Sankhyā: The Indian Journal of Statistics
Special Issue on Sample Surveys
1999, Volume 61, Series B, Pt. 1, pp. 166-186

PARAMETRIC AND SEMI-PARAMETRIC ESTIMATION OF REGRESSION MODELS FITTED TO SURVEY DATA*

By DANNY PFEFFERMANN
and
MICHAIL SVERCHKOV
Hebrew University, Jerusalem

SUMMARY. This paper proposes two new classes of estimators for regression models fitted to survey data. The proposed estimators account for the effect of nonignorable sampling schemes which are known to bias standard estimators. Both classes derive from relationships between the population distribution and the sample distribution of the sample measurements. The first class consists of parametric estimators. These are obtained by extracting the sample distribution as a function of the population distribution and the sample selection probabilities and applying maximum likelihood theory to this distribution. The second class consists of semi-parametric estimators, obtained by utilizing existing relationships between moments of the two distributions. New tests for sampling ignorability based on these relationships are developed. The proposed estimators and other estimators in common use are applied to real data and further compared in a simulation study. The simulations enable also to study the performance of the sampling ignorability tests and bootstrap variance estimators.

1. Introduction

Regression models are routinely fitted to data collected from survey samples. The sampling designs underlying the sample selection often involve unequal selection probabilities, at least at some stages of the selection process. When these probabilities are related to the values of the regression dependent variable, even after controlling for the regressor variables, the use of ordinary least squares (OLS) estimators or other estimators that ignore the sample selection process can yield large biases and hence mislead the inference.

AMS (1991) *subject classification.* 62D05, 62F10

Key words and phrases. Bootstrap, nonignorable sampling, probability weighted estimators, randomization distribution, sample distribution.

* Work supported by the Isreal Science Foundation, grant 841/95-1.

In this article we develop new classes of regression estimators based on the parametric distribution of the sample observations, $f_s(y_i|\mathbf{x}_i) = f(y_i|\mathbf{x}_i, i \in s)$, or moments of this distribution. The use of the sample distribution permits the application of classical inference procedures like maximum likelihood estimation (MLE) or residual analysis for model diagnostics. The functional form of the sample distribution and its relationship to the population distribution $f_p(y_i|\mathbf{x}_i)$ (before sampling), are discussed in Section 2. For earlier references see Krieger and Pfeffermann (1992, 1997) and Pfeffermann, Krieger and Rinott (1998, hereafter PKR). In the rest of this section we review briefly other approaches for regression estimation that account for sample selection effects (informative sampling). To simplify the discussion we consider the case of linear regression and single stage sampling.

The OLS estimators have in this case the familiar form

$$\mathbf{b}_{ols} = (X_s^T X_s)^{-1} X_s^T \mathbf{y}_s = (\sum_{i \in S} \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i \in S} \mathbf{x}_i y_i. \qquad \ldots(1.1)$$

A common procedure to account for possible sampling effects is 'probability weighted least squares' (PWLS) which in the linear case yields the estimators

$$\mathbf{b}_w = (X_s^T W_s X_s)^{-1} X_s^T W_s \mathbf{y}_s = (\sum_{i \in S} \frac{1}{\pi_i} \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i \in S} \frac{1}{\pi_i} \mathbf{x}_i y_i \qquad \ldots(1.2)$$

where $W_s = diag[\frac{1}{\pi_i}, i \in s]$ and the π_i represent the sample inclusion probabilities. The same estimators are obtained by use of the 'pseudo likelihood' approach assuming normality of the population error terms. See Skinner, Holt and Smith, (1989) for description and applications of this approach.

The prominent advantage of these approaches is that the PWLS estimators are approximately design unbiased and consistent for the corresponding OLS 'census coefficients' $B = (\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i$. Since the latter are free of sampling effects and converge to the hypothetical coefficients under an appropriate linear regression model, the PWLS estimators are likewise consistent for the model coefficients with respect to the compound distribution induced by the model and the sampling process. The PWLS estimators are known also to be robust to certain model misspecifications, although in a limited sense.

The use of PWLS, however, has three serious limitations. First, the small sample distribution of these estimators is generally unknown, thus hindering the use of test statistics or confidence intervals for the unknown regression coefficients in such cases. Second, the distribution of the sample residuals computed by application of the PWLS procedure is different under informative sampling from the distribution of the census residuals which restricts their use for model diagnostics on the population model. For example, assuming that the census residuals are normal with mean zero and constant variance, the sample residuals may no longer possess any of these properties if the sample selection probabilities

depend on these residuals. Hence, the use of classical diagnostic statistics aimed at assessing these properties may give false indications. The third limitation of the use of the PWLS approach is that it does not lend itself to prediction problems. For example, it is not clear how one would predict $(Y_i|\mathbf{x}_i, i \notin s)$ if the sample selection is informative. See, e.g., Pfeffermann (1993) for discussion of the use of PWLS, with references.

A different approach proposed in the past for controlling the sample selection effects is to extract the joint distribution $f(\mathbf{y}_s, X_s, Z)$ of the sample measurements $(\mathbf{y}_s, X_s)$ and the population values Z of design variables used for the selection process, by integrating over the unobserved y-values. The regression coefficients of $E(Y|\mathbf{x})$ can be estimated under this approach as functions of the estimators of the parameters indexing the distribution $f(\mathbf{y}_s, X_s, Z)$, (See, e.g., Smith (1981)). A necessary condition for the application of this approach is that the sample selection is independent of the y and $\mathbf{x}$-values given Z, that is, $Pr(S|\mathbf{y}_s, X_s, Z) = Pr(S|Z)$ for all S . It requires also that the population values of Z are known, but this problem can be handled by modelling instead the joint distribution $f(\mathbf{y}_s, X_s, \pi)$ where $\pi^T = (\pi_1, ..., \pi_N)$, provided that π is an 'adequate summary' of Z in the sense that $Pr(S|Z, \pi) = Pr(S|\pi)$. See, e.g., Sugden and Smith (1984) and Chambers, Dorfman and Wang (1998).

The use of this approach permits in principle the computation of MLE under the model and conditions stated above, but it is restricted since modelling the joint distribution $f(\mathbf{y}_s, X_s, Z)$ or even $f(\mathbf{y}_s, X_s, \pi)$ is generally very complicated given the nature of the design variables or the sample selection probabilities. Indeed, all the known applications assume exact or approximate multivariate normality.

A third approach, proposed by Skinner (1994), attempts to deal with these difficulties by extracting the model holding in the population from models identified and fitted to the sample data. Denote, in general, by $f_p(u_i|\mathbf{v}_i)$ the conditional density of a measurement u_i corresponding to population unit i, given a vector variable $\mathbf{v}_i$, and define the sample pdf of $u_i|\mathbf{v}_i$ as, $f_s(u_i|\mathbf{v}_i) = f_p(u_i|\mathbf{v}_i, i \in s)$. Let $w_i = \frac{1}{\pi_i}$ denote the i-th sampling weight, viewed as a random realization of a random variable w. The following two relationships are shown to hold:
(a) $\mathrm{f_p(y_i|\mathbf{x}_i, w_i) = f_s(y_i|\mathbf{x}_i, w_i)}$; (b) $\mathrm{f_p(w_i|\mathbf{x}_i) = w_i f_s(w_i|\mathbf{x}_i)/E_s(w_i|\mathbf{x}_i)}$
where E_s defines expectation with respect to the sample distribution. Skinner's approach consists of the following four steps:

(1) Identify $f_p(y|\mathbf{x}, w) = f_s(y|\mathbf{x}, w)$ from data $\{(y_i, \mathbf{x}_i, w_i), i \in s\}$
(2) Identify $f_s(w|\mathbf{x})$ from data $\{(w_i, \mathbf{x}_i), i \in s\}$
(3) Extract $f_p(w|\mathbf{x})$ from $f_s(w|\mathbf{x})$ using (b) above
(4) Combine $f_p(y|\mathbf{x}, w)$ in (1), with $f_p(w|\mathbf{x})$ in (3), to obtain $f_p(y|\mathbf{x})$.

