PREDICTION OF FINITE POPULATION TOTALS BASED ON THE SAMPLE DISTRIBUTION

MICHAIL SVERCHKOV, DANNY PFEFFERMANN

ABSTRACT

In this article we study the use of the sample distribution for the prediction of finite population totals under single-stage sampling. The proposed predictors condition on the sample values of the target outcome variable, the sampling weights of the sample units and possibly on known population values of auxiliary variables.

The prediction problem is solved by estimating the expectation of the outcome values for units outside the sample as a function of the corresponding expectation under the sample distribution and the sampling weights. The prediction variance is estimated by a combination of an inverse sampling procedure and the bootstrap method. An important outcome of the present analysis is that several familiar estimators in common use are shown to be special cases of the proposed approach, thus providing them a new interpretation. The performance of the new and some old predictors in common use is evaluated and compared by a Monte Carlo simulation study using a real data set.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M03/06

University
of Southampton

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 79-92
Statistics Canada

Prediction of Finite Population Totals Based on the Sample Distribution

MICHAIL SVERCHKOV and DANNY PFEFFERMANN[1]

ABSTRACT

This article studies the use of the sample distribution for the prediction of finite population totals under single-stage sampling. The proposed predictors employ the sample values of the target study variable, the sampling weights of the sample units and possibly known population values of auxiliary variables. The prediction problem is solved by estimating the expectation of the study values for units outside the sample as a function of the corresponding expectation under the sample distribution and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. An interesting outcome of the present analysis is that several familiar estimators in common use are shown to be special cases of the proposed approach, thus providing them a new interpretation. The performance of the new and some old predictors in common use is evaluated and compared by a Monte Carlo simulation study using a real data set.

KEY WORDS: Bootstrap; Design consistency; Informative sampling; Sample-complement distribution.

1. INTRODUCTION

The sample distribution is the parametric distribution of the outcome values for units included in the sample. This distribution is different from the population distribution if the sample selection probabilities are correlated with the values of the study variable even when conditioning on the values of concomitant variables included in the population model. It is also different from the randomization (design) distribution that accounts for all the possible sample selections with the population values held fixed. The sample distribution is defined and discussed with examples in Pfeffermann, Krieger and Rinott (1998), and is further investigated in Pfeffermann and Sverchkov (1999) who use it for the estimation of linear regression models. Krieger and Pfeffermann (1997) use the sample distribution for testing population distribution functions and Pfeffermann and Sverchkov (2003a) discuss its use for fitting Generalized Linear Models. Chambers, Dorfman and Sverchkov (2003) utilize the sample distribution for nonparametric estimation of regression models, and Kim (2002) and Pfeffermann and Sverchkov (2003b) apply it for small area estimation problems.

In this article we study the use of the sample distribution for the prediction of finite population totals under single-stage sampling. It is assumed that the population outcome values (the *y*-values) are random realizations from some distribution that conditions on known values of auxiliary variables (the *x*-values). The problem considered is the prediction of the population total *Y* based on the sample *y*-values, the sampling weights for units in the sample and the population *x*-values. The use of the sample distribution permits conditioning on all these values, which is not possible under the randomization (design) distribution, and the prediction of *Y* is equivalent therefore to the prediction of the *y*-values for units outside the sample.

The prediction problem is solved by estimating the conditional expectation of the *y*-values (given the *x*-values) for units outside the sample as a function of the conditional sample expectation (the expectation under the sample distribution) and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. As it turns out, several familiar estimators in common use and in particular, classical design based estimators are special cases of the proposed procedure, thus providing them a new interpretation. The performance of the new and old predictors is evaluated and compared by mean of a Monte Carlo simulation study using a real data set.

2. THE SAMPLE AND SAMPLE-COMPLEMENT DISTRIBUTIONS

2.1 The Sample Distribution

Suppose that the population values $\{\mathbf{y}, X\} = \{(y_1 ... y_N)', [\mathbf{x}_1 ... \mathbf{x}_N]'\}$ are random realizations with conditional probability density function (*pdf*) $f_p(y_i \mid \mathbf{x}_i)$ that may be discrete or continuous. The *y*-values are assumed to be scalars but the *x*-values can be vectors. We consider single stage sampling with sample inclusion probabilities $\pi_i = \Pr(i \in s) = g(\mathbf{y}, X, Z, i)$ for some function *g*, where *Z* defines the population values of design variables used for the sampling process. Note that the *y*-values are random and we also consider the design variables as random so that the

[1] Michail Sverchkov, The Bureau of Labor Statistics, Washington D.C. 20212, U.S.A.; Danny Pfeffermann, Hebrew University, Israël and University of Southampton, U.K.

g-values are random as well. Let $I_i = 1$ if $i \in s$ and $I_i = 0$, if $i \notin s$. The conditional marginal *sample pdf* is defined as,

$$f_s(y_i|\mathbf{x}_i) \overset{\text{def}}{=} f(y_i|\mathbf{x}_i, I_i = 1) = \frac{\Pr(I_i = 1|\, y_i, \mathbf{x}_i) f_p(y_i|\, \mathbf{x}_i)}{\Pr(I_i = 1|\, \mathbf{x}_i)} \tag{2.1}$$

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 actual sample selection probability $\pi_i = g(\mathbf{y}, X, Z, i)$ (see Remark 1 below). 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. When the sample distribution differs from the population distribution it becomes *informative*, and the sampling scheme can not be ignored at the inference process.

Remark 1. It is important to emphasize that the definition and use of the sample distribution does not assume that the sample selection probabilities are function of only $(y_i, \mathbf{x}_i)$. As mentioned earlier and highlighted by expressing the selection probabilities as $\pi_i = g(\mathbf{y}, X, Z, i)$, the actual selection probabilities may depend on all the population values $(\boldsymbol{y}, X, Z)$. However, as shown in Pfeffermann and Sverchkov (1999), $E_p(\pi_i \,|\, y_i, \mathbf{x}_i) = \Pr(I_i = 1 \,|\, y_i, \mathbf{x}_i)$. Thus, although the selection probabilities may depend on all the population values $(\boldsymbol{y}, X, Z)$, for given values $(y_i, \mathbf{x}_i)$ they equal $\Pr(I_i = 1 \,|\, y_i, \mathbf{x}_i)$ 'on average'. In fact, π_i may not depend directly on $\boldsymbol{y}$ at all and only be a function of (X, Z), and still the expectation $E_p(\pi_i \,|\, y_i, x_i)$ equals $\Pr(I_i = 1 \,|\, y_i, \mathbf{x}_i)$. The reason why the expectation may depend on y_i in this case is that *Z* may be correlated with ***y***. For example, the 1999 Canadian Workplace and Employee Survey uses a disproportionate stratified sample with the strata defined by region, activity, and the size of the workplace. The size information is obtained from tax records from 1998; see, Patak, Hidiroglou and Lavallée (2000) for details. When modeling the payrolls in 1999 against the number of employees, the sampling design is found to be informative, which is explained by the fact that the stratification is based in part on the size obtained from the tax records in the previous year, which are correlated with the payroll the year after. See Fuller (2003) for details of the analysis.

The discussion above should not be understood to mean that π_i is never a function of $(y_i, \mathbf{x}_i)$ only. A classical example for the latter case is retrospective sampling. Thus, in a case control study, the selection probabilities of the cases and controls usually only depend on the respective *y* and *x* values (and often just on the *y* values). In the empirical study of this paper we use a real data set where the sample was drawn by a disproportionate stratified sample with the strata boundaries defined by the values of the dependent variable.

