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 $\{(y_i, z_i), i \epsilon s\}$ 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 $\sum_Y$ 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 Y_i are independent draws from a common distribution $f(Y;\theta)$ and let $\ell_N(\theta; Y) = \sum_{i=1}^N \log f(Y_i;\theta)$ define the census log likelihood function. Under some regularity conditions, the MLE, $\hat{\theta}$, solves the equations

$$U(\theta) = d\ell_N(\theta; Y)/d\theta = \sum_{i=1}^N u(\theta;y_i) = 0, \qquad (2.6)$$

where "d" defines the derivative operator and $u(\theta,y_i) = d\log f(Y_i;\theta)/d\theta$. The pseudo MLE of θ is defined as the solution of $\hat{U}(\theta) = 0$ where $\hat{U}(\theta)$ is a design consistent estimator of $U(\theta)$ in the sense that $\text{plim}_{n\to\infty,N\to\infty}[\hat{U}(\theta) - U(\theta)] = 0$ for all $\theta\epsilon\Theta$. The commonly used estimator of $U(\theta)$ is the Horvitz-Thompson (1952) estimator so that the pseudo MLE of θ is the solution of $\hat{U}(\theta) = \sum_{i=1}^n w_i^* u(\theta; y_i) = 0$ 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 μ_Y and $\sum_Y$ are

$$\tilde{\mu}_Y = \sum_{i=1}^n w_i^* y_i \Big/ \sum_{i=1}^n w_i^*; \quad \tilde{\Sigma}_Y = \sum_{i=1}^n w_i^* (y_i - \tilde{\mu}_Y)(y_i - \tilde{\mu}_Y)' \Big/ \sum_{i=1}^n w_i^*. \qquad (2.7)$$

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 $U(\theta)$. 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 $\sum_{i=1}^n w_i^* (y_i,z_i) / \sum_{i=1}^n w_i^*$, 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 $i \epsilon s$. The pdf of X_i^w is then

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

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 $k = 1, \ldots, n$, $P(j \epsilon s) = w(\underset{\sim}{x}_j;\underset{\sim}{\alpha})$, $j = 1, \ldots, N$. 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

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

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 Σ_{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 $\underset{\sim}{x}_i' = (\underset{\sim}{y}_i', \underset{\sim}{z}_i')$, $i = 1, \ldots, n$ and the sample selection probabilities $P_i = t_i / N\bar{T}$. Under the assumption $\bar{T} = \mu_T$, the likelihood for $[\underset{\sim}{\mu}_X, \Sigma_{XX}]$ can be written using (3.3) as

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

where $\phi(\underset{\sim}{x}; \underset{\sim}{\mu}_X, \Sigma_{XX})$ is the normal pdf with mean $\underset{\sim}{\mu}_X$ and $V - C$ matrix Σ_{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

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

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 $\underset{\sim}{\alpha}^* = (\alpha_0^*, \underset{\sim}{\alpha}_1^*, \alpha_2^*)'$ 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 $i \epsilon s$ is in this case [compare with (3.2)]

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

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

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

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 $\mu_T = E(T) = \underset{\sim}{\alpha}'\underset{\sim}{\mu}_X$, $\sigma_T^2 = \text{Var}(T) = \underset{\sim}{\alpha}' \Sigma_{XX} \underset{\sim}{\alpha}$ 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

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

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

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, $t^{(1)} = ½(t_{(n_1)} + t_{(n_1+1)}) \ldots t^{(L-1)} = ½(t_{(n^*)} + t_{(n^*+1)})$ 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 $\hat{\sigma}_T^2 = \hat{\underset{\sim}{\alpha}}' \sum_{XX} \hat{\underset{\sim}{\alpha}}$. 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 Σ^{-1} where Σ is the covariance matrix among Y_1, Y_2 and Z_1. Furthermore, to insure that the six parameters that define Σ^{-1} are unconstrained, we use the elements of the upper triangular matrix B so that $B'B = \Sigma^{-1}$. Any choice of the values for B leads to a matrix Σ^{-1} that is positive semi-definite.

4.4 Results

We present the results obtained when estimating $\mu_1 = E(Y_1)$, $\sigma_1^2 = \mathrm{Var}(Y_1)$ 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	
	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$
$ML(Z_1,Z_2)$	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
$WDML(X^*,\underset{\sim}{t}_s)$	–	–	0.74	0.43
$WDML(X^*,\underset{\sim}{t}_s,Z_1)$	–	–	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	
	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$
$ML(Z_1,Z_2)$	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
$WDML(X^*,\underset{\sim}{t}_s)$	–	–	26.18*	15.46
$WDML(X^*,\underset{\sim}{t}_s,Z_1)$	–	–	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	
	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$	$t_i = 0.5\underset{\sim}{z}_i$	$t_i = 0.25\underset{\sim}{x}_i$
$ML(Z_1,Z_2)$	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
$WDML(X^*,\underset{\sim}{t}_s)$	–	–	0.048	0.045
$WDML(X^*,\underset{\sim}{t}_s,Z_1)$	–	–	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 $t_i = 0.25\underset{\sim}{x}_i$, 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 $t_i = 0.5\underset{\sim}{z}_i$, 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.