The appealing feature of this approach is that it permits in principle the use of standard model diagnostics and estimation procedures applied to models fitted to the sample data. On the other hand, for cases where the form of the population model is known, it is not always clear how to define sample models

that are consistent with this model. The distribution of the regression estimators obtained under this approach and in particular, the variances of these estimators have yet to be established. See Section 5 for the performance of bootstrap variance estimators.

In this paper we likewise employ the relationships holding between the population distribution and the sample distribution but in different directions. Instead of extracting the population distribution as a function of the sample distribution, we extract the sample distribution as a function of the population distribution and the first order sample inclusion probabilities. This permits application of likelihood based theory and other classical regression techniques, including residual analysis methodology, directly to the sample observations. We consider also a semi-parametric approach which only uses the form of the expectation $E_p(Y|\mathbf{x})$ in the population in order to extract the expectation in the sample, and hence estimate the regression coefficients.

In Section 2 we extend on the notions of population and sample distributions and establish the relationships between the two distributions and their moments. Test statistics for comparing the two sets of moments and hence detecting possible sampling effects are proposed. Section 3 defines the new parametric and semi-parametric regression estimators implied by these relationships. Section 4 contains the details of a Monte Carlo simulation study designed to illustrate the performance of the new estimators and test statistics, and compare them with some of the other estimators proposed in the literature and described in the introduction. The main findings of this study are discussed in Section 5 which contains also the results obtained for a real data set. We conclude with a brief summary in Section 6.

2. Relationships Between Population and Sample Distributions

2.1 Definitions and basic relationships. Suppose that the population values $[\mathbf{y}, X] = \{(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)\}$ are independent realizations with continuous conditional probability density functions (pdf) $f_p(y_i|\mathbf{x}_i)$. We consider single stage sampling, (see comment 1 below) with inclusion probabilities $\pi_i = Pr(i \in s) = g(\mathbf{y}, X, Z)$ for some function g, where Z denotes as before the population values of design variables used for the sampling process. Let $I_i = 1$ if $i \in s$ and $I_i = 0$, otherwise. The conditional *sample pdf* is defined as,

$$f_s(y_i|\mathbf{x}_i) = f(y_i|\mathbf{x}_i, I_i = 1) = \frac{Pr(I_i = 1|y_i, \mathbf{x}_i) f_p(y_i|\mathbf{x}_i)}{Pr(I_i = 1|\mathbf{x}_i)} \qquad \ldots(2.1)$$

with the second equality obtained by application of Bayes theorem. Note that $Pr(I_i = 1|y_i, \mathbf{x}_i)$ is not necessarily the same as the sample inclusion probability π_i. It follows from (2.1) that the population and sample pdf's are different,

unless $Pr(I_i = 1|y_i, \mathbf{x}_i) = Pr(I_i = 1|\mathbf{x}_i)$ for all y_i, in which case the sampling process can be ignored for inference when conditioning on the $\mathbf{x}$'s.

In what follows we regard the probabilities π_i as realizations of random variables. Let $w_i = \frac{1}{\pi_i}$ define the sampling weight of unit i. In the Appendix, the following additional relationships between the sample and population pdf's are shown to hold for pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$, where E_p and E_s denote expectations under the population and sample pdf's respectively.

$$f_s(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p(\pi_i|\mathbf{v}_i)} \qquad \ldots(2.2)$$

$$f_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i|\mathbf{u}_i, \mathbf{v}_i) f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)} \qquad \ldots(2.3)$$

$$E_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)}. \qquad \ldots(2.4)$$

As special cases of (2.4) we also have,

$$\begin{array}{ll} (a) & E_s(w_i|\mathbf{v}_i) = \frac{1}{E_p(\pi_i|\mathbf{v}_i)}; \\ (b) & E_p(\mathbf{u}_i) = \frac{E_s(w_i\mathbf{u}_i)}{E_s(w_i)}; \\ (c) & E_s(w_i) = \frac{1}{E_p(\pi_i)} \end{array} \qquad \ldots(2.5)$$

The relationship (2.2) is employed in Krieger and Pfeffermann (1997) and in PKR. Equation (2.4) can be shown to follow from Proposition 2 in Skinner(1994). Chambers establishes a similar relationship in an unpublished manuscript.

The relationships (2.2) and (2.3) form the basis for parametric inference on the population distribution by reference to the sample distribution. In this article we restrict to the relationship (2.2) and it is seen that for given (hypothesized) forms of the population pdf and the expectations $E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i)$, the form of the sample pdf is uniquely defined. Furthermore, for independent population measurements PKR establish asymptotic independence of the sample values under commonly used sampling schemes for selection with unequal probabilities. The asymptotics requires that the population size increases (but with the sample size held fixed), and that the (random) size variable used to determine the sample selection probabilities has bounded moments of all order, in addition to some other mild technical conditions. Thus, the relationship (2.2) permits the use of standard inference procedures like MLE or residual analysis applied to the sample measurements. In practice, the forms of the expectations $E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i)$ and $E_p(\pi_i|\mathbf{v}_i)$ are seldom known but they can be identified (approximated) and estimated from the sample data via (2.5a). See Sections 3 and 4.

Equation (2.4) forms the basis for semi-parametric estimation of the population regression $E_p(y_i|\mathbf{x}_i)$, by use of the sample regressions of $y_i w_i$ and w_i. In

Section 3 we develop a new class of estimators arising from this equation. We provide also a model based least squares justification for the PWLS estimators discussed in the introduction.

Comments. (1) We focus throughout this paper on single stage sampling but much of the analysis follows through to multistage sampling. Consider, for example, the case of a two stage sample with the first stage clusters (PSU's) selected with probabilities π_c determined by a size variable Z, (e.g., the cluster size), and the ultimate sampling units selected with probabilities π_{cj}. In regression modelling of data arising from such designs, it is common to allow for independent random cluster effects which account for the homogeneity within the clusters. These models pertain to the population, before sampling. Extension of the parametric and semi-parametric approaches considered in the present paper to two stage sampling requires therefore the extraction of the sample distribution (moments) of the response variable given the cluster effects and extraction of the sample distribution (moments) of the cluster effects. Notice that by the independence results of PKR mentioned before, the sample cluster effects are again independent (asymptotically) although they may have a different distribution. Work in this direction is currently under way. See Section 5.2 of PKR for further discussion with an example.

(2) It was mentioned in the introduction that classical sampling theory does not provide a prediction paradigm under informative sampling. This problem can be resolved under the present approach by extracting the distribution $f_{\bar{s}}(y_i|\mathbf{x}_i) = f(y_i|\mathbf{x}_i, i \notin s)$, similarly to the derivation of $f_s(y_i|\mathbf{x}_i)$ in the Appendix. This issue is not pursued further in this paper.

2.2 Testing for sampling ignorability. The moments relationship in (2.4) suggests a formal test for testing that the sample distribution of the regression residuals is indeed different from the corresponding population distribution. Clearly, when this is not the case, one can ignore the sampling scheme and apply standard regression techniques as in the case of simple random sampling. Let $\epsilon_i = y_i - E_p(y_i|\mathbf{x}_i)$ denote the residual term associated with unit i. Classical test procedures for comparing two distributions are not applicable to this problem since no observations are available for the population distribution of the residuals. Considering, however, that under general conditions the set of all moments of a distribution, when they exist, determine the distribution, it is plausible to test instead hypotheses of the form, $E_p(\epsilon_i^k) = E_s(\epsilon_i^k)$, $k = 1, 2, \ldots$. By (2.5b), $E_p(\epsilon_i^k) = \frac{E_s(\epsilon_i^k w_i)}{E_s(w_i)}$ so that an equivalent set of hypotheses is,

$$H_{0k}: \quad Corr_s(\epsilon_i^k, w_i) = 0, \quad k = 1, 2, \ldots \qquad \ldots(2.6)$$

where $Corr_s$ is the correlation under the sample distribution. In practice, it would normally suffice to test the first 2-3 correlations.