In what follows we regard the probabilities π_i as random realizations of the random variable $g(\mathbf{y}, X, Z, i)$. Let $w_i = 1/\pi_i$ define the sampling weight of unit *i*. The following relationships, established in Pfeffermann and Sverchkov (1999) hold for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$, with E_p and E_s defining expectations under the population and sample *pdf*s respectively. (As a special case, $\mathbf{u}_i = y_i$, $\mathbf{v}_i = \mathbf{x}_i$).

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

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

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

It follows from (2.4) that

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) $$E_s(w_i) = \frac{1}{E_p(\pi_i)}. \tag{2.5}$$

For a detailed discussion of the sample distribution with illustrations, see Pfeffermann *et al.* (1998).

2.2 The Sample-Complement Distribution

Similar to (2.1), we define the conditional *pdf* for units outside the sample as,

$$f_c(y_i|\mathbf{x}_i) \overset{\text{def}}{=} f_p(y_i|\mathbf{x}_i, I_i = 0) = \frac{\Pr(I_i = 0|y_i, \mathbf{x}_i) f_p(y_i|\, \mathbf{x}_i)}{\Pr(I_i = 0|\mathbf{x}_i)}. \tag{2.6}$$

The relationships (2.2) – (2.5) and the equality $\Pr(I_i = 0 \,|\, \mathbf{u}_i, \mathbf{v}_i) = 1 - \Pr(I_i = 1|\, \mathbf{u}_i, \mathbf{v}_i) = 1 - E_p(\pi_\mathrm{i} \,|\mathbf{u}_i, \mathbf{v}_i)$ imply the following representations of the sample-complement distribution for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$.

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p[(1-\pi_i)|\mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]}$$

$$= \frac{E_p[(1-\pi_i)|\mathbf{u}_i, \mathbf{v}_i]}{E_p[(1-\pi_i)|\mathbf{v}_i]} \frac{E_p[\pi_i|\mathbf{v}_i]}{E_p[\pi_i|\mathbf{u}_i, \mathbf{v}_i]} f_s(\mathbf{u}_i|\mathbf{v}_i) \tag{2.7}$$

$$f_c(\mathbf{u}_i|\mathbf{v}_i)=\frac{E_s[(w_i-1)|\mathbf{u}_i,\mathbf{v}_i]\,f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s[(w_i-1)|\mathbf{v}_i]}. \qquad (2.8)$$

(Equation (2.8) follows by application of (2.5a) to the second expression in (2.7)). Also, by (2.8) and the first equation in (2.7),

$$E_c(\mathbf{u}_i|\mathbf{v}_i)=\frac{E_p[(1-\pi_i)\mathbf{u}_i|\mathbf{v}_i]}{E_p[(1-\pi_i)|\mathbf{v}_i]}=\frac{E_s[(w_i-1)\mathbf{u}_i|\mathbf{v}_i]}{E_s[(w_i-1)|\mathbf{v}_i]}. \qquad (2.9)$$

Remark 2. In practical applications the sampling fraction is often very small and hence the sample selection probabilities are small for at least most of the population units. If $\pi_i<\delta$ with probability 1,

$$\begin{aligned} f_c(\mathbf{u}_i|\mathbf{v}_i) &= \frac{E_p[(1-\pi_i)|\mathbf{u}_i,\mathbf{v}_i]\,f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]} \\ &= f_p(\mathbf{u}_i|\mathbf{v}_i)+\\ &\qquad \frac{E_p\{[E_p(\pi_i|\mathbf{v}_i)-\pi_i]|\mathbf{u}_i,\mathbf{v}_i\}\,f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]} \\ &= f_p(\mathbf{u}_i\mid\mathbf{v}_i)(1+\Delta) \qquad (2.10) \end{aligned}$$

where $-\delta<\Delta<\delta/(1-\delta)$. It follows from (2.10) that for δ sufficiently small, the difference between the population *pdf* and the sample-complement *pdf* is accordingly small, which is not surprising.

3. OPTIMAL PREDICTION OF FINITE POPULATION TOTALS

Let $Y=\sum_{i=1}^{N}y_i$ define the population total. The problem considered is how to predict *Y* based on the sample data and possibly population values of auxiliary variables. Denote the 'design information' available for prediction by $D_s=\{(y_i,w_i),i\in s;\ (\mathbf{x}_j,I_j),\ j=1...N\}$ and let $\hat{Y}=\hat{Y}(D_s)$ define the predictor. The MSE of $\hat{Y}$ with respect to the population *pdf* given D_s is,

$$\begin{aligned} \mathrm{MSE}(\hat{Y}\mid D_s) &= E_p[(\hat{Y}-Y)^2|D_s] \\ &= E_p\{[\hat{Y}-E_p(Y|D_s)]^2|D_s\}+V_p(Y\mid D_s) \\ &= [\hat{Y}-E_p(Y|D_s)]^2+V_p(Y|D_s) \qquad (3.1) \end{aligned}$$

since $[\hat{Y}-E_p(Y\,|D_s)]$ is fixed given D_s. It follows from (3.1) that $\mathrm{MSE}(\hat{Y}\mid D_s)$ is minimized when $\hat{Y}=E_p(Y\,|D_s)$. The latter expectation can be decomposed as,

$$\begin{aligned} E_p(Y|D_s) &= \textstyle\sum_{i=1}^{N}E_p(y_i|D_s) \\ &= \sum_{i\in s}E_p(y_i|D_s,I_i=1)+\sum_{j\notin s}E_p(y_j|D_s,I_j=0) \\ &= \sum_{i\in s}y_i+\sum_{j\notin s}E_c(y_j|D_s) \\ &= \sum_{i\in s}y_i+\sum_{j\notin s}E_c(y_j|\mathbf{x}_j) \qquad (3.2) \end{aligned}$$

where in the last equality we assume that y_j for $j\notin s$ and D_s are uncorrelated given x_j. The prediction problem reduces therefore to the estimation of the expectations $E_c(y_j\mid\mathbf{x}_j)$. In section 4 we consider semi-parametric estimation of these expectations.

4. SEMI-PARAMETRIC PREDICTION OF FINITE POPULATION TOTALS

Suppose that the sample-complement model takes the form,

$$\begin{aligned} y_j &= C_\beta(\mathbf{x}_j)+\varepsilon_j, \\ E_c(\varepsilon_j|\mathbf{x}_j) &= 0,\ E_c(\varepsilon_j^2|\mathbf{x}_j)=\sigma^2 v(\mathbf{x}_j), \\ E_c(\varepsilon_k\varepsilon_j|\mathbf{x}_k,\mathbf{x}_j) &= 0,\ k\neq j \qquad (4.1) \end{aligned}$$

where $C_\beta(\mathrm{x})$ is a known (possibly nonlinear) function of **x** that depends on an unknown vector parameter β. The variances $\sigma^2\ v(\mathbf{x_j})$ are assumed known except for σ^2.

