strengths and weaknesses of the procedure. In Section 4 we study the special case of a population total. We propose an estimating equations (EE) approach in Section 5 for handling complex parameters such as ratios and "census" regression parameters.

[1] School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B.
[2] Department of Statistics, University of Auckland, Auckland, New Zealand.
[3] Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6.

2. INVERSE SAMPLING ALGORITHMS

In this section we summarize some of the inverse sampling schemes, proposed by Hinkins, Oh and Scheuren (1997). These schemes include both exact and approximate methods in terms of matching simple random sampling (SRS) unconditionally. Suppose we have a sample s_0 of observations drawn from the finite population of size N according to a specified complex design. We wish to draw a subsample s^* of size m from s_0 such that the unconditional probability of s^*, $p(s^*)$, matches simple random sampling with $p(s^*) = 1/\binom{N}{m}$, either exactly or approximately. We have

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0)\, p(s^* \mid s_0), \tag{2.1}$$

where $p_0(s_0)$ is the probability of selecting s_0 and $p_0(s^* \mid s_0)$ is the conditional probability of choosing s^*. If $p_0(s^* \mid s_0)$ does not depend on s_0, then it follows from (2.1) that

$$p(s^* \mid s_0) = p_2(s^*) = \frac{p(s^*)}{\sum_{s_0 \supset s^*} p(s_0)}. \tag{2.2}$$

Denote the first-order and second-order inclusion probabilities corresponding to s^* and s_0 as (π_i^*, π_{il}^*) and (π_i, π_{il}) respectively, where $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/N(N-1), i \neq l$. Similarly, denote the conditional inclusion probabilities as $(\tilde{\pi}_i(s_0), \tilde{\pi}_{il}(s_0))$. If the conditional inclusion probabilities do not depend on s_0, then we write them as $(\tilde{\pi}_i, \tilde{\pi}_{il})$. It is readily seen that

$$\pi_i^* = \sum_{s_0 \ni i} p_0(s_0)\tilde{\pi}_i(s_0); \quad \pi_{il}^* = \sum_{s_0 \ni i,l} p_0(s_0)\tilde{\pi}_{il}(s_0). \tag{2.3}$$

If $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ and $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$, then it follows from (2.3) that

$$\pi_i^* = \pi_i \tilde{\pi}_i, \quad \pi_{il}^* = \pi_{il}\tilde{\pi}_{il}. \tag{2.4}$$

We use (2.4) in Section 4 to study the properties of inverse sampling for estimating a population total. Note that (π_i^*, π_{il}^*) may correspond to some other simpler sampling design if it is not feasible to match simple random sampling (SRS).

2.1 Stratified Simple Random Sampling

Suppose that the original sample s_0 is a stratified simple random sample, i.e.,

$$p_0(s_0) = \prod_1^L \binom{N_h}{n_h}^{-1}, \tag{2.5}$$

where $N_h(n_h)$ denotes the number of population (sample) units in stratum $h\ (=1,\ldots,L)$. We wish to draw a subsample s^* of size m such that $p(s^*) = 1/\binom{N}{m}$, where $N = \sum_h N_h$. Clearly, m cannot be larger than $\min(n_h)$. Let $m = (m_1,\ldots,m_L)^T$ denote the (random) number of units in each stratum that belong to s^*,

$0 \le m_h \le m$, $\sum_h m_h = m$. Noting that the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_h \binom{N_h - m_h}{n_h - m_h}$, it follows from (2.2) that

$$p(s^* \mid s_0) = \frac{\prod_h \binom{N_h}{m_h}}{\binom{N}{m}} \frac{1}{\prod_h \binom{n_h}{m_h}} \tag{2.6}$$

The subsampling scheme readily follows from (2.6): (i) Generate m from the hypergeometric distribution $f(m) = \prod_h \binom{N_h}{m_h} / \binom{N}{m}$; (ii) Draw a simple random sample of size m_h, without replacement, from the n_h sample units in stratum h, independently across strata $h\ (=1, \ldots, L)$. HOS specify $p(s^* \mid s_0)$ first and then verify that it gives $p(s^*) = \binom{N}{n}^{-1}$. Our approach provides the subsampling scheme from the specification of $p_0(s_0)$ and $p(s^*)$.

2.2 One-stage Cluster Sampling

HOS studied the case of one-stage cluster sampling in detail. Three sampling designs for s_0 were investigated: (1) Equal cluster sizes, M, and clusters sampled with equal design probability; (2) Unequal cluster sizes, M_i, and clusters sampled with equal probability; (3) Unequal cluster sizes, M_i, and clusters sampled with probability proportional to size M_i and with replacement.