Testing hypotheses on correlation coefficients is a familiar problem. In the simulation study we use as test statistic a standardized form of Fisher transfor-

mation $FT(k) = (1/2)log[(1 + r_k)/(1 - r_k)]$, namely,

$$FTS(k) = FT(k) \Big/ \widehat{SD}(FT(k)) \qquad \ldots(2.7)$$

where r_k is the empirical correlation $\widehat{Corr}(\hat{\epsilon}_i^k, w_i)$, $\hat{\epsilon}_i^k = (y_i - \mathbf{x}_i^T \hat{\beta})^k$ and $\widehat{SD}(FT(k))$ is the bootstrap standard deviation of $FT(k)$, (see Section 4.4). For bivariate normal measurements (u_i, v_i) with $Corr(u, v) = 0$, Fisher's transformation has an asymptotic normal distribution with mean zero. The use of the bootstrap standard deviation was found to yield better results than the common approximation $Var[FT(k)] \approx 1/(n - 3)$ (Kendall and Stuart (1969, Vol 1, p.390)). An alternative test procedure is to regress w_i against $\hat{\epsilon}_i^k$ and test that the corresponding slope coefficient is zero, using the conventional t-statistic. The use of this procedure yields similar results to the use of $FTS(k)$ in our study. See Section 5 for the performance of the statistics $FTS(k)$.

Pfeffermann (1993) reviews and discusses several other tests proposed in the literature for testing sampling ignorability. These tests compare the expectations of unweighted and probability weighted statistics and hence provide only partial answers to the ignorability issue.

2.3 Relationship between the sample distribution and the randomization distribution. We conclude this section by comparing the sample distribution as defined by (2.1) and (2.2) with the randomization (design) distribution underlying classical survey sampling inference. The two distributions are conceptually different as the former accounts for both the distribution generating the population values and the process of sample selection, whereas the latter only accounts for the sample selection with the population values held fixed. Modelling the sample distribution requires therefore the modelling of the population distribution but it permits extraction of the marginal and joint distribution of the sample measurements. This is generally not feasible under the randomization distribution since it requires knowledge of all the population values. As a result, the use of the sample distribution lends itself to the application of standard inference methods. Note again that for the case of independent population measurements, the sample measurements are asymptotically independent under the sample distribution for commonly used sampling schemes. This validates also the use of the Bootstrap method for variance estimation, see Section 4.4 and the empirical results in Section 5.

Another important advantage of the use of the sample distribution is that it can be applied in a conditional set up which, at the present state of art, is restricted under the randomization distribution. In fact, our definition of the sample distribution already employs a conditional formulation. See Rao (1984) for the use of conditional randomization based inference. As already mentioned, the proposed approach can be extended to prediction problems.

Although the use of the sample distribution for inference is appealing in light of the advantages listed above, it should be emphasized that modelling the

population distribution and the conditional expectations of the sample inclusion probabilities, needed for extracting the sample distribution (equation 2.2), is not straightforward. See Section 5 for an illustration with real data. Note on the other hand that hypothesized models can be tested by standard goodness of fit diagnostic procedures. Krieger and Pfeffermann (1997) apply the Kolmogorov-Smirnov and Chi-square test statistics to the sample measurements.

3. New Estimators of Regression Coefficients

3.1 Parametric estimation. Hereafter we consider for convenience linear regression with normal error terms but the theory extends to general regression models of the form $y_i = k(\mathbf{x}_i) + \epsilon_i$, for some function $k(\mathbf{x})$ and independent disturbances $\{\epsilon_i\}$. See also PKR.

Suppose that the population measurements $(y_i, \mathbf{x}_i)$ are independent random realizations such that

$$f_p(y_i|\mathbf{x}_i) = \frac{1}{\sigma}\phi[(y_i - \mathbf{x}_i^T\beta)/\sigma] \qquad \ldots(3.1)$$

where ϕ is the standard normal pdf.

Consider for example the following alternative expressions (approximations) for the expectations $\pi_p(y_i, \mathbf{x}_i) = E_p(\pi_i|y_i, \mathbf{x}_i)$,

$$\pi_p(y_i, \mathbf{x}_i) = exp[A_0 + A_1 y_i + h(\mathbf{x}_i)], \qquad \ldots(3.2a)$$

$$\pi_p(y_i, \mathbf{x}_i) = \sum_{j=0}^{J} A_j y_i^j + h(\mathbf{x}_i); \qquad \ldots(3.2b)$$

where $h(\mathbf{x}_i)$ is some fixed function of $\mathbf{x}_i$.

Substituting (3.1) and (3.2a) in (2.2) yields the sample pdf as

$$f_s(y_i|\mathbf{x}_i) = \frac{1}{\sigma}\phi[(y_i - \mathbf{x}_i^T\beta - A_1\sigma^2)/\sigma]. \qquad \ldots(3.3).$$

Hence, the regression model in the sample is the same as in the population, (including the error variance), except for the intercept term that changes by the constant $A_1\sigma^2$. See PKR for derivation of (3.3) and extensions of this result to other distributions belonging to the exponential family.

When (3.2a) is replaced by (3.2b), the sample pdf has the form

$$f_s(y_i|\mathbf{x}_i) = \frac{\sum_{j=1}^{J}(A_j E^{(j)}) f_p^{(j)}(y_i|\mathbf{x}_i) + [A_0 + h(\mathbf{x}_i)] f_p(y_i|\mathbf{x}_i)}{\sum_{j=1}^{J}(A_j E^{(j)}) + [A_0 + h(\mathbf{x}_i)]} \qquad \ldots(3.4)$$

where $E^{(j)} = E_p(y_i^j|\mathbf{x}_i)$, $f_p^{(j)}(y_i|\mathbf{x}_i) = y_i^j f_p(y_i|\mathbf{x}_i)/E^{(j)}$ and, for our application, $f_p(y_i|\mathbf{x}_i)$ is the normal pdf defined by (3.1). Thus, the sample pdf is in

this case a mixture of the densities $f_p^{(j)}(y_i|\mathbf{x}_i)$, $j = 0, ..., J$. Note that changing the function $h(\mathbf{x})$ to another function $h^*(\mathbf{x})$ only affects the mixture coefficients. The regression coefficients β can be estimated by MLE, with the likelihood computed as the product of the pdf's in (3.3) or (3.4). As already mentioned, the sample measurements are asymptotically independent.

In practice, the expectations $\pi_p(y_i, \mathbf{x}_i)$ are usually unknown but they can be identified and estimated from the sample data by regressing $w_i = \frac{1}{\pi_i}$ against $(y_i, \mathbf{x}_i)$ and substituting $\pi_p(y_i, \mathbf{x}_i) = \frac{1}{E_s(w_i|y_i, \mathbf{x}_i)}$, exploiting the relationship (2.5a). See Section 4.3 for the implementation of this step in the simulation study. By substituting the estimated expectations $\hat{\pi}_p(y_i, \mathbf{x}_i)$ in (3.3) or (3.4), the maximization of the likelihood is only with respect to β and σ^2. This two step estimation process facilitates the maximization process. It is imperative under distributions such as (3.3) because of parameter identification problems. See PKR for discussion.

3.2 Semi-parametric estimation. Next we drop the normality assumption and only assume

$$y_i = \mathbf{x}_i^T\beta + \epsilon_i, \quad E_p(\epsilon_i|\mathbf{x}_i) = 0, \quad Var_p(\epsilon_i|\mathbf{x}_i) = \sigma^2. \qquad \ldots(3.5)$$

By a well known property, $\min_{\tilde{\beta}} E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 = E_p(y_i - \mathbf{x}_i^T\beta)^2$ where β is defined by (3.5). The following relationship is proved in the Appendix,

$$\beta = arg\min_{\tilde{\beta}} E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 = arg\min_{\tilde{\beta}} E_s\Big(\frac{w_i(y_i - \mathbf{x}_i^T\tilde{\beta})^2}{w_s(\mathbf{x}_i)}\Big) \qquad \ldots(3.6)$$

where $w_s(\mathbf{x}_i) = E_s(w_i|\mathbf{x}_i)$.