Remark 3. In actual applications the model (4.1) can be identified by a two-step procedure, utilizing the equality $E_c(y_i|\mathbf{x}_i)=E_s(r_iy_i|\mathbf{x}_i)$ with $r_i=(w_i-1)/E_s[(w_i-1)|\,\mathbf{x}_i]$ (follows from Equation 2.9). First, estimate $E_s(w_i\,|\,\mathbf{x}_i)$ and hence r_i by regressing w_i against $\mathbf{x}_i$ using the sample data. Let $\hat{r}_i=(w_i-1)/[\hat{E}_s(w_i\,|\,\mathbf{x}_i)-1]$ and transform $y_i^*=\hat{r}_iy_i$. Second, study the relationship in the sample between y_i^* and $\mathbf{x}_i$ for identifying the form of $C_\beta(\mathbf{x}_i)$. See Pfeffermann and Sverchkov (1999, 2003a) for examples of estimating $E_s(w_i\,|\,\mathbf{x}_i)$. A similar procedure can be applied for identifying the variance function $v(\mathbf{x}_i)$, using the empirical residuals $\hat{\varepsilon}_i=y_i-\hat{E}_s(\hat{r}_iy_i\,|\,\mathbf{x}_i)$.

The function $C_\beta(\mathbf{x}_j)$ in (4.1) with the true vector parameter β satisfies for all $j\notin s$,

$$\begin{aligned} C_\beta(\mathbf{x}_j) &= \arg\min_{C_{\tilde{\beta}}(\mathbf{x}_j)} E_c\left(\frac{[y_j-C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Bigg|\mathbf{x}_j\right) \\ &= \arg\min_{C_{\tilde{\beta}}(\mathbf{x}_j)} E_s\left(r_i\,\frac{[y_j-C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right). \qquad (4.2) \end{aligned}$$

(The second equality follows from (2.9)). Hence, by substituting the sample expectation outside the curved brackets by the sample mean (a straightforward application of the method of moments) and estimating r_i by $\hat{r}_i$ (see Remark 3), the vector β can be estimated as,

$$\hat{\beta}_1 = \arg\min_{\tilde{\beta}} \sum_{i\in s}\left(\hat{r}_i \frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \right). \tag{4.3}$$

The predictor of the population total takes then the form,

$$\hat{Y}_1 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_1}(\mathbf{x}_j). \tag{4.4}$$

Alternatively, it follows from (4.1) that,

$$\begin{aligned} E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \Big| \mathbf{x}_j \right) &= E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right) \\ &= E_s\left(\left[\frac{w_j - 1}{E_s(w_j) - 1} \right] \frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right) \end{aligned} \tag{4.5}$$

where the right hand side expectation is with respect to the joint distribution of $(y_i, \mathbf{x}_j)$. Thus, β can be estimated as,

$$\hat{\beta}_2 = \arg\min_{\tilde{\beta}} \sum_{i\in s} (w_i - 1) \frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \tag{4.6}$$

since $E_s(w_i) = constant$. The predictor of Y with β estimated by $\hat{\beta}_2$ is therefore,

$$\hat{Y}_2 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_2}(\mathbf{x}_j). \tag{4.7}$$

Remark 4. A notable advantage of the use of the predictor $\hat{Y}_2$ over the use of the predictor $\hat{Y}_1$ is that it does not require the identification and estimation of the expectation $w(\mathbf{x}) = E_s(w\,|\,\mathbf{x})$. On the other hand, in situations where this expectation can be estimated properly, the predictor $\hat{Y}_1$ is likely to be more accurate since the weights $r_i = (w_i - 1)/[E_s(w_i|\mathbf{x}_i) - 1]$ will often be less variable than the weights $(w_i - 1)$. This is because the weights r_i only account for the net effect of the sampling process on the target conditional distribution $f_c(y_i|\mathbf{x}_i)$, whereas the weights $(w_i - 1)$ account for the effect of the sampling process on the joint distribution $f_c(y_i, \mathbf{x}_i)$. In particular, when w_i is a deterministic function of $\mathbf{x}_i$ such that $w_i = w(\mathbf{x}_i)$, the sampling process is noninformative and $f_c(y_i|\mathbf{x}_i) = f_s(y_i|\mathbf{x}_i) = f_p(y_i|\mathbf{x}_i)$. In this case the estimator $\hat{\beta}_1$ (but not $\hat{\beta}_2$) coincides with the optimal generalized least square (GLS) estimator of β since $r_i = 1$ and the model (4.1) holds for the sample data. (For the data analysed in section 7, the empirical variance of the weights r_i is 1.36, whereas the empirical variance of the weights w_i is 2.66). In contrast to this, when the sampling weights w_i are independent of $\mathbf{x}_i$, the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and hence the predictors $\hat{Y}_1$ and $\hat{Y}_2$ are equal since $w(\mathbf{x}_i) = constant$.

An interesting special case of the predictor $\hat{Y}_2$ arises when the working model postulated for the sample-complement is linear with an intercept term and constant variance. Let $\mathbf{x}_i' = (1, \tilde{\mathbf{x}}_i')$. As easily verified, the estimator in this case takes the form,

$$\hat{Y}_{2,\ \mathrm{Reg}} = \sum_{i\in s} y_i + \hat{Y}_C + \tilde{B}_c'\left[\tilde{X}(c) - \hat{X}_C\right] \tag{4.8}$$

where $\tilde{X}(c) = \sum_{i\notin s} \tilde{\mathbf{x}}_i$, $(\hat{Y}_C, \hat{X}_C) = [(N-n)/\sum_{i\in s}(w_i - 1)]$ $[\sum_{i\in s}(w_i - 1)(y_i, \tilde{\mathbf{x}}_i)]$ and $\tilde{B}_c$ is the probability weighted estimator of the vector coefficient of $\tilde{\mathbf{x}}_i$ but with the weights $(w_i - 1)$ instead of w_i.

Remark 5. The predictor $\hat{Y}_{2,\ \mathrm{Reg}}$ can be obtained as a special case of the Cosmetic predictors proposed by Brewer (1999). It should be emphasized, however, that the development of the cosmetic predictors and the derivation of their MSE assumes explicitly *noninformative* sampling.

An important property of $\hat{Y}_{2,\ \mathrm{Reg}}$ is that under general conditions it is *design consistent* for *Y*, irrespective of the true sample-complement model (see Lemma 1 below). Many analysts view ‘design consistency’ as an essential requirement from any predictor; see the discussion in Hansen, Madow and Tepping (1983) and Särndal (1980). The following Lemma 1 defines conditions under which the more general predictor $\hat{Y}_2$ of (4.7) is design consistent for *Y*.

Lemma 1. The predictor $\hat{Y}_2$ is design consistent for *Y* if the working model used for the computation of $\hat{\beta}_2$ satisfies the conditions, *i*- $C_\beta(\mathbf{x})$ has an intercept term, *ii*- $C_\beta(\mathbf{x})$ is differentiable with respect to β in the neighborhood of $\hat{\beta}_2$ and *iii*- $v(\mathbf{x}) =$ constant.

Proof: By (4.6) and condition *iii*, $\hat{\beta}_2 = \arg\min_{\tilde{\beta}}$ $\sum_{i\in s}(w_i - 1)[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2$ and by condition *i*, $C_\beta(\mathbf{x}) = \beta_0 + C_{\beta_1 \ldots \beta_p}(\tilde{\mathbf{x}})$, so that by condition *ii*, $\partial/\partial\beta_0$ $\{\sum_{i\in s} (w_i - 1)[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2\}_{\beta=\hat{\beta}_2} = 0$, which implies $\sum_{i\in s}(w_i - 1)[y_i - C_{\hat{\beta}_2}(\mathbf{x}_i)] = 0$ or,

$$\sum_{i\in s} w_i y_i = \sum_{i\in s} y_i + \sum_{i\in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i) - \sum_{i\in s} C_{\hat{\beta}_2}(\mathbf{x}_i). \tag{4.9}$$

The proof is completed by noting that under mild regularity conditions $\sum_{i\in s} w_i y_i$ is design consistent for *Y*, and $\sum_{i\in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i)$ is design consistent for $\sum_{j=1}^{N} C_{\hat{\beta}_2}(\mathbf{x}_i)$. Thus, the right hand side of (4.9) converges in probability to $\hat{Y}_2$ while the left hand side converges in probability to *Y*.