Case 1. Exact matching with SRS is difficult to implement in the case of equal cluster sizes, M, and clusters sampled with equal probability. Suppose s_0 contains k clusters drawn from K clusters in the population $N = KM$. A simple approximate method of subsampling selects one element at random from each sample cluster so that the size of s^* is k. Hoffman, Sen and Weinberg (2001) used a similar method for biostatistical applications. HOS used systematic sampling to select one case from each sample cluster.

Case 2. Hoffman, Sen and Weinberg (2001) selected one unit at random from each cluster in the case of unequal cluster sizes, under a model-based framework for clustered data. For sampling applications, this method does not work in the sense that it is not possible to obtain SRS of fixed sizes by subsampling, even approximately. HOS proposed an alternative method that artificially enlarges the population to equal cluster size case and then applies subsampling used in Case 1. We first force all clusters to have the same size by adding an appropriate number of pseudo-unit to bring them up to the size of the largest sample cluster. Then we take one unit at random from each sample cluster, and discard any pseudo-units to obtain the final sample. This approximate method makes $p(s^* \mid s_0)$ depend on s_0 because the conditional probability depends on $M(s_0)$, the size of the largest sample cluster.

Case 3. For the case of probability proportional to size (PPS) sampling with replacement of unequal size clusters, HOS proposed a simple method of subsampling which gives $p(s^*) = (1/N)^k$, where s^* now denotes an ordered simple random sample with replacement selected from the $N = \sum_i M_i$, units in the population. Viewing the

sample clusters as ordered, we select one unit at random from each sample cluster. Note that the same cluster might appear more than once in the ordered sample. Denote the size of the cluster drawn in the i-th PPS draw by $M_i^{'}$, then

$$p(s^*) = \left[\prod_i^k \frac{M_i^{'}}{N}\right]\left[\prod_i^k \frac{1}{M_i^{'}}\right] = \left(\frac{1}{N}\right)^k \tag{2.7}$$

where $\prod_i^k (M_i^{'} / N)$ is the probability of drawing the ordered cluster sample. Note that s_0 is the ordered PPS sample and we have only one term in the summation in (2.1).

If the clusters are drawn with inclusion probabilities $\pi_i = kM_i / N$ and without replacement, then it is not possible to match SRS. However, we can treat the clusters as if they were drawn with replacement, as done in practice, and then apply the scheme for Case 3. This will lead to overestimation of variance, but the overestimation is not serious if the sampling fraction k / K is small (see Section 4.3)

2.3 Two-stage Cluster Sampling

HOS also studied two-stage sampling for the following cases: (1) Equal cluster sizes, M and k clusters sampled with equal probability in the first stage; simple random subsample of equal size, m, drawn independently within each sampled cluster (PSU). (2) Unequal cluster sizes, M_i, and k clusters sampled with PPS and with replacement; simple random subsamples of unequal sizes, m_i, drawn independently within each cluster in the with replacement sample.

Case 1. As in the case of one-stage cluster sampling, exact method of inverse sampling is difficult to implement. A simple approximate method of inverse sampling selects one unit at random from each of the k subsamples.

Case 2. As in Case 3 of uni-stage cluster sampling, we simply select one unit at random from each of the ordered subsamples. HOS suggested a different method: Take a simple random sample with replacement of k clusters first and then with each selected cluster take one unit at random from the corresponding subsample. It appears that the first stage inverse sampling of clusters is not necessary. To see this, we note that

$$p_0(s_0) = \prod_{i=1}^k \left[\left(\frac{M_i^{'}}{N}\right)\frac{1}{\binom{M_i^{'}}{m_i^{'}}}\right],$$

where $m_i^{'}$ is the subsample size associated with the cluster selected in the i-th draw $(i = 1, \ldots, k).$. We wish to draw a subsample s^* of size k such that $p(s^*) = (1/N)^k$, where $N = \sum_i^k M_i$. Also the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{i=1}^k \binom{M_i^{'} - 1}{m_i^{'} - 1}$ and

$$\sum_{s_0 \supset s^*} p_0(s_0) = \prod_{i=1}^k \left[\left(\frac{M_i^{'}}{N}\right)\frac{\binom{M_i^{'} - 1}{m_i^{'} - 1}}{\binom{M_i^{'}}{m_i^{'}}}\right] = \prod_{i=1}^k \frac{m_i^{'}}{N}.$$

It follows from (2.2) that $p(s^* \mid s_0) = \prod_i^k \left(1 / m_i^{'}\right)$ and hence the subsampling scheme readily follows.