Thus, the proposed semi-parametric estimator of the vector coefficient β is obtained by the following two step procedure:

Step 1: Estimate $\hat{w}_s(\mathbf{x}_i)$ by regressing w_i against $\mathbf{x}_i$ using the sample measurements,

Step 2: Compute $\mathbf{b}_{sp} = arg\min_{\tilde{\beta}} \Big\{\frac{1}{n}\sum_{i\in S}\big[w_i(y_i - \mathbf{x}_i^T\tilde{\beta})^2/\hat{w}_s(\mathbf{x}_i)\big]\Big\}$.

The estimator $\mathbf{b}_{sp}$ has the form

$$\mathbf{b}_{sp} = (X^TQX)^{-1}X^TQ\mathbf{y} = \Big(\sum_{i\in S} q_i\mathbf{x}_i\mathbf{x}_i^T\Big)^{-1}\Big(\sum_{i\in S} q_i\mathbf{x}_i y_i\Big) \qquad \ldots(3.7)$$

where $Q = diag[q_1, ..., q_n]$ and $q_i = [w_i/\hat{w}_s(\mathbf{x}_i)]$. Note that the estimator $\mathbf{b}_{sp}$ uses basically the same data and assumptions underlying the use of the PWLS estimator $\mathbf{b}_w$ defined by (1.2). The two estimators coincide when the π_i's are not dependent on the $\mathbf{x}_i$'s such that $E_s(w_i|\mathbf{x}_i) = [1/E_p(\pi_i|\mathbf{x}_i)] = const.$. Chris Skinner informed us that by dividing w_i by $w_s(\mathbf{x}_i)$, the resulting weights (q_i) account for the net sampling effect on the conditional distribution of $y_i|\mathbf{x}_i$, whereas

the standard sampling weights (w_i) account for the sampling effect on the joint distribution of $(y_i, \mathbf{x}_i)$. As such, the use of the q-weights is expected to be more efficient for estimating the regression coefficients than the use of the sampling weights. Notice in this respect that when w_i is a deterministic function of $\mathbf{x}_i$, the sampling process is ignorable and the use of the estimator $\mathbf{b_W}$ results in increased variance, whereas in this case $q_i = 1$ and $\mathbf{b}_{sp} = \mathbf{b}_{ols}$.

The estimator $\mathbf{b}_{sp}$ is not strictly unbiased for β under the sample distribution, but it is consistent, provided that the mild conditions guaranteeing the convergence of probability weighted statistics to the corresponding population functionals are satisfied. This can be shown by writing

$$(\mathbf{b}_{sp} - \beta) = \Big(\frac{1}{n}\sum_{i\in S} q_i \mathbf{x}_i \mathbf{x}_i^T\Big)^{-1}\Big(\frac{1}{n}\sum_{i\in S} q_i \mathbf{x}_i \epsilon_i\Big).$$

The lack of strict unbiasedness follows from the fact that ϵ_i and q_i may be dependent which occurs when the sampling scheme is nonignorable such that π_i and hence w_i depends on ϵ_i. The consistency follows from the fact that for fixed values $\hat{w}_s(\mathbf{x}_i)$, $(\mathbf{b}_{sp}-\beta)$ converges to $\mathbf{d}_{sp} = \Big(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{w}_s(\mathbf{x}_i)}\mathbf{x}_i\mathbf{x}_i^T\Big)^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{w}_s(\mathbf{x}_i)}\mathbf{x}_i\epsilon_i\Big)$ and $E_p(\mathbf{d}_{sp}) = \mathbf{0}$. Notice that this property holds for general functions $\hat{w}_s(\mathbf{x}_i)$, illustrating the robustness of the estimator to misspecification of the expectations $w_s(\mathbf{x}_i)$.

We conclude this section by providing a new model based justification for the PWLS estimator $\mathbf{b}_w$. Consider again the model defined by (3.5) and assume that the parameters indexing $w_s(\mathbf{x}_i)$ and the marginal pdf of $\mathbf{x}_i$, and hence $E_s(w_i)$, are distinct of β. By (2.5b),

$$\beta = arg\min_{\tilde{\beta}} E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 = arg\min_{\tilde{\beta}} E_s\Big(\frac{w_i(y_i - \mathbf{x}_i^T\tilde{\beta})^2}{E_s(w_i)}\Big) =$$

$$= arg\min_{\tilde{\beta}} E_s\Big(w_i(y_i - \mathbf{x}_i^T\tilde{\beta})^2\Big). \qquad \ldots(3.8)$$

Thus, replacing the expectation in (3.8) by the sample mean yields $\mathbf{b}_w$ as the optimal (least squares) solution.

Notice the difference between (3.6) and (3.8). In (3.8) we make use of the relationship $E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 = E_s\Big(w_i(y_i - \mathbf{x}_i^T\tilde{\beta})^2 \Big/ E_s(w_i)\Big)$ which holds for every vector $\tilde{\beta}$ and hence also for the optimal choice. The relationship (3.6) on the other hand only refers to the optimal choice of β since in general, $E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 \neq E_s\Big(\frac{w_i(y_i-\mathbf{x}_i^T\tilde{\beta})^2}{E_s(w_i|\mathbf{x}_i)}\Big)$. See the proof of (3.6) in the Appendix.

4. Monte Carlo Study with Simulated Data, Design and Estimators Considered

4.1 *Objectives.* In order to assess the performance of the parametric and semi-parametric estimators and compare them with the other estimators described in the Introduction, we designed a small Monte Carlo study. The study consists of generating populations from the normal density defined by (3.1), selecting samples with unequal selection probabilities and computing the various estimators for each of the selected samples. In addition to evaluating the performance of the regression estimators, we assess the performance of plausible variance estimators and the sample ignorability tests discussed in Section 2.2. In what follows we describe the various stages of the simulation study and the statistics computed in more detail. The results are reported in Section 5, where we present also the results obtained when sampling from some real data.

4.2 *Population values and sample selection.* We generated univariate populations of x-values of size N, $(N = 1000, 3000)$, from $Gamma(1,1)$. Values of Y were generated accordingly as

$$y_i = 1 + x_i + \epsilon_i; \quad \epsilon_i \sim N(0,1), \quad i = 1, ..., N, \qquad \ldots(4.1)$$

so that $\beta_0 = \beta_1 = \sigma^2 = 1$. Samples of size n, $(n = 100, 300)$, were selected by probability proportional to size systematic sampling, (Cochran, 1977, p.265), with the size variable z, defined in three different ways;

$$z_i = exp(-.1y_i - .08y_i^2 + .08x_i^2 + 0.3u_i) \qquad \ldots(4.2a)$$

$$z_i = 5 + 5y_i + 3y_i^2 + 10x_i + 3x_i^2 + u_i \qquad \ldots(4.2b)$$

$$z_i = 10x_i + 3x_i^2 + u_i \qquad \ldots(4.2c)$$

with $u_i \sim U(0,1)$. For N sufficiently large such that $\bar{z} = N^{-1}\sum_{i=1}^{N} z_i \approx constant$, the selection probabilities obtained from (4.2a) have expectations of the form (3.2a) but with the addition of a quadratic term of y in the exponent. The selection probabilities obtained from (4.2b) and (4.2c) satisfy for large populations the relationship (3.2b).