It is important to emphasize again that the Lemma does not assume that the working model is the correct sample-complement model.

The use of the predictors $\hat{Y}_1$ and $\hat{Y}_2$ requires a specification of the sample-complement model. Next we develop another predictor that only requires the identification and estimation of the sample model. The approach leading to this predictor is a sample-complement analogue of the 'bias correction method' proposed by Chambers *et al.* (2003). The proposed predictor is based on the following relationship,

$$\begin{aligned}\sum_{j\notin s} E_c(y_j \mid \mathbf{x}_j) &= \sum_{j\notin s} E_s(y_j \mid \mathbf{x}_j) \\ &\quad + (N-n)\left\{\frac{1}{N-n}\sum_{j\notin s} E_c\left\{\left[y_j - E_s(y_j \mid \mathbf{x}_j)\right] \Big| \mathbf{x}_j\right\}\right\} \\ &\cong \sum_{j\notin s} E_s(y_j \mid \mathbf{x}_j) \\ &\quad + (N-n)\left\{\frac{1}{N-n}\sum_{j\notin s} E_c\left[y_j - E_s(y_j \mid \mathbf{x}_j)\right]\right\} \qquad (4.10)\end{aligned}$$

where in the second row we replaced the sample-complement average of the conditional expectations $E_c(y_j \mid \mathbf{x}_j)$ by its expectation over the sample-complement distribution of the **x**-values (n denotes the sample size). By (2.9),

$$\begin{aligned}E_c[y_j - E_s(y_j \mid \mathbf{x}_j)] \\ = E_s\left\{\frac{w_j - 1}{[E_s(w_j) - 1]}\left[y_j - E_s(y_j \mid \mathbf{x}_j)\right]\right\} \qquad (4.11)\end{aligned}$$

implying that the sample-complement mean in the second row of (4.10) can be estimated as $\hat{M}_c = 1/n \sum_{i\in s}\{[(w_i - 1)/(\bar{w}_s - 1)][y_i - \hat{E}_s(y_i \mid \mathbf{x}_i)]\}$, where $\bar{w}_s = \sum_{i\in s} w_i / n$. The proposed predictor therefore takes the form,

$$\hat{Y}_3 = \sum_{i\in s} y_i + \sum_{j\notin s} \hat{E}_s(y_j \mid \mathrm{x}_j) + (N-n)\hat{M}_c \qquad (4.12)$$

with $\hat{E}_s(y_j \mid \mathbf{x}_j)$ estimated from the sample data. The use of $\hat{Y}_3$ only requires the identification and estimation of the sample regression $E_s(y_j \mid \mathbf{x}_j)$, which can be carried out using conventional regression techniques. Moreover, under mild conditions $\hat{Y}_3$ is *design consistent* for Y even if the expectation $E_s(y_j \mid \mathbf{x}_j)$ is misspecified. This property follows from the fact that $\sum_{j\notin s} \hat{E}_s(y_j \mid \mathrm{x}_j)$ is design consistent for $\sum_{j\notin s} E_s(y_j \mid \mathrm{x}_j)$ and $(N-n)\hat{M}_c$ is design consistent for $M_c = \sum_{j\notin s}[y_j - E_s(y_j \mid \mathrm{x}_j)]$.

Remark 6. If the model fitted to the sample data is linear regression with an intercept and constant residual variance, the difference between the predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ defined by (4.8) and the predictor $\hat{Y}_3$ is that $\hat{Y}_{2,\,\mathrm{Reg}}$ uses a consistent estimator for the regression coefficients defining the linear approximation to the model holding for the sample-complement, whereas in $\hat{Y}_3$ the regression coefficients are estimated by ordinary least squares (OLS), thus estimating the linear approximation to the sample model.

Finally, rather than only predicting the sample-complement values as with the previous predictors, one could instead predict all the population values by their estimated expectations under the population model. Assuming that the latter model is linear regression with an intercept term and constant residual variance, application of (2.5b) yields,

$$\begin{aligned}\beta &= \arg\min_{\tilde{\beta}} E_p(y_k - \mathbf{x}_k'\tilde{\beta})^2 \\ &= \arg\min_{\tilde{\beta}} \frac{E_s[w_k(y_k - \mathbf{x}_k'\tilde{\beta})^2]}{E_s(w_k)}. \qquad (4.13)\end{aligned}$$

Estimating the sample expectation in the numerator of (4.13) by the corresponding sample mean (application of the method of moments) and minimizing the sample mean with respect to β yields the familiar probability weighted estimator $\hat{B}_{pw} = (X'_{[s]}\, W_s\, X_{[s]})^{-1}(X'_{[s]}\, W_s\, Y_s)$, where $(X_{[s]}, Y_s) = \{[\mathbf{x}_1 \ldots \mathbf{x}_n]', (y_1 \ldots y_n)'\}$ and $W_s = \mathrm{Diag}[\mathrm{w}_1 \ldots w_n]$. Let $\mathbf{x}_i' = (1,\ \tilde{\mathbf{x}}_i')$. Estimating $\hat{E}_p(y_k \mid \mathbf{x}_k) = \mathbf{x}_k'\hat{B}_{pw} = \hat{B}_0 + \tilde{\mathbf{x}}_k'\tilde{B}_{pw}$ and summing over all the population values yields the familiar generalized regression (GREG) estimator (Särndal 1980),

$$\hat{Y}_{\mathrm{GREG}} = N\frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i} + \tilde{B}'_{pw}\left[\tilde{X}(p) - N\frac{\sum_{i\in s} w_i \mathbf{x}_i}{\sum_{i\in s} w_i}\right];$$

$$\tilde{X}(p) = \sum_{k=1}^{N} \tilde{\mathbf{x}}_k. \qquad (4.14)$$

Remark 7. By considering the estimation of Y as a prediction problem, the use of the predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ in (4.8) requires the prediction of (N - n) values whereas the use of the GREG requires the prediction of N values. Hence, in situations where both the sample-complement model and the population model can be approximated fairly well by linear regression models with intercept terms (but possibly with different vectors of coefficients for the two models), one expects that for sufficiently large sampling fractions n/N the predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ will be superior (see the empirical results in section 7).