3. BASIC PROPERTIES

The results in this section are quite general and apply equally to sample surveys and the type of clustered situation considered by Hoffman, Sen and Weinberg (2001). Suppose that we are interested in estimating some population parameter, θ, and we have a sample, s_0, of observations drawn from the population according to some complex design. We assume that we have a subsampling algorithm that can produce samples from some simpler design. This design will often be simple random sampling, but we can extend the range of applications considerably by allowing for the possibility of more general (sub-) designs; for example, stratified SRS when the original sample is a stratified two-stage sample. Our only requirement for the simpler design is that we can produce an estimate of the quantity of interest, θ together with an estimate of its variance. Let $\hat{\theta}_j^*$ and $\hat{V}_j^*$ denote the estimate and variance estimate produced from the j-th subsample when we generate a sequence of g independent subsamples $s_j^* \ (j = 1, \ldots, g)$. Note that the $\hat{\theta}_j^*$'s are not unconditionally independent when averaged over the distribution of the initial sample, s_0. An estimate of θ based on the g subsamples is

$$\hat{\theta}_g = \frac{1}{g} \sum_{j=1}^{g} \hat{\theta}_j^* \tag{3.1}$$

We denote the estimator based on s_0 as $\hat{\theta}$. Theorem 1 gives some results on $\hat{\theta}_g$.

Theorem 1

1. *Conditional on the original sample,* s_0, $\hat{\theta}_g$ *converges almost surely to* $E(\hat{\theta}_1^* \mid s_0) = \hat{\theta}_\infty$, *say, as* $g \to \infty$.
2. $E(\hat{\theta}_g) = E(\hat{\theta}_1^*)$.
3. $Var(\hat{\theta}_g) = Var(\hat{\theta}_\infty) + \frac{1}{g} E[Var(\hat{\theta}_1^* \mid s_0)]$.
4. *If* $r_g = \frac{Var(\hat{\theta}_g)}{Var(\hat{\theta}_\infty)}$, *then* $r_g = 1 + \frac{r_1 - 1}{g}$.

Result 4 of Theorem 1 demonstrates that increasing the number of subsamples, g, does indeed increase the efficiency of $\hat{\theta}_g$. More precisely, the variance ratio r_g has the form $a + b/g$. If the subsample estimator, $\hat{\theta}_1^*$, is unbiased for θ, then so is the resampling estimator, $\hat{\theta}_g$. However, if $\hat{\theta}_1^*$ has bias of order m^{-1}, where m denotes the subsample size, then $\hat{\theta}_g$ has exactly the same bias. Since m will usually be very much smaller than the original sample size, this bias can be appreciable. This is a serious limitation of $\hat{\theta}_g$ in the nonlinear cases, such as ratios and regression coefficients. In Section 5, we propose an alternative estimator of θ based on the estimating equations (EE) approach. This estimator is asymptotically unbiased for any m as the size of s_0 increases, unlike $\hat{\theta}_g$.

The fact that $\hat{\theta}_1^*, \ldots, \hat{\theta}_g^*$ are not unconditionally independent means that estimating $Var(\hat{\theta}_g)$ is not completely straightforward. However, a relatively simple variance estimator may be obtained using Theorem 2 below.

Theorem 2

$$Var(\hat{\theta}_g) = Var(\hat{\theta}_1^*) - \frac{g-1}{g} E[Var(\hat{\theta}_1^* \mid s_0)]. \tag{3.2}$$

We can estimate the first term of (3.2) by $\hat{V}_j^*$ for $j = 1, \ldots, g$, and hence by their average $g^{-1}\sum \hat{V}_j^*$. In addition, the quantity

$$s_{\theta g}^2 = \frac{1}{g-1}\sum_{j=1}^{g}(\hat{\theta}_j^* - \hat{\theta}_g)^2$$

gives an unbiased estimator of $E[Var(\hat{\theta}_1^* \mid s_0)]$ since $\hat{\theta}_1^*, \ldots, \hat{\theta}_g^*$ are conditionally independent given the initial sample, s_0. This leads to an estimator of $Var(\hat{\theta}_g)$ of the form

$$\hat{V}_g = \frac{1}{g}\sum_{j=1}^{g}\hat{V}_j^* - \frac{1}{g}\sum_{j=1}^{g}(\hat{\theta}_j^* - \hat{\theta}_g)^2 . \tag{3.3}$$

The properties of the variance estimator $\hat{V}_g$ depend on the properties of the subsample estimator $\hat{V}_j^*$. For example, if $\hat{V}_j^*$ is unbiased, then $\hat{V}_g$ is also unbiased.