For the model defined by (4.1) and (4.2a) and large N, the conditional sample pdf is approximately normal with intercept $\beta_{0s} = (.9/1.16) \approx 0.776$, slope $\beta_{1s} = (1/1.16) \approx 0.862$ and residual variance $\sigma_s^2 = (1/1.16) \approx 0.862$. See PKR for the computation of these values. For the model defined by (4.1) and (4.2b) and large N, the conditional sample pdf is a mixture of normal densities, (equation 3.4). Under (4.2c), the sample selection scheme is ignorable given the $\mathbf{x}$'s, hence the conditional sample pdf is the same as in (4.1). We generated 100 populations with values (y_i, x_i, z_i) for each of the three cases defined by (4.2) and selected one sample from each population after randomly ordering the population units.

The random ordering was needed to secure independent samples since the x-values were fixed for all populations of the same size. Also, PPS systematic sampling from randomly ordered populations yields asymptotically independent measurements under the sample distribution, (see PKR).

4.3 *Estimators considered.* Estimators of (β_0, β_1) were computed for each sample using five different methods:

(A) OLS, (equation 1.1);

(B) PWLS, (equation 1.2); C - Skinner (1994) method, (see introduction);

(D) The 'parametric method' described in Section 3.1;

(E) The 'semi-parametric method', (equation 3.7).

Skinner's method has been implemented by fitting separately the (false) linear relationships $E_s(y_i|x_i, w_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 w_i$ and $E_p(w_i|x_i) = \delta_0 + \delta_1 x_i$, with γ_0, γ_1 and γ_2 estimated by OLS and δ_0 and δ_1 estimated by PWLS. This differs from the original method proposed by Skinner and deviates from his general philosophy that population models should be identified from the sample data, which is difficult to implement in a simulation study. Forcing linear expectations has the advantage in the present context that it induces a linear model for the population values which is the correct model assumed also for the other methods.

The computation of the parametric estimators requires the identification and estimation of the expectation $\pi_p(y, x) = E_p(\pi|y, x)$. This was implemented in two steps. First we applied the Box-Cox transformation, (Draper and Smith, 1981, p.225) in order to select between the linear expectation,

$$E_s(w|y, x) = \alpha_0 + \alpha_1 y + \alpha_2 y^2 + \theta_1 x + \theta_2 x^2 \qquad \ldots(4.3a)$$

and the exponential expectation,

$$E_s(w|y, x) = exp(\alpha_0 + \alpha_1 y + \alpha_2 y^2 + \theta_1 x + \theta_2 x^2). \qquad \ldots(4.3b)$$

In the second step we tested the significance of the quadratic terms in the selected model using ordinary F-tests. Notice that by (2.5a), the relationship (4.3b) is the correct one when z satisfies the model (4.2a), but the relationship (4.3a) is at best an approximation under either one of the models (4.2a) - (4.2c).

Denoting the resulting estimated relationship by $\hat{w}_s(y, x)$, the expectations $\pi_p(y_i, x_i)$ were estimated as $1/\hat{w}_s(y_i, x_i)$ exploiting the relationship (2.5a). The expectations $\pi_p(x_i)$ were estimated as $1/\hat{w}_s(x_i)$, with $\hat{w}_s(x_i) = \hat{E}_s(w_i|x_i)$ obtained from $\hat{w}_s(y_i, x_i)$ by integration over y. ($\hat{w}_s(x_i)$ is thus a function of β_0 and β_1). Substituting $\hat{\pi}_p(y_i, x_i)$ for $\pi_p(y_i, x_i)$ and $\hat{\pi}_p(x_i)$ for $\pi_p(x_i)$ in (2.2) (with $\mathbf{u}_i = y_i$, $\mathbf{v}_i = x_i$, and $f_p(y_i|x_i)$ defined by 4.1), and taking the product over the sample observations yields an approximation to the sample likelihood which was then maximized with respect to β_0, β_1 and σ^2, using PROC NLIN of the SAS software.

The computation of the semi-parametric estimators requires the identification and estimation of the expectation $w_s(x_i)$. This was implemented similar to the identification and estimation of the expectation $w_s(y_i, x_i)$ described above, selecting between 4.3a and 4.3b but without the y terms.

It is important to emphasize that in all cases the choice of the expectations was 'data driven', using an automated search procedure, with no bearing to the true relationships (4.2) used to generate the probabilities π_i.

4.4 *Variance estimation.* Estimation of the variances has two aspects. First is the estimation of the residual variance $\sigma^2 = Var(\epsilon_i)$ and second is the estimation of the variances $V(\hat{\beta})$.

The following estimates of σ^2 have been computed for the various estimation methods where we denote $\mathbf{x}_i = (1, x_i)$. (Hereafter we use the abbreviation "S" for Skinner's method, "P" for the parametric method and "SP" for the semi-parametric method.) $\hat{\sigma}^2(OLS) = \sum_{i=1}^n (y_i - \mathbf{x}_i \mathbf{b}_{ols})^2/n$; $\hat{\sigma}^2(PWLS) = \sum_{i=1}^n w_i(y_i - \mathbf{x}_i \mathbf{b_w})^2 \Big/ \sum_{i=1}^n w_i$; $\hat{\sigma}^2(S) = \sum_{i=1}^n w_i(y_i - \mathbf{x}_i \mathbf{b_s})^2 \Big/ \sum_{i=1}^n w_i$; $\hat{\sigma}^2(P)$ = the estimator obtained from the inverse information matrix, (computed with respect to β_0, β_1 and σ^2); $\hat{\sigma}^2(SP) = \sum_{i=1}^n w_i(y_i - \mathbf{x}_i \mathbf{b}_{sp})^2 \Big/ \sum_{i=1}^n w_i$. Note that by use of the weights $\{w_i\}$ for $\hat{\sigma}^2(SP)$ rather than the weights $\{q_i\}$ used for the computation of $\mathbf{b}_{sp}$, the estimator is consistent for σ^2 under the same mild conditions warranting the consistency of $\mathbf{b}_{sp}$. (See Section 3.2).

In order to estimate $V(\hat{\beta})$ we used bootstrap variance estimates. As mentioned before, the sample vector observations $(y_i, \mathbf{x}_i, \pi_i)$ can be viewed for large populations as independent realizations from the corresponding sample distribution, so that the use of the bootstrap method is well founded. To apply the method, we selected from each 'parent sample' 100 simple random samples with replacement of the same size as the parent sample. We computed the five regression estimates and the sample ignorability test statistics for each of these samples and then computed the variances over the corresponding 100 replicate estimates. Denoting in general by $\hat{\theta}_{(r)}$ any of the estimates computed from bootstrap sample r selected from a parent sample s , the corresponding variance estimate is computed as,

$$\hat{V}(\hat{\theta}) = \frac{1}{100} \sum_{r=1}^{100} (\hat{\theta}_{(r)} - \bar{\theta})(\hat{\theta}_{(r)} - \bar{\theta})^T; \quad \bar{\theta} = \frac{1}{100} \sum_{r=1}^{100} \hat{\theta}_{(r)}. \tag{4.5}$$

Comment: The bootstrap variances used in this study differ in principle from the bootstrap variances proposed in the sampling literature (see, Sitter, 1992 for an overview of bootstrap methods for sample surveys). The former assume independent observations from the sample distribution so that they coincide with the classical bootstrap estimators. The latter refer to the randomization distribution and hence are modified to account for finite population corrections and other aspects of the sampling process.

5. Results of Simulations and When Sampling from Real Data

5.1 *Results of simulation study.* In this section we report and discuss the results obtained for the simulation study described in Section 4. To save in space, we only present the results obtained for the case $N = 3000$, $n = 300$. The results obtained for the smaller sample size ($n = 100$) are in general very similar but occasionally less stable for some of the statistics considered.