5. EXAMPLES

5.1 Prediction with No Concomitant Variables

Let $\mathbf{x}_i = 1$ for all i. By (3.2),

$$\hat{Y} = \sum_{i\in s} y_i + \sum_{j\notin s} \hat{E}_c(y_j) = \sum_{i\in s} y_i + (N-n)\hat{E}_s\left(\frac{w_j - 1}{\hat{E}_s(w_j) - 1} y_j\right). \qquad (5.1)$$

Estimating the two sample expectations in the right hand side of (5.1) by the respective sample means yields the estimator,

$$\hat{Y}_{\mathrm{EI}} = \sum_{i\in s} y_i + (N-n)\frac{1}{n}\sum_{i\in s}\frac{w_i - 1}{\overline{w}_s - 1} y_i = \sum_{i\in s} y_i + \frac{(N-n)}{\sum_{i\in s}(w_i - 1)}\sum_{i\in s}(w_i - 1)y_i\,. \qquad (5.2)$$

In (5.2), $\sum_{i\in s}(w_i - 1)y_i$ is a 'Horvitz-Thompson estimator' of $\sum_{j\notin s} y_j$. The multiplier $(N-n)/\sum_{i\in s}(w_i - 1)$ is a 'Hajek type correction' for controlling the variability of the sampling weights. Notice that $\hat{Y}_{EI}$ is a special case of the predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ defined in (4.8), obtained by setting $\mathbf{x}_i = 1$ for all *i*. It is also a special case of the predictor $\hat{Y}_3$ if one estimates $\hat{E}_s(y_j) = \bar{y} = \sum_{i\in s} y_i/n$. For sampling designs such that $\sum_{i\in s} w_i = N$ for all *s*, or if one estimates $\hat{E}_s(w_i) = N/n$, the predictor $\hat{Y}_{EI}$ reduces to the familiar Horvitz-Thompson estimator of the population total, $\hat{Y}_{\mathrm{H-T}} = \sum_{i\in s} w_i y_i$.

As with the GREG estimator considered in section 4, rather than predicting the sample-complement total $Y_c = \sum_{j\notin s} y_j$ and using the predictor $\hat{Y}_{EI}$, one could predict all the population *y*-values by estimating their expectations under the population model. By (2.5b), $E_p(y_i) = E_s(w_i y_i)/E_s(w_i)$. Estimating the two sample expectations by the corresponding sample means yields the familiar Hajek estimator,

$$\hat{Y}_{\mathrm{Hajek}} = \sum_{k=1}^{N} \hat{E}_p(y_k) = N\hat{E}_s\left(\frac{w_i y_i}{\hat{E}_s(w_i)}\right) = \frac{N}{\sum_{i\in s} w_i}\sum_{i\in s} w_i y_i\,. \qquad (5.3)$$

Here again, we anticipate $\hat{Y}_{EI}$ to be more precise than $\hat{Y}_{\mathrm{Hajek}}$ as the sampling fraction increases (see also the empirical results in section 7). Note that $\hat{Y}_{EI}$ and $\hat{Y}_{\mathrm{Hajek}}$ are the same and coincide with the Horvitz-Thompson estimator for sampling designs satisfying $\sum_{i\in s} w_i = N$.

5.2 Optimal Prediction with Concomitant Variables, Comparison with Optimal Predictors Under Noninformative Sampling

Let the population model be,

$$y_i = H_\beta(\mathbf{x}_i) + \varepsilon_i,\ E_p(\varepsilon_i | \mathbf{x}_i) = 0,$$
$$E_p(\varepsilon_i^2 | \mathbf{x}_i) = v(\mathbf{x}_i),\ E_p(\varepsilon_i\varepsilon_j | \mathbf{x}_i, \mathbf{x}_j) = 0,\ i \neq j \qquad (5.4)$$

and suppose that the sample inclusion probabilities can be modeled as,

$$\pi_i = K \times [y_i\ g(\mathbf{x}_i) + \delta_i],\ E_p(\delta_i | \mathbf{x}_i, y_i) = 0 \qquad (5.5)$$

where $H_\beta(\mathbf{x})$, $v(\mathbf{x})$ and $g(\mathbf{x})$ are positive functions and *K* is a normalizing constant. (Below we consider the special case of 'regression through the origin'). This sampling scheme is considered for illustration only, although in section 2 we mention several practical situations where the sample selection probabilities depend directly on the *y* and **x**-values. In particular, this is the case with the data set analysed in section 7. Under (5.4) and (5.5), $\pi(\mathbf{x}_i) = E_p(\pi_i | \mathbf{x}_i) = KH_\beta(\mathbf{x}_i)g(\mathbf{x}_i)$. Hence, by (2.9), (5.4) and (5.5),

$$E_c(y_j | \mathbf{x}_j) = E_p\left(\frac{1-\pi_j}{1-\pi(\mathrm{x}_j)} y_j \Big| \mathbf{x}_j\right) = E_p\left(\frac{1-\pi(\mathbf{x}_j) - K\varepsilon_j g(\mathbf{x}_j) - K\delta_j}{1-\pi(\mathbf{x}_j)} y_j \Big| \mathbf{x}_j\right) = E_p(y_j | \mathbf{x}_j) - \frac{Kg(\mathbf{x}_j)v(\mathbf{x}_j)}{1-\pi(\mathbf{x}_j)}. \qquad (5.6)$$

The last expression in (5.6) shows that $E_c(y_j | \mathbf{x}_j) < E_p(y_j | \mathbf{x}_j) = H_\beta(\mathbf{x}_j)$, which is clear since for the inclusion probabilities defined by (5.5), the sample-complement tends to include the units with the smaller *y*-values for any given **x**-values. Note, however, that as $n/N \to 0$, $K \to 0$ and $E_p(y_j | \mathbf{x}_j) - E_c(y_j | \mathbf{x}_j) \to 0$ (see Remark 2).

As a special case of (5.4), consider the case of a single auxiliary variable *x* and let $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$ ('regression through the origin with variance proportional to *x*'). For *noninformative* sampling and known β, the optimal unbiased predictor of *Y* minimizing $E_p[(\hat{Y} - Y)^2 | D_s]$ is in this case, $\hat{Y} = \sum_{i\in s} y_i + \beta\sum_{j\notin s} x_j$. In the practical case of unknown β, the optimal unbiased predictor of *Y* is the familiar Ratio estimator $\hat{Y}_R = N\bar{y}(\overline{X}/\bar{x})$ with $\bar{y}$ denoting the sample mean of *Y* and ($\bar{x}$, $\overline{X}$) denoting the sample and population means of *x* (Brewer 1963, Royall 1970).

Now let $g(x) = 1$ in (5.5) for all *x*, so that $\pi_i = n(y_i + \delta_i)/\sum_{j=1}^{N}(y_j + \delta_j)$. For sufficiently large *N*, we can approximate $\pi_i \approx n(y_i + \underline{\delta_i})/(N\beta\overline{X})$, implying that $\pi(x_i) = E_p(\pi_i | x_i) \approx nx_i/(N\overline{X})$. By (5.6), $E_c(y_j | x_j) = x_j\beta - \sigma^2 x_j/[\beta(f^{-1}\overline{X} - x_j)]$ where $f = n/N$ is the sampling fraction, so that for known β and σ^2 the optimal predictor of *Y* is,

$$\hat{Y}_{E,\,\mathrm{Reg}} = \sum_{i\in s} y_i + \beta\sum_{j\notin s} x_j - \frac{\sigma^2}{\beta}\sum_{j\notin s}\frac{x_j}{f^{-1}\overline{X} - x_j}. \qquad (5.7)$$

Lemma 2: Let the population model be defined by (5.4) with $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$. Assume also

$E_p(\varepsilon_i^3 \mid x_i) = 0.$ Suppose that the sample units are selected independently with probabilities defined as in (5.5), with $g(x) = 1.$ Then,

$$\mathrm{MSE}_p\left(\hat{Y}_{E,\ \mathrm{Reg}} \middle| D_s\right) = \sigma^2 \sum_{j \notin s} x_j - (\sigma^2/\beta)^2 \sum_{j \notin s} [x_j^2 / (f^{-1}\bar{X} - x_j)^2]. \quad (5.8)$$