4. ESTIMATION OF A TOTAL

4.1 Exact Matching

As shown in Section 3, resampling increases the efficiency of an estimator, but this does not necessarily mean that the resampling estimator, $\hat{\theta}_g$, converges to the original full sample estimator, $\hat{\theta}$, as $g \to \infty$, even when we start with an unbiased estimator for the subsample. In this section, we study the special case of a total $\theta = Y$ and consider the Horvitz-Thompson (H-T) unbiased estimator, $\hat{Y} = \sum_{i \in s_0} y_i / \pi_i$, based on the original full sample. Theorem 3 below establishes conditions under which the corresponding resampling estimator

$$\hat{Y}_g = \frac{1}{g}\sum_{j=1}^{g}\hat{Y}_j^* \tag{4.1}$$

converges to the H-T estimator, $\hat{Y}$, for the original design as $g \to \infty$.

Theorem 3

Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i-th unit is selected in the subsample for a given initial sample, s_0. Suppose that $\hat{\theta}_j^* = \hat{Y}_j^*$ is the H-T estimator of a total $\theta = Y$ for the j-th subsample. Then the limiting resampling estimator, $\hat{\theta}_\infty^* = \hat{Y}_\infty$, will be the H-T estimator, $\hat{Y}$, for the original design if and only if the conditional inclusion probabilities $\tilde{\pi}_i(s_0)$ are constant for all s_0 containing the i-th unit, i.e., $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ for all $s_0 \supset i$.

The condition $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ is a fairly natural one for most sampling designs for which the H-T estimator is used. If the subsamples are all simple random samples of fixed size m, then the H-T estimator for a subsample is simply the subsample mean, which is the natural estimator.

Theorem 4 below establishes conditions under which the resampling variance estimator, $\hat{V}_{g,HT}$, of $\hat{Y}_g$ converges to $\hat{V}_{HT}$, the H-T variance estimator of $\hat{Y}$ for the original design, as $g \to \infty$, where

$$\hat{V}_{HT} = \sum_{i,\ l \in s_0}\sum \frac{\pi_{il} - \pi_i\pi_l}{\pi_i\pi_l\pi_{il}} y_i y_l \tag{4.2}$$

(see Cochran (1977), *p*.261) and

$$\hat{V}_{g,HT} = \frac{1}{g}\sum_{j=1}^{g}\hat{V}_{j,HT}^{*} - \frac{1}{g}\sum_{j=1}^{g}(\hat{Y}_j^{*} - \hat{Y}_g)^2$$

with

$$\hat{V}_{j,HT}^{*} = \sum_{i,}\sum_{l\in s_j^{*}}\frac{\pi_{il}^{*} - \pi_i^{*}\pi_l^{*}}{\pi_i^{*}\pi_l^{*}\pi_{il}^{*}} y_i y_l \ . \tag{4.3}$$

Note that $\hat{V}_{j,HT}^{*}$ is the H-T variance estimator of $\hat{Y}_j^{*}$, and $\pi_{ii}^{*} = \pi_i^{*}$, $\pi_{ii} = \pi_i$.

Theorem 4

If $\hat{V}_{j,HT}^{*}$ *is the Horvitz-Thompson (H-T) variance estimator of* $\hat{Y}_j^{*}$ *for the j-th subsample, then conditional on* s_0, $\hat{V}_{g,HT}$, *converges to the Horvitz-Thompson (H-T) variance estimator of* $\hat{Y}$ *for the original design, as* $g \to \infty$, *if the conditional joint inclusion probabilities are constant for all* s_0 *containing a given pair* (i,l) *of units, i.e.,* $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$ *for all* $s_0 \supset \{i,l\}$.

4.2 Exact Matching: PPS Estimates

(i) Unistage Cluster Sampling

For the case of PPS sampling with replacement of clusters with unequal sizes M_i, we have exact matching with SRS with replacement. The estimates of Y is given by $\hat{Y}_{pps} = (N/k)\sum_{i=1}^{k}\overline{Y}_i^{'}$, where N is the total number of population elements and $\overline{Y}_i^{'}$ is the mean of the cluster selected on the i-th draw. The estimator $\hat{Y}_{pps}$ is not equal to the H-T estimator of Y. The resampling estimator corresponding to $\hat{Y}_{pps}$ is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^{*}$, where $\hat{Y}_j^{*}$ denotes the estimator of Y from the j-th inverse sample. It is easy to verify that $\hat{Y}_{\infty} = \hat{Y}_{pps}$, noting that $\hat{Y}_j^{*} = (N/k)\sum_{i=1}^{k} y_i^{'}$ where $y_i^{'}$ denotes the value of the element of an inverse sample selected from the cluster in the i-th draw.