Tables 1-3 contain summary statistics for the five estimation methods considered in this study. The three tables correspond to the three models for the design variable, 'exponential' (equation 4.2a), 'polynomial' (equation 4.2b) and 'ignorable' (equation 4.2c). The rows entitled $A(\hat{\beta}_0), A(\hat{\beta}_1)$ and $A(\hat{\sigma}^2)$ show the averages of the three estimators by method of estimation over the 100 samples. The rows entitled $SD(\hat{\beta}_0)$ and $SD(\hat{\beta}_1)$ show the corresponding empirical standard deviations whereas the rows entitled $\widehat{ASD}(\hat{\beta}_0)$ and $\widehat{ASD}(\hat{\beta}_1)$ show the averages of the bootstrap standard deviation estimates over the 100 parent samples (equation 4.5). The rows entitled Ar(k) show the average of the empirical sample correlations $r_k = Corr(\epsilon_i^k, w_i), k = 1, 2$ by method of estimation of ϵ_i^k. The rows entitled SDFT(k) and $\widehat{ASDFT}(k)$ $k = 1, 2$ show the standard deviations, and the average of the bootstrap standard deviation estimates of Fisher correlation transformations, whereas the rows entitled SIG FTS(k) show the proportion of samples for which the statistics $FTS(k)$ (equation 2.7) were found significant at the 5% level. The last row of each table, entitled POPMSE, shows the mean square of the prediction errors when predicting all the population values.

Table 1. MEAN VALUES AND STANDARD DEVIATIONS OF REGRESSION ESTIMATES AND SAMPLING IGNORABILITY TEST STATISTICS. POPULATION SIZE=3000, SAMPLE SIZE=300, $E(z_i|y_i, x_i)$ = "EXPONENTIAL".

	Method				
	OLS	PWLS	S	P	SP
$A(\hat{\beta}_0)$	0.78	1.02	0.91	1.01	1.02
$SD(\hat{\beta}_0)$	0.07	0.10	0.11	0.10	0.10
$\widehat{ASD}(\hat{\beta}_0)$	0.07	0.10	0.10	0.09	0.10
$A(\hat{\beta}_1)$	0.86	0.98	1.09	0.99	0.96
$SD(\hat{\beta}_1)$	0.06	0.08	0.10	0.07	0.08
$\widehat{ASD}(\hat{\beta}_1)$	0.06	0.08	0.10	0.07	0.08
$A(\hat{\sigma}^2)$	0.86	0.98	0.99	0.99	0.98
Ar(1)	0.71	0.67	0.63	0.67	0.68
$SDFT(1)$	0.13	0.15	0.15	0.14	0.15
$\widehat{ASDFT}(1)$	0.11	0.12	0.12	0.12	0.12
SIG FTS(1)	1.00	0.97	0.97	0.97	0.97
Ar(2)	0.40	-0.01	-0.02	-0.01	0.00
$SDFT(2)$	0.11	0.06	0.07	0.09	0.07
$\widehat{ASDFT}(2)$	0.11	0.06	0.07	0.08	0.07
SIG FTS(2)	1.00	0.06	0.08	0.07	0.04
POPMSE	1.15	1.01	1.02	1.01	1.01

Table 2. MEAN VALUES AND STANDARD DEVIATIONS OF REGRESSION ESTIMATES AND SAMPLING INGNORABILITY TEST STATISTICS. POPULATION SIZE=3000, SAMPLE SIZE=300, $E(z_i|y_i, x_i)$ = "POLYNOMIAL".

	Method				
	OLS	PWLS	S	P	SP
$A(\hat{\beta}_0)$	1.48	1.02	1.06	1.01	1.01
$SD(\hat{\beta}_0)$	0.10	0.14	0.13	0.11	0.11
$\widehat{ASD}(\hat{\beta}_0)$	0.09	0.13	0.12	0.10	0.10
$A(\hat{\beta}_1)$	0.92	0.99	0.99	0.98	0.99
$SD(\hat{\beta}_1)$	0.04	0.06	0.05	0.04	0.04
$\widehat{ASD}(\hat{\beta}_1)$	0.04	0.06	0.05	0.04	0.04
$A(\hat{\sigma}^2)$	1.01	0.99	1.00	1.03	1.03
Ar(1)	-0.41	-0.35	-0.35	-0.35	-0.34
$SDFT(1)$	0.08	0.05	0.04	0.03	0.05
$\widehat{ASDFT}(1)$	0.06	0.07	0.05	0.04	0.05
SIG FTS(1)	1.00	1.00	1.00	1.00	1.00
Ar(2)	0.10	-0.10	-0.10	-0.10	-0.09
$SDFT(2)$	0.08	0.05	0.05	0.05	0.05
$\widehat{ASDFT}(2)$	0.07	0.05	0.05	0.04	0.04
SIG FTS(2)	0.15	0.60	0.58	0.55	0.55
POPMSE	1.17	1.01	1.01	1.01	1.01

Table 3. MEAN VALUES AND STANDARD DEVIATIONS OF REGRESSION ESTIMATES AND SAMPLING IGNORABILITY TEST STATISTICS. POPULATION SIZE=3000, SAMPLE SIZE=300, $E(z_i|y_i, x_i)$ = "IGNORABLE".

	Method				
	OLS	PWLS	S	P	SP
$A(\hat{\beta}_0)$	1.00	1.01	1.00	1.00	1.00
$SD(\hat{\beta}_0)$	0.10	0.16	0.12	0.11	0.11
$\widehat{ASD}(\hat{\beta}_0)$	0.10	0.16	0.11	0.11	0.11
$A(\hat{\beta}_1)$	1.00	0.99	1.00	1.00	1.00
$SD(\hat{\beta}_1)$	0.04	0.08	0.04	0.04	0.04
$\widehat{ASD}(\hat{\beta}_1)$	0.04	0.07	0.04	0.05	0.04
$A(\hat{\sigma}^2)$	1.00	0.99	1.01	1.00	1.00
Ar(1)	0.000	0.000	0.001	0.001	0.000
$SDFT(1)$	0.05	0.01	0.05	0.04	0.05
$\widehat{ASDFT}(1)$	0.04	0.01	0.04	0.05	0.04
SIG FTS(1)	0.04	0.08	0.04	0.05	0.04
Ar(2)	0.000	-0.011	-0.004	-0.003	0.000
$SDFT(2)$	0.06	0.05	0.05	0.05	0.06
$\widehat{ASDFT}(2)$	0.05	0.05	0.06	0.05	0.05
SIG FTS(2)	0.09	0.10	0.08	0.09	0.08
POPMSE	1.01	1.02	1.01	1.01	1.01

The main findings from the three tables are as follows:

(1) The OLS estimators are highly biased for the two informative sampling schemes. (The true regression coefficients are $\beta_0 = \beta_1 = 1$ and also $\sigma^2 = 1$.) The

bias in the estimation of the regression coefficients translates into a relatively large bias in the prediction of the population values (the statistics POPMSE).

(2) The four other estimation methods eliminate the OLS biases or at least reduce them substantially. Notice the close values of the POPMSE statistics to the residual variance $\sigma^2 = 1$, which of course is supported by theory.

(3) The standard deviations (SD) of the OLS coefficients are always the lowest or among the lowest. This result is expected because for the exponential and ignorable cases the regression in the sample is again linear, (although with different coefficients in the exponential case), for which the OLS is optimal.

(4) The SD of the PWLS estimators are in most cases larger, and in some cases much larger than the SD of the other three estimators that account for sample selection effects. (This holds particularly for the case $n = 100$ not shown here). The explanation to this outcome is that unlike the PWLS method, the other methods identify and exploit the relationship between the sampling weights and the survey variables. Notice the good performance of Skinner's method despite the fact that the method was applied by forcing linear expressions for the expectations $E_s(y_i|\mathbf{x}_i, w_i)$ and $E_p(w_i|\mathbf{x}_i)$. Note also that the parametric and semi-parametric methods perform equally well albeit the extra information employed by the former method. This result is important considering that the semi-parametric method is much simpler but it may not hold for different models (say, nonnormal distributions) or parameter values.

(5) The performance of the bootstrap SD estimates is generally satisfactory for all the methods and statistics, except for $\widehat{ASDFT}(1)$ in the exponential case. This is an important outcome because except for the OLS and PWLS estimators, the other methods are two-step procedures for which the use of resampling techniques for variance estimation seems very appealing.