Proof: By the independence of the population values and of the sample selections,

$$\begin{aligned} &\mathrm{MSE}_p(\hat{Y}_{E,\,\mathrm{Reg}} | D_s) \\ &= E_p[(\hat{Y}_{E,\,\mathrm{Reg}} - Y)^2 | D_s] \\ &= \sum_{j \notin s} E_c\{[y_j - E_c(y_j | x_j)]^2 \mid x_j\}. \end{aligned}$$

By (5.6), $[y_j - E_c(y_j | x_j)]^2 = \{\varepsilon_j + x_j^* / [1 - \pi(x_j)]\}^2$ where $x_j^* = K\sigma^2 x_j$, $K = n / \beta N\bar{X}$ and $\pi(x_j) = E_p(\pi_j \mid x_j) \approx nx_j / (N\bar{X})$. Hence,

$$\begin{aligned} &E_c\{[y_j - E_c(y_j | x_j)]^2 | x_j\} \\ &= E_c(\varepsilon_j^2 | x_j) + 2x_j^* / (1 - \pi(x_j)) E_c(\varepsilon_j | x_j) \\ &\quad + [x_j^* / (1 - \pi(x_j))]^2. \end{aligned}$$

Now,

$$\begin{aligned} &E_c(\varepsilon_j^2 | x_j) \\ &= E_p[1 - \pi_j / (1 - \pi(x_j))\varepsilon_j^2 | x_j] \\ &= E_p[1 - \pi(x_j) - \mathrm{K}\varepsilon_\mathrm{j} - K\delta_j / (1 - \pi(x_j))\ \varepsilon_j^2 | x_j] \\ &= E_p(\varepsilon_j^2 | x_j) = \sigma^2 x_j \end{aligned}$$

and

$$\begin{aligned} &E_c(\varepsilon_j | x_j) \\ &= E_p[1 - \pi(x_j) - K\varepsilon_j - K\delta_j / \left(1 - \pi(x_j)\right)\ \varepsilon_j | x_j] \\ &= -x_j^* / (1 - \pi(x_j)). \end{aligned}$$

It follows therefore that $\mathrm{MSE}_p(\hat{Y}_{E,\,\mathrm{Reg}} \mid D_s) = \sigma^2 \sum_{j \notin s} x_j - \sum_{j \notin s} [x_j^* / (1 - \pi(x_j))]^2.$ Q.E.D.

Remark 8: For *noninformative* sampling and with known $\beta,$ the prediction MSE of the optimal predictor $\hat{Y} = \sum_{i \in s} y_i + \beta \sum_{j \notin s} x_j$ is, $E_p[(\hat{Y} - Y)^2 \mid D_s] = \sigma^2 \sum_{j \notin s} x_j$. This MSE is larger than the MSE obtained under the informative sampling scheme defined by the Lemma, which is obvious since the latter scheme tends to sample the units with the larger *y*-values and hence also with the larger *x*-values and the larger standard deviations.

6. MEAN SQUARE ERROR ESTIMATION

Estimating $\mathrm{MSE}(\hat{Y} \mid D_s) = E_p[(\hat{Y} - Y)^2 \mid D_s]$ for the predictors $\hat{Y}$ considered in section 4 requires strict model assumptions that could be hard to validate. This is largely due to the conditioning on the design information D_s. In order to deal with this problem, we propose to estimate instead the unconditional MSE, $\mathrm{MSE}(\hat{Y}) = E[(\hat{Y} - Y)^2] = E_{D_s}\{E_p[(\hat{Y} - Y)^2 \mid D_s]\}$, where $E_{D_s} = E_D E_s$ defines the expectation over the sample distribution (given the selected sample) and over all possible sample selections. Notice that $E_p[(\hat{Y} - Y)^2 \mid D_s]$ can be viewed as a random variable $u(D_s)$, so that $\mathrm{MSE}(\hat{Y}) = E_{D_s}[u(D_s)]$ defines its 'best predictor' with respect to the mean square loss function under the distribution f_{D_s} over which the expectation E_{D_s} is taken. By changing the order of the expectations, the unconditional MSE can be expressed as,

$$\begin{aligned} \mathrm{MSE}(\hat{Y}) &= E_s E_p E_D[(\hat{Y} - Y)^2 | \boldsymbol{y}] \\ &= E_p E_D[(\hat{Y} - Y)^2 | \boldsymbol{y}] \qquad (6.1) \end{aligned}$$

where $\boldsymbol{y} = \{y_i;\ i \in U\}$. Estimating the unconditional MSE of any of the predictors $\hat{Y}$ can be carried out therefore by estimating its randomization MSE, see Pfeffermann (1993) for further discussion. Estimation of the randomization MSE of the various predictors has the additional advantage of allowing their use under the design based approach.

Estimation of randomization variances of design based estimators is considered extensively in the literature and many diverse methods are in routine use. However, in view of the complicated structure of some of the predictors considered in this study and in order not to restrict to particular sampling schemes, we propose below the use of a two-step procedure that combines an inverse sampling process (Step 1) and what can be viewed as a bootstrap resampling algorithm (Step 2). A notable advantage of this procedure is that it is general and applies 'equally' to all the predictors. Also, unlike other variance estimation methods in common use, it does not require knowledge of the pair wise joint selection probabilities $\pi_{ij} = \Pr(i,\ j \in s).$ As discussed later, a valid application of the first step requires sufficiently large samples. The two steps of the proposed procedure are as follows:

Step 1- Generate a single 'pseudo population' by selecting *with replacement N* units from the original sample with probabilities proportional to $w_i = 1/\pi_i$, where N is the population size. The justification for this step is given below, see also Remark 10. Denote by Y_{pp} the sum of the *y*-values in the pseudo population.

Step 2- Select independently a large number *B* of bootstrap samples from the pseudo population generated in Step 1, using the same sampling scheme as used for the selection of the original sample, and re-estimate the population total.

Let $\hat{Y}$ represent any of the predictors and denote the predictor obtained for bootstrap sample *b* by $\hat{Y}_{pp}^b$. Estimate,

$$\hat{E}_D(\hat{Y} - Y)^2 = \frac{1}{B} \sum_{b=1}^{B} (\hat{Y}_{pp}^b - Y_{pp})^2. \qquad (6.2)$$

The performance of the estimator (6.2) in estimating the randomization MSE depends obviously on the 'closeness' of the pseudo population generated in Step 1 to the actual population from which the original sample was drawn. The closeness of the two populations can be verified in part by noting that the marginal distribution of $y_i \mid \mathbf{x}_i$ in the pseudo population is the same as in the original population. To see this, note that the pseudo population generated in Step 1 is a 'sample with replacement' from the original sample with selection probabilities Cw_i on each draw, where $C = 1/\sum_{i=1}^{n} w_i$. Denoting by $f_{pp}(y_i \mid \mathbf{x}_i)$ the marginal pseudo population distribution we find using (2.2) and (2.5a),

$$f_{pp}(y_i|\mathbf{x}_i) = \frac{E_s(Cw_i|y_i,\mathbf{x}_i)\, f_s(y_i|\mathbf{x}_i)}{E_s(Cw_i|\mathbf{x}_i)}$$

$$= \frac{E_p(\pi_i|\mathbf{x}_i)\, f_s(y_i|\mathbf{x}_i)}{E_p(\pi_i|y_i,\mathbf{x}_i)} = f_p(y_i|\mathbf{x}_i). \tag{6.3}$$

Remark 9. Equation (6.3) only refers to the marginal distribution of $y_i \mid \mathbf{x}_i$. Like with the standard bootstrap method, a successful application of the proposed procedure requires that the original sample size is sufficiently large and that the sample measurements are approximately independent. Pfeffermann *et al.* (1998) establish conditions under which for independent population measurements the sample measurement are 'asymptotically independent' under commonly used sampling schemes with unequal selection probabilities.