The variance estimator of $\hat{Y}_{pps}$ is given by

$$\hat{V}_{pps} = \frac{N^2}{k}\frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{Y}_i^{'} - \frac{1}{k}\sum_{i=1}^{k}\overline{Y}_i^{'}\right)^2 .$$

It is easy to verify that $\hat{V}_{\infty} = \hat{V}_{pps}$. Thus, resampling preserves both the estimator and the variance estimator.

(ii) Two-stage Cluster Sampling

Turning to the case of unequal cluster sizes, M_i, we select the clusters with PPS and with replacement, and then draw simple random subsampling of equal size, m, independently within each cluster in the with-replacement sample. The estimator of Y is $\hat{Y}_{pps} = (N/k)\sum_{i=1}^{k}\bar{y}_i^{'}$ where $\bar{y}_i^{'}$ is the sample mean of the cluster selected in the i-th draw. The variance estimator of $\hat{Y}_{pps}$ is given by

$$\hat{V}_{pps} = \frac{N^2}{k}\frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{y}_i^{'} - \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i^{'}\right)^2 .$$

The resampling estimator is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$, where $\hat{Y}_j^* = (N/k)\sum_{i=1}^{k} y_i^{'}$, and $y_i^{'}$ is defined as above. It is easy to verify that $\hat{Y}_\infty = \hat{Y}_{pps}$ and $\hat{V}_\infty = \hat{V}_{pps}$. Thus, resampling preserves both the estimator and the variance estimator.

4.3 Approximate Matching

In Section 2 we noted that exact matching with SRS is difficult to implement when the original sampling design involves clusters. We proposed several approximate matching methods to overcome this difficulty. In this subsection we study the properties of the approximate matching methods.

4.3.1 Unistage Cluster Sampling

In Section 2.2, we considered the case of equal clusters, M, and proposed to select one element at random from each sample cluster $i\ (=1,\ldots,k)$ selected with equal probabilities and without replacement. The estimator of total Y is given by $\hat{Y} = (K/k)\sum_{i=1}^{k} Y_i$, where Y_i is the i-th sample cluster total and K is the number of population clusters. The corresponding resampling estimator is $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$ with $\hat{Y}_j^* = N\bar{y}_j^*$ denoting the estimator of Y from the j-th inverse sample. It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator $\hat{Y}$ in the limit. However, the variance estimator of $\hat{Y}$, namely

$$\hat{V}_{pps} = \frac{K^2}{k}\left(1-\frac{k}{K}\right)\frac{1}{k-1}\sum_{i=1}^{k}\left(Y_i - \frac{1}{k}\sum_{i=1}^{k}Y_i\right)^2 ,$$

is not preserved. It is easy to verify that

$$\hat{V}/\hat{V}_\infty = 1 - k/K \tag{4.5}$$

It now follows from (4.5) that $\hat{V}_\infty$ leads to overestimation of the variance if the sampling fraction k/K is not small.

4.3.2 Two-stage Cluster Sampling

Consider the case of two-stage cluster sampling with equal cluster sizes, M, and SRS without replacement in both stages. As noted in Section 2.3, exact matching is difficult to implement. The approximate inverse sampling method consists of selecting one element at random from the m sample elements in each sample cluster $i\ (=1,\ldots,k)$. Denote the values of the elements by $y_1^{'},\ldots,y_k^{'}$. The H-T estimator of Y is given by $\hat{Y} = (K/k)\sum_{i=1}^{k}\hat{Y}_i$, where $\hat{Y}_i = M\bar{y}_i$ and $\bar{y}_i$ is the sample mean of the i-th sample cluster. The resampling estimator, based on approximate matching, is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$, where $\hat{Y}_j^* = (N/k)\sum_{i=1}^{k} y_i^{'}$. It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator $\hat{Y}$ in the limit.

The variance estimator of $\hat{Y}$ is given by

$$\hat{V} = N^2\left\{\frac{1}{k}\left(1-\frac{k}{K}\right)s_{1y}^2 + \frac{k}{K}\left(1-\frac{m}{M}\right)\frac{1}{km}s_{2y}^2\right\}, \tag{4.6}$$

where $s_{1y}^2 = \sum_i (\bar{y}_i - \bar{y})^2 / (k-1)$, $s_{2y}^2 = \sum_i s_{2i}^2 / k$ with s_{2i}^2 denoting the sample variance in the i-th cluster, $\bar{y}_i$ is the i-th cluster sample mean and $\bar{y} = \sum_i \bar{y}_i / k$ is the overall sample mean (see Cochran (1977), *p.*276 - 278). The resampling variance estimator $\hat{V}_g$ tends to