(6) The test statistics FTS(1) that employ the empirical sample residuals perform well, having very large powers for the informative sampling schemes and yielding empirical significance levels that are close to the nominal 5% level for the ignorable scheme. Testing the normality of $FT(1)$ in the latter case yields high p-values except for the statistics that use the OLS and PWLS residuals where the p-values are .08 and .09 respectively. The statistics $FTS(2)$ on the other hand that use the squared residuals behave very erratically even though the normality test statistics for $FT(2)$ yields high p-values. We are not able at this stage to provide a satisfactory explanation to this outcome.

5.2 *A real data example.* In order to further compare the parametric and semi-parametric estimators with the other methods, we use a real data set, previously analyzed by Korn and Graubard (1995). The data were collected as part of the 1988 U.S. National Maternal and Infant Health Survey which uses a stratified random sample of vital records corresponding to live births, late fetal deaths and infant deaths in the United States. The strata were defined based on the mother's race and child's birthweight, with different sampling fractions in different strata. Korn and Graubard use the sample data corresponding to the

live births in order to illustrate that OLS and PWLS estimators may differ when the population model is misspecified. Here we use the same data for studying the performance of the OLS, PWLS and the newly proposed estimators in a somewhat more complicated context than considered in the simulation study. (Skinner's method is not applied).

To this end, we considered the sample data as 'population' and selected independent samples with probabilities proportional to the original selection probabilities. For each sample we estimated the regression of *birthweight* (Y, measured in grams) on the first, second and third powers of the *gestational age* (x, measured in weeks), using the four estimation methods. All three powers of x were included as regressors to ensure an appropriate fit at the population (original sample) level. The populational model (fitted by OLS) is,

$$Y_i = 17886 - 1827.7x_i + 61.2x_i^2 - 0.61x_i^3 + \epsilon_i; \quad i = 1, ..., 9447. \qquad ...(5.1)$$

All the coefficients are highly significant (p-value$\leq 10^{-4}$) with $R^2 = 0.61$ and $\sigma_\epsilon^2 = 603.2$. (We dropped 506 sample records because of missing data.) Note that the model defined by (5.1) is not representative of the actual population of vital records because of the informativeness of the original sampling scheme. For the original sample, $r(1) = Corr(\tilde{\epsilon}_i, w_i) = 0.30$ where $\tilde{\epsilon}_i$ are the estimated residuals obtained from (5.1).

Application of the parametric method requires a prior specification of the form of the expectations $\pi(y_i, x_i) = E_p(\pi_i|y_i, x_i)$ and $\pi(x_i) = E_p(\pi_i|x_i)$. By the given structure of the original sampling design, $\pi(y_i, x_i) = c_1$ for $y < 1500$, $\pi(y_i, x_i) = c_2$ for $1500 \leq y < 2500$ and $\pi(y_i, x_i) = c_3$ for $y \geq 2500$, irrespective of the x-values, where c_1, c_2 and c_3 are three constants which were re-estimated for each of the new samples s. (The constants c_j are actually the unconditional means over the two races).

Denoting the polynomial of x_i in the right hand side of (5.1) by $g(x_i)$ and assuming $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, it follows that

$$\pi(x_i) = E_{\epsilon_i|x_i}\{E_p(\pi_i|y_i, x_i)\} = c_1\Phi\Big[\frac{1500 - g(x_i)}{\sigma_\epsilon}\Big] +$$

$$c_2\Big(\Phi\Big[\frac{2500 - g(x_i)}{\sigma_\epsilon}\Big] - \Phi\Big[\frac{1500 - g(x_i)}{\sigma_\epsilon}\Big]\Big) + c_3\Big(1 - \Phi\Big[\frac{2500 - g(x_i)}{\sigma_\epsilon}\Big]\Big). \quad ...(5.2)$$

Substituting $\pi(y_i, x_i)$ and $\pi(x_i)$ given by (5.2) in (2.2), (with $u_i = y_i$ and $v_i = x_i$) and taking the product over the sample records defines the sample likelihood. Notice that the unknown regression parameters appear both in the population normal pdf of $\epsilon_i = y_i - g(x_i)$ and in the functions Φ defining $\pi(x_i)$.

The maximization of the likelihood (using PROC NLIN of SAS) over all the regression parameters turned out to be very unstable. (See also the discussion below Table 4.) Hence we set $\sigma_\epsilon^2 = \hat{\sigma}_\epsilon^2(WLS)$ and maximized the likelihood with respect to the four regression coefficients by fixing in sequence two of the

coefficients and maximizing with respect to the other two coefficients until convergence.

Application of the semi-parametric method requires the specification and estimation of the expectation $w_s(x_i) = E_s(w_i|x_i)$, (equation 3.7). Following similar considerations we find that $E_s(w_i|y_i, x_i) = [1/E_p(\pi_i|y_i, x_i)]$ is constant in the same ranges of the Y-values as for $\pi(y, x)$. Hence, $w_s(x_i)$ has the same form as (5.2), but with different constants c_i and with the cumulative normal pdf postulated for the sample residuals ϵ_{si}. For estimating $w_s(x_i)$, we estimated $g(x_i)$ and $\sigma^2_{s\epsilon} = Var_s(\epsilon_{si})$ by OLS and approximated $\Phi(t)$ as $\hat{\Phi}(t) \approx 1 - 0.5(1 + 0.1968t + 0.1152t^2 + 0.0003t^3 + 0.0195t^4)^{-4}$ (Johnson and Kotz, 1970, Vol 1, equation 27). Note that $w_s(x_i)$ refers to the sample distribution and hence the use of OLS.

Table 4 shows the means and standard deviations of the estimated coefficients as obtained over 100 samples selected by Poisson sampling. We use the same notation as in the previous tables with β_j defining the coefficient of x_i^j. The mean sample size is $E(n) = 233.6$, with standard deviation $SD(n) = 15.3$. Below we discuss also the results obtained when increasing the sample size such that $E(n) = 870.3$ and $SD(n) = 25.4$.

Table 4. MEAN VALUES AND STANDARD DEVATIONS OF REGRESSION ESTIMATES OVER 100 SAMPLES SELECTED FROM REAL DATA.

		Method			
True coefficients	Means, SD	OLS	PWLS	P	SP
$\beta_0 = 17886$	$A(\hat{\beta}_0)$	13625.2	19035.7	17630.8	17556.3
	$SD(\hat{\beta}_0)$	4539.3	8106.0	7217.5	7332.8
$\beta_1 = -1827.7$	$A(\hat{\beta}_1)$	-1382.8	-1952.7	-1813.7	-1809.5
	$SD(\hat{\beta}_1)$	458.4	764.1	686.3	689.7
$\beta_2 = 61.2$	$A(\hat{\beta}_2)$	45.9	65.50	60.21	61.07
	$SD(\hat{\beta}_2)$	15.13	23.51	21.26	21.23
$\beta_3 = -0.61$	$A(\hat{\beta}_3)$	-0.45	-0.66	-0.60	-0.62
	$SD(\hat{\beta}_3)$	0.16	0.24	0.22	0.21

The results shown in table 4 favour the use of the parametric and semi-parametric estimators quite strikingly. The OLS estimators are extremely biased (although with the smallest variances) indicating the high degree of sample informativeness. The WLS estimators have the largest variances and biases in the order of 7%, although the biases are not significant based on the conventional t-statistics. The parametric and semi-parametric estimators on the other hand have negligible biases (less then 2%) and smaller variances then the WLS estimators, with the parametric estimators performing only slightly better.

We repeated the same analysis using sample sizes of mean 870.3 and standard deviation 25.4. As expected, the SD of all the estimators decrease by a factor of about $0.5 \approx (233.6/870.3)^{1/2}$, with the biases of the OLS estimators remaining almost unchanged. Unexpectedly, the biases of the WLS estimators also decrease only marginally although they are still nonsignificant at the 5% level. Increasing

the sample size decreases the (already negligible) biases of the semi-parametric estimators even further, but the parametric estimators of β_2 and β_3 become biased with relative biases of 8% and 17% respectively. (The means of $\hat{\beta}_0(P)$ and $\hat{\beta}_1(P)$ are almost the same as for the smaller sample sizes).