Remark 10. Step 1 is similar and asymptotically equivalent to duplicating sample unit i w_i times. Notice, however, that the use of this duplication procedure does not yield pseudo populations of size N unless $\sum_{i=1}^{n} w_i = N$. It is also not clear how to establish the relationship (6.3) when using this procedure.

7. EMPIRICAL ILLUSTRATIONS

7.1 Description of Empirical Study

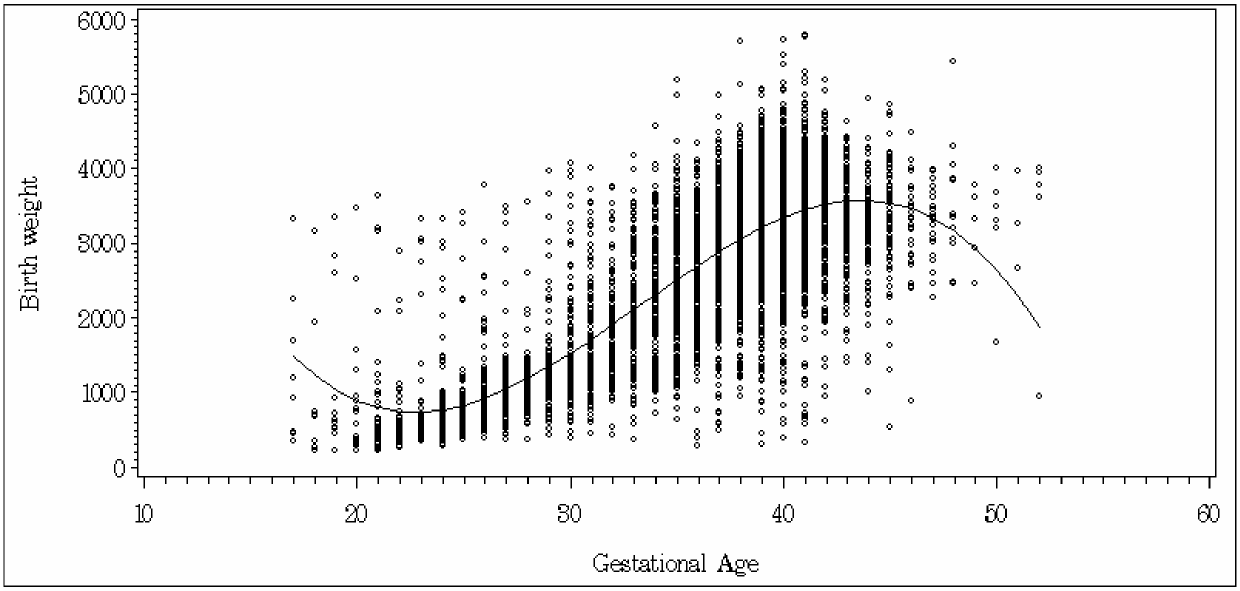
In order to illustrate the performance of the predictors and the associated MSE estimates discussed in previous sections we use a real data set, collected as part of the 1988 U.S. National Maternal and Infant Health Survey. The survey uses a disproportionate stratified random sample of vital records with the strata defined by *mother's race* and *child's birth weight;* see Korn and Graubard (1995) for details. For the empirical study in this section we considered the sample data as 'population' and selected independently 1,000 samples with probabilities proportional to the inverse of the original sampling weights, using a systematic PPS sampling scheme. The list of 'population units' was randomly ordered before every sample selection. For each sample we predicted the population total of *birth weight* (measured in *grams*, divided by 10,000 in the present study), using *gestational age* as the auxiliary variable (measured in *weeks*). The sample inclusion probabilities depend therefore on the values of the study variable that defines the original strata. Notice that although the original sample was supposedly a stratified random sample, the sampling weights actually vary within the strata, which is why we used systematic PPS sampling for the simulation study. We considered three different sample sizes, $n = 232$, 1,145, 2,429. The 'population' (original sample) size is $N = 9{,}948$. (For $n = 232$, $0.002 < \pi_i = \Pr(i \in s) < 0.15$. For $n = 1{,}145$, $0.01 < \pi_i < 0.73$. For $n = 2{,}429$, $0.03 < \pi_i < 0.99$ with mean $\bar{\pi} = 0.26$ and standard deviation $Std(\pi_i) = 0.29$. In the latter case some of the units were drawn almost with certainty).

Some of the predictors considered for this study (see below) require the specification of either the sample model or the sample-complement model. We assumed for both models the third order polynomial regression,

$$y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3 + \varepsilon_k \tag{7.1}$$

with independent residuals and constant variance. This model was found by Pfeffermann and Sverchkov (1999) to give a good fit to the 'population' (original sample) data with $R^2 = 0.61$ (see Figure 1), and it was found also to fit fairly well the sample data (with different coefficients) for several samples selected from this 'population'. Notice, on the other hand, that with this strongly informative sampling scheme, it is unlikely that the sample model, the population model and the sample-complement model are all from the same family even if with different parameters. The present study enables therefore studying the performance of the various predictors when some or all of the three models are misspecified. This important robustness question is further examined by fitting simple regression models instead of the third order polynomial regressions that is, by omitting the second and third powers of the auxiliary variable. The only exception is the model dependent predictor $\hat{Y}_1$ (Equation 4.4) where no coherent estimator for the expectation $E_s(w_j \mid x_j)$ could be found when restricting to simple regression. (The method considered in Pfeffermann and Sverchkov (1999) for the estimation of this expectation assumes normality of the population model residuals. This is a valid assumption when fitting the third order polynomial regression model but is clearly violated when dropping the second and third powers of the auxiliary variable).

U.S. National Maternal and Infant Health Survey,1988.



Model Fitted: $y_i = 17886 - 1827.7 x_i + 61.2 x_i^2 - 0.61 x_i^3 + \varepsilon_i$

$\text{Var}(\varepsilon_i) = 603.2,\ R^2 = 0.61$

Figure 1. Scatterplot of Birth Weight against Gestational Age in ‘Population’ (original Sample), and Predicted Values Under 3rd Order Polynomial Regression.

The predictors considered for this study divide therefore into three groups. The first group consists of predictors that only use the sample *y*-values and the sampling weights. Included in this group are the Horvitz-Thompson estimator $\hat{Y}_{\text{H-T}} = \sum_{i \in s} w_i y_i$, the predictor $\hat{Y}_{EI}$ defined by (5.2) and Hajek’s estimator $\hat{Y}_{\text{Hajek}}$ defined by (5.3). The second group consists of predictors that use the working model defined by (7.1). Included in this group are the two regression predictors $\hat{Y}_1$ and $\hat{Y}_{2,\text{ Reg}}$ defined by (4.4) and (4.8) respectively, the bias corrected predictor $\hat{Y}_3$ defined by (4.12) and the GREG estimator defined by (4.14). The third group contains the same predictors as the second group (except for $\hat{Y}_1$, see above), but based on the simple regression model (only the first power of *x*).