$$\hat{V}_\infty = N^2 \frac{1}{k} s_{1y}^2 \tag{4.7}$$

as $g \to \infty$. It follows from (4.6) and (4.7) that

$$\begin{aligned} \frac{\hat{V}}{\hat{V}_\infty} &= 1 - \frac{k}{K}\left[1 - \left(1 - \frac{m}{M}\right)\frac{1}{km}\frac{s_{2y}^2}{s_{1y}^2}\right] \\ &\approx 1 - \frac{k}{K}, \end{aligned} \tag{4.8}$$

because the neglected term in (4.8) is of order $(mK)^{-1}$. It follows that $\hat{V}_\infty$ again leads to overestimation of the variance if the sampling fraction k/K is not small.

5. ESTIMATING EQUATIONS APPROACH

In this section, we study an estimating equations approach to inverse sampling. This approach permits valid inferences on nonlinear parameters such as ratios and "census" linear regression and logistic regression parameters. As noted in Section 3, the resampling estimator $\hat{\theta}_g$, given by (3.1), has exactly the same bias as $\hat{\theta}_1^*$, and the bias of $\hat{\theta}_1^*$ is of order m^{-1}, where m is the subsample size. As a result, the bias of $\hat{\theta}_g$ can be appreciable because m is usually very much smaller than the original sample size n. In fact, m could be as small as 2 for stratified two-stage cluster sampling designs with two sample clusters in each stratum. Moreover, for logistic regression and other cases, the calculation of $\hat{\theta}_j^*$ and $\hat{\theta}$ involves iterative solutions. As a result, the implementation of $\hat{\theta}_g$, and the resampling variance estimator $\hat{V}_g$, given by (3.3), could become computationally very cumbersome when the number of resamples, g, is large. We avoid these difficulties using an estimating equation approach.

A finite population parameter vector θ_N may be regarded as the solution to "census" estimating equations (EE's):

$$S(\theta) = \sum_{k \in U} u_k(\theta) = 0 \tag{5.1}$$

where $\sum_{k \in U}$ denotes the summation over the finite population U of size N, and the estimating functions $u_k(\theta)$ are suitably chosen (Binder (1983), Godambe and Thompson (1986)). For example, consider the scalar case of (5.1) and let $u_k(\theta) = y_k - \theta$ in (5.1). This gives the population mean $\theta_N = \bar{Y}$. Similarly, letting $u_k(\theta) = y_k - \theta x_k$ we get the ratio of means: $\theta_N = R = \bar{Y} / \bar{X}$. The choice $u_k(\theta) = x_k(y_k - \mu_k(\theta))$ with $\mu_k(\theta) = x_k^T \theta$ gives the least squares regression vector

$$\theta_N = \left(\sum_{k \in U} x_k x_k^T\right)^{-1} \sum_{k \in U} x_k y_k \ .$$

The choice $u_k(\theta) = x_k(y_k - \mu_k(\theta))$ with $\log[\mu_k(\theta)/(1-\mu_k(\theta))] = x_k^T \theta$ gives the logistic regression vector θ_N.

The sample estimating equation are given by

$$\hat{S}(\theta) = \sum_{k \in s_0} w_k u_k(\theta) = 0, \tag{5.2}$$

where w_k is the survey weight attached to $k \in s_0$; in particular, $w_k = 1/\pi_k$ if the H-T estimator of $S(\theta)$ is used. The solution to (5.2) gives the estimator $\hat{\theta}$ which, in general, is nonlinear and hence biased. We assume that the size of the original sample, s_0, is large enough to neglect the bias of $\hat{\theta}$. For logistic regression and other complex cases, it is necessary to solve (5.2) iteratively to obtain the solution $\hat{\theta}$. The Newton-Raphson algorithm is commonly used to solve (5.2).

Under regularity conditions, Binder (1983) obtained a Taylor linearization variance estimator of $\hat{\theta}$ as

$$\hat{V}_L(\hat{\theta}) = [\hat{J}(\hat{\theta})]^{-1} \hat{V}[\hat{S}(\hat{\theta})][\hat{J}(\hat{\theta})]^{-1}, \tag{5.3}$$