The failure of the parametric method to yield unbiased estimators for the last two coefficients in the case of the larger sample sizes can be attributed to numerical problems in reaching the global maximum values. Plotting the likelihood as a function of β_2 and β_3, (with β_0 and β_1 held fixed), reveals that it is quite flat, particularly with respect to β_3. Although such problems are not unexpected with complex likelihoods as in the present study, it is not easy to handle them in an "automatic" simulation study. The semi-parametric estimators on the other hand are free of any computational problems, at least in the present application, and perform very well under both sample sizes.

6. Conclusions

In this study we consider two new classes of regression estimators and another class proposed by Skinner (1994), for regression analysis from complex surveys. All three methods account for the relationship between the sample selection probabilities and the regression variables. These estimators are shown to perform well in the empirical study and outperform the PWLS estimator which is the estimator in common use. The choice between the three alternative estimators is not clearcut. The parametric method has the theoretical advantage of using information about the form of the population distribution which permits in principle the application of likelihood based inference, but as illustrated with the real data, the derivation of the maximum likelihood estimators can become quite complicated. Also, the robustness of the parametric method against possible misspecification of the population model has yet to be studied. Skinner's method only uses sample models, which is an appealing property, but the combination of these models may yield incomprehensible population models. The semi-parametric method is somewhere in between the other two methods as it only uses information about the first two moments in the population. This method has the further advantage of being simple and robust, and it is shown to perform very well in the empirical study with both the simulated and the real data.

Another important finding of our study is the good performance of the bootstrap variance estimators. Admittedly, we were surprised to see how well the procedure works given the complicated nature of the sampling process and most of the statistics considered, derived as combinations of two separate estimation processes. The use of the bootstrap method in conjunction with the randomization (design) distribution is a long standing problem. As noted in Section 4.4, by focusing instead on the sample distribution, the sample measurements

obtained from single stage sampling can often be considered as independent, thus validating the use of the bootstrap method. Finally, we mention the good performance of the statistics $FTS(1)$ in testing the sampling ignorability. The statistics $FTS(2)$ on the other hand perform poorly, and the use of them and test statistics that employ higher powers of the residual terms requires further investigation.

The empirical study of this paper is of limited scope, but we hope that the good results obtained for the parametric, and in particular the semi-parametric estimators will encourage further theoretical and empirical research in these directions.

Appendix

(A1) PROOF OF (2.2): Define $I_i = 1$ if $i \in S$, $I_i = 0$, otherwise

$$f_s(\mathbf{u}_i|\mathbf{v}_i) \stackrel{def}{=} f_p(\mathbf{u}_i|\mathbf{v}_i, I_i = 1)$$
$$= Pr(I_i = 1|\mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i|\mathbf{v}_i) \Big/ Pr(I_i = 1|\mathbf{v}_i)$$
$$= E_p[Pr(I_i = 1|\mathbf{u}_i, \mathbf{v}_i, \pi_i)|\mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i|\mathbf{v}_i) \Big/ E_p[Pr(I_i = 1|\mathbf{v}_i, \pi_i)|\mathbf{v}_i]$$
$$= E_p\left[\pi_i|\mathbf{u}_i, \mathbf{v}_i\right] f_p(\mathbf{u}_i|\mathbf{v}_i) \Big/ E_p\left[\pi_i|\mathbf{v}_i\right]$$

since $Pr(I_i = 1|..., \pi_i) = \pi_i$.

(A2) PROOF OF (2.3): By (2.2),

$$f_s(w_i|\mathbf{v}_i) = E_p(\pi_i|w_i, \mathbf{v}_i) f_p(w_i|\mathbf{v}_i) \Big/ E_p(\pi_i|w_i, \mathbf{v}_i)$$
$$= (1/w_i) f_p(w_i|\mathbf{v}_i) / E_p(\pi_i|\mathbf{v}_i).$$

Hence $E_s(w_i|\mathbf{v}_i) = 1/E_p(\pi_i|\mathbf{v}_i)$ which is the relationship (2.5a). Substituting $E_p(\pi_i|\mathbf{v}_i) = 1/E_s(w_i|\mathbf{v}_i)$ and $E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i) = 1/E_s(w_i|\mathbf{u}_i, \mathbf{v}_i)$ in (2.2) gives (2.3).

(A3) PROOF OF (2.4): By (2.3),

$$E_p(\mathbf{u}_i|\mathbf{v}_i) = E_s[\mathbf{u}_i E_s(w_i|\mathbf{u}_i, \mathbf{v}_i)|\mathbf{v}_i] / E_s(w_i|\mathbf{v}_i)$$
$$= E_s(w_i \mathbf{u}_i|\mathbf{v}_i) / E_s(w_i|\mathbf{v}_i).$$

(A4) PROOF OF (3.6): Denote $\tilde{\epsilon}_i = (y_i - \mathbf{x}_i^T \tilde{\beta})$ for given $\tilde{\beta}$ and $w_s(\mathbf{x}_i) = E_s(w_i|\mathbf{x}_i)$.

$$E_s[w_i \tilde{\epsilon}_i^2 / w_s(\mathbf{x}_i)] = E_s\left(E_s[w_i \tilde{\epsilon}_i^2 / w_s(\mathbf{x}_i)]\Big|\mathbf{x}_i\right) = E_s\left(E_p[\tilde{\epsilon}_i^2]\Big|\mathbf{x}_i\right)$$
$$= E_s\left(E_p[y_i - \mathbf{x}_i^T\beta + \mathbf{x}_i^T\beta - \mathbf{x}_i^T\tilde{\beta}]^2\Big|\mathbf{x}_i\right) = \sigma^2 + E_s[\mathbf{x}_i^T(\beta - \tilde{\beta})]^2$$
$$= \min_{\tilde{\beta}} E_p(y_i - \mathbf{x}_i^T\tilde{\beta})^2 + E_s[\mathbf{x}_i^T(\beta - \tilde{\beta})]^2.$$

References

CHAMBERS, R. L., DORFMAN, A. H., AND WANG, S. (1998). Limited information likelihood analysis of survey data. *Jour. Royal Statist. Soc.*, Series B, **60**, 397-411.

COCHRAN, W. G. (1977). *Sampling Techniques,* (third edition). Wiley, New York.

DRAPER, N. R. AND SMITH, H. (1981). *Applied Regression Analysis,* (second edition). Wiley, New York.

JOHNSON, N. I. AND KOTZ, S. (1970). *Continuous Univariate Distributions* -1. Houghton Mifflin Company, Boston.

KENDALL, M. G. AND STUART, A. (1969). *The Advanced Theory of Statistics,* Vol 1, (third edition). Hafner Publishing company, New York.

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

KRIEGER, A. M. AND PFEFFERMANN, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology,* **18**, 225-239.

KRIEGER, A. M. AND PFEFFERMANN, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics,* **13**, 123-142.

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

PFEFFERMANN, D., KRIEGER, A. M. AND RINOTT, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica,* **8**, 1087-1114.

RAO, J. N. K. (1984). Conditional inference in survey sampling. *Survey Methodology,* **11**, 15-31.

SITTER, R. V. (1992). Bootstrap methods for survey data. *Canadian Journal of Statistics,* **20**, 135-154.

SKINNER, C. J. (1994). Sample models and weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 133-142.

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

SMITH, T. M. F. (1981). Regression analysis for complex surveys. In: D. Krewski, R. Platek and J.N.K. Rao, Eds. *Current Topics in Survey Sampling.* Academic Press, New York, pp. 267-292.

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

DANNY PFEFFERMANN AND MICHAIL SVERCHKOV
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM, ISRAEL 91905
e-mail: msdanny@mscc.huji.ac.il/ msmisha@mscc.huji.ac.il