where $\hat{V}[\hat{S}(\hat{\theta})]$ is the variance estimator of the estimated vector of totals, $\hat{S}(\theta)$, of the $u_k(\theta)$'s evaluated at $\theta = \hat{\theta}$ and $\hat{J}(\hat{\theta})$ is the observed information matrix obtained by evaluating $\hat{J}(\theta) = -\partial \hat{S} / \partial \theta^T = -\sum_{k \in s_0} w_k \partial u_k(\theta) / \partial \theta^T$ at $\theta = \hat{\theta}$. For example, if $u_k(\theta) = y_k - \theta x_k$ then $\hat{\theta} = \sum_{s_0} w_k y_k / \sum_{s_0} w_k x_k = \hat{R}$ is the estimated ratio, and (5.3) reduces to the usual formula

$$\hat{V}_L(\hat{\theta}) = \left(\sum_{k \in s_0} w_k x_k \right)^{-2} \hat{V}\left[\sum_{k \in s_0} w_k u_k(\hat{\theta}) \right],$$

noting that $\hat{J}(\theta) = -\sum_{k \in s_0} w_k x_k$.

As noted in Section 3, the estimator $\hat{\theta}_g$, based on the average of the subsample estimator $\hat{\theta}_j^*$, can be seriously biased if the subsample size m is not large. To avoid this difficulty, we now propose a combined resampling estimator $\hat{\theta}_{gc}$ that leads to valid inference regardless of the subsample size m. A combined resampling equation is given by

$$\hat{S}_{gc}(\theta) = \frac{1}{g} \sum_{j=1}^{g} \hat{S}_j^*(\theta) = 0, \tag{5.4}$$

where $\hat{S}_j^*(\theta) = \frac{N}{m} \sum_{k \in s_j^*} u_k(\theta)$. In general, we solve (5.4) using the Newton-Raphson algorithm. At convergence, we get the estimator $\hat{\theta}_{gc}$ as well as the observed information matrix $\hat{J}_{gc}(\hat{\theta}_{gc})$ given by $\hat{J}_{gc}(\theta) = -\partial \hat{S}_{gc}(\theta) / \partial \theta^T = -\frac{1}{g} \sum_{j=1}^{g} \frac{N}{m} \sum_{k \in s_j^*} \partial u_k(\theta) / \partial \theta^T$ evaluated at $\theta = \hat{\theta}_{gc}$. Note that we solve the EE equations only once to get $\hat{\theta}_{gc}$.

To illustrate the proposed method, consider the special case of ratio $\theta_N = R$, in which case $u_k(\theta) = y_k - \theta x_k$. The solution $\hat{\theta}_{gc} = \hat{R}_{gc}$ is then given by

$$\hat{\theta}_{gc} = \hat{R}_{gc} = \frac{\sum_{j=1}^{g}\sum_{k\in s_j^*} y_k}{\sum_{j=1}^{g}\sum_{k\in s_j^*} x_k},$$

a combined ratio estimator of R.

Assuming first moment matching, it follows from (5.4) that $\hat{\theta}_{gc}$ is a solution of

$$\hat{S}_{\infty c}(\theta) = E[\hat{S}_1^* \mid s_0] = \hat{S}(\theta) = 0. \tag{5.5}$$

As a result, $\hat{\theta}_{\infty c} = \hat{\theta}$ regardless of the subsample size m. Thus, the bias of $\hat{\theta}_{gc}$ is of the same order as the bias of $\hat{\theta}$ for large number of resamples, g, regardless of the subsample size, m.

We now apply Binder's (1983) method to $\hat{S}_{gc}(\theta)$ to get a linearization resampling variance estimator. It follows from (5.3) that

$$\hat{V}_L(\hat{\theta}_{gc}) = [\hat{J}_{gc}(\hat{\theta}_{gc})]^{-1}\hat{V}[\hat{S}_{gc}(\hat{\theta}_{gc})][\hat{J}_{gc}(\hat{\theta}_{gc})]^{-1}, \tag{5.6}$$

where $\hat{V}[\hat{S}_{gc}(\hat{\theta}_{gc})]$ is the variance estimator of the estimated vector of totals, $\hat{S}_{gc}(\theta)$, of the $u_k(\theta)$'s evaluated at $\hat{\theta} = \hat{\theta}_{gc}$. Note that $\hat{J}_{gc}(\hat{\theta}_{gc})$ is obtained at the convergence of the N-R algorithm applied to (5.4). Since $\hat{S}_{gc}(\theta)$ is the resampling estimator of the total $S(\theta)$, it follows that the resampling covariance matrix of $\hat{S}_{gc}(\theta)$ is given by

$$\tilde{V}_{gS} = \frac{1}{g}\sum_{j=1}^{g}\tilde{V}_{iS}^* - \frac{1}{g}\sum_{j=1}^{g}\left[\hat{S}_j^*(\theta) - \hat{S}_{gc}(\theta)\right]\left[\hat{S}_j^*(\theta) - \hat{S}_{gc}(\theta)\right]^T \tag{5.7}$$

where $\tilde{V}_{jS}^*$ is the SRS variance estimator from the j-th resample, assuming second moment matching. If the matching is with respect to SRS without replacement, then

$$\tilde{V}_{jS}^* = \frac{N^2}{m}\left(1-\frac{m}{N}\right)\frac{1}{m-1}\sum_{k\in s_j^*}\left[u_k(\theta) - \frac{1}{m}\sum_{k\in s_j^*}u_k(\theta)\right]\left[u_k(\theta) - \frac{1}{m}\sum_{k\in s_j^*}u_k(\theta)\right]^T. \tag{5.8}$$

In the case of matching to SRS with replacement, we replace $(1-m/N)$ by 1 in (5.8).

Now substituting $\hat{\theta}_{gc}$ for θ in (5.7) we get the resampling covariance matrix $\hat{V}_{gS}$ given by

$$\hat{V}_{gS} = \frac{1}{g}\sum_{j=1}^{g}\hat{V}_{jS}^* - \frac{1}{g}\sum_{j=1}^{g}\hat{S}_j^*(\hat{\theta}_{gc})\hat{S}_j^*(\hat{\theta}_{gc})^T \tag{5.9}$$

where $\hat{V}_{jS}^*$ is obtained from (5.8) by substituting $\hat{\theta}_{gc}$ for θ. Note that $\hat{S}_{gc}(\hat{\theta}_{gc}) = 0$. The linearization resampling variance estimator, for a given g, is now given by

$$\hat{V}_L(\hat{\theta}_{gc}) = [\hat{J}_{gc}(\hat{\theta}_{gc})]^{-1}\hat{V}_{gS}[\hat{J}_{gc}(\hat{\theta}_{gc})]^{-1}. \tag{5.10}$$

Under second moment matching with SRS, it is easy to verify that the variance estimator (5.10) as $g\to\infty$ equals the Binder's variance estimator $\hat{V}_L(\hat{\theta})$ given by (5.3). This follows by noting that $\hat{\theta}_{\infty c} = \hat{\theta}$, $\hat{J}_{\infty c}(\hat{\theta}) = \hat{J}(\theta)$ and $\tilde{V}_{\infty S} = \hat{V}[\hat{S}(\theta)]$ under second moment matching with SRS. Thus, the variance estimator $\hat{V}_L(\hat{\theta}_{gc})$ provides valid inferences on θ for large number of resamples, g, regardless of the subsample size, m.

To illustrate the calculation of the resampling variance estimator $\hat{V}_L(\hat{\theta}_{gc})$, given by (5.6), consider the special case of a ratio $\theta_N = R$ with $u_k(\theta) = y_k - \theta x_k$. We have

$$\tilde{V}_{jS}^{*} = \frac{N^2}{m}\left(1-\frac{m}{N}\right)\frac{1}{m-1}\sum_{k\in s_j^*}(e_k - \bar{e}_j^*)^2 ,$$

where $e_k = y_k - \theta x_k$, $\bar{e}_j^* = \bar{y}_j^* - \theta \bar{x}_j^*$ and $(\bar{y}_j^*, \bar{x}_j^*)$ are the j-th subsample means. Further, $\hat{J}_{gc}(\hat{\theta}) = -\frac{N}{g}\sum_{j=1}^{g}\bar{x}_j^*$ and $\hat{S}_j^*(\theta) = N(\bar{y}_j^* - \theta\bar{x}_j^*)$.

It is important to note again that the estimator $\hat{\theta}_{gc}$ and the associated variance estimator $\hat{V}_L(\hat{\theta}_{gc})$ can be implemented from a micro data file providing g subsamples, each of size m. Neither the survey weights w_k nor the cluster identifiers are needed so that confidentiality of micro data may be preserved.

REFERENCES

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51, pp. 279-292.

Binder, D. A. (1992), "Fitting Cox's Proportional Hazard Models from Survey Data", *Biometrika*, 79, pp. 139-147.

Cochran, W. (1977), *Sampling Techniques*, New York: Wiley.

Hinkins, S., Oh, H. L. and Scheuren, F. (1997), "Inverse Sampling Design Algorithms", *Survey Methodology*, 23, pp. 11-21.

Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001), "Within-Cluster Resampling", *Biometrika*, 88, pp. 1121-34.

Rao, J. N. K., Scott, A. J. and Benhin, E. (2002), "Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling", submitted to *Survey Methodology*.

Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.)(1989), *Analysis of Complex Surveys*,. Chichester: Wiley.