The MSEs of all the predictors considered in this study have been estimated by use of the two-step procedure described in section 6. However, because of computing time limitations, the MSE estimators were only computed for a random selection of 200 out of the 1,000 samples and are based on only 200 bootstrap samples from each pseudo population. For assessing the performance of the MSE estimators we computed the corresponding empirical MSEs based on the 1,000 samples selected from the study population. Thus, the ‘true’ MSE of a generic predictor $\hat{Y}$ was computed as,

$$\text{MSE}(\hat{Y}) = \frac{1}{1{,}000} \sum_{r=1}^{1{,}000} (\hat{Y}_{(r)} - Y)^2 \qquad (7.2)$$

where $\hat{Y}_{(r)}$ denotes the predictor computed from the r^{th} sample. Notice that since the population values are fixed, the MSE in (7.2) is the randomization MSE over all possible sample selections, which is what the estimator (6.2) is intended to estimate.

7.2 Results of Empirical Study

The main results of this study are exhibited in Tables 1.1 – 1.3 (one table for each sample size). The third column of each table shows for every predictor $\hat{Y}$ the empirical bias, $[(\sum_{r=1}^{R} \hat{Y}_{(r)} / R) - Y]$, and the standard deviation (S*td*) of the empirical bias, computed as $[\sum_{r=1}^{R} (\hat{Y}_{(r)} - \bar{Y}_R)^2 / R^2]^{1/2}$; $\bar{Y}_R = \sum_{r=1}^{R} \hat{Y}_{(r)} / R,\ R = 1{,}000.$ The next two columns show respectively the ‘true’ (empirical) RMSE (square root of Equation 7.2), and the square root of the mean of the corresponding Bootstrap estimators defined by (6.2).

The main conclusions from Tables 1.1 – 1.3 are as follows:

1- All the predictors considered for this study are virtually design unbiased with all three sample sizes, irrespective of the underlying working model. The predictor $\hat{Y}_1$ has a statistically significant bias when tested by use of the conventional *t*-statistic but the actual bias is negligible when compared to the true population total. (The predictor $\hat{Y}_1$ is the only predictor considered in this study that is not design consistent).

The next three comments refer to the RMSE of the various predictors.

2- The predictors in Groups 2 and 3 that use the auxiliary values perform much better than the predictors in Group 1, particularly for the smaller sample sizes. The predictors in Group 2 that employ the 3rd order polynomial regression model (7.1) perform better than the corresponding predictors in Group 3 that employ the simple regression model as the working model, but the differences diminish as the sample size increases.

3- An important result emerging from this study is that the predictors $\hat{Y}_{2,\,\mathrm{Reg}}$ and $\hat{Y}_{EI}$ (and also $\hat{Y}_3$ for the larger sample sizes), that only predict the *y*-values for units outside the sample indeed perform better than the other predictors in their respective groups (see also below). As surmised in Remark 7, this holds particularly with the larger sample sizes. Notice that the differences between $\hat{Y}_{2,\,\mathrm{Reg}}$ and the GREG estimator for $n = 1{,}145$ and $n = 2{,}250$ are smaller under the polynomial model (Group 2) than under the simple regression model (Group 3), which is explained by the tight relationship between the study variable and auxiliary variables under the polynomial model. The predictor $\hat{Y}_3$ is less stable than $\hat{Y}_{2,\,\mathrm{Reg}}$ for $n = 232$ but for the other two sample sizes the two predictors perform similarly.

4- The predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ performs somewhat better than the model dependent predictor $\hat{Y}_1$ that employs the expectations $E(w_i \mid x_i)$ to adjust the sampling weights. We have no clear explanation for this result because as illustrated in Pfeffermann and Sverchkov (1999) using the same data, adjusting the sampling weights improves the estimation of the regression coefficients very significantly.

Next consider the MSE estimators.

5- The MSE estimators developed in section 6 perform very well for all the predictors and with all the sample sizes. For the sample size $n = 232$ there is a systematic under-estimation of the RMSE by up to 3%, which is explained by the fact that the pseudo population is in this case less variable than the actual study population (see Remark 9). The MSE estimators are almost unbiased for the other sample sizes with the largest difference between the estimated and true RMSE being again in the magnitude of 3%.

Another way of assessing the bias of the various predictors and their MSE estimation is by studying the coverage properties of confidence intervals defined by these predictors. Tables 2.1 – 2.3 compare the empirical percentage coverage of the standard confidence intervals $\hat{Y} \pm \; Z_{1-\alpha/2}\sqrt{\hat{\mathrm{MSE}}}$ with the corresponding nominal percentages for selected values of α (one table for each sample size). The empirical percentages are somewhat erratic with $n = 232$ sample units but they stabilize as the sample size increases, particularly with the use of the predictors in the second and third group. The empirical percentages are close to the nominal percentages with all the predictors when $n = 2{,}250$.

Table 1.1
Bias, RMSE and Square Root of Mean of MSE Estimators, n = 232

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\mathrm{MSE}}}$
1 No *x*-values	$\hat{Y}_{\mathrm{H-T}}$	-4.5 (11.6)	365.1	355.0
	$\hat{Y}_{EI}$	1.5 (2.9)	91.1	89.8
	$\hat{Y}_{\mathrm{Hajek}}$	1.7 (2.9)	93.0	91.6
2 3rd order polynomial regression	$\hat{Y}_1$	4.4 (2.0)	64.0	63.0
	$\hat{Y}_{2,\,\mathrm{Reg}}$	3.5 (2.0)	63.4	62.4
	$\hat{Y}_3$	-0.3 (2.1)	65.4	65.0
	$\hat{Y}_{\mathrm{GREG}}$	3.4 (2.1)	63.6	62.6
3 Simple Regression	$\hat{Y}_{2,\,\mathrm{Reg}}$	-2.3 (2.2)	68.0	66.2
	$\hat{Y}_3$	-0.3 (2.2)	68.6	67.4
	$\hat{Y}_{\mathrm{GREG}}$	-2.3 (2.2)	68.3	66.5

True ‘population’ total= 2710.7

Table 1.2
Bias, RMSE and Square Root of Mean of MSE Estimators, n = 1,145

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\text{MSE}}}$
1 No *x*-values	$\hat{Y}_{\text{H-T}}$	-9.1 (5.0)	157.1	156.1
	$\hat{Y}_{EI}$	0.0 (1.1)	35.2	34.9
	$\hat{Y}_{\text{Hajek}}$	-0.1 (1.3)	39.5	39.3
2 3rd order polynomial regression	$\hat{Y}_1$	3.0 (0.9)	27.6	28.1
	$\hat{Y}_{2,\text{ Reg}}$	2.0 (0.9)	27.4	27.3
	$\hat{Y}_3$	0.5 (0.9)	27.4	27.7
	$\hat{Y}_{\text{GREG}}$	1.7 (0.9)	27.8	27.8
3 Simple Regression	$\hat{Y}_{2,\text{ Reg}}$	0.0 (1.0)	28.3	28.7
	$\hat{Y}_3$	0.1 (1.0)	28.2	28.9
	$\hat{Y}_{\text{GREG}}$	0.0 (2.0)	29.1	29.6

True ‘population’ total= 2710.7

Table 1.3
Bias, RMSE and Square Root of Mean of MSE Estimators, n=2,250

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\text{MSE}}}$
1 No *x*-values	$\hat{Y}_{\text{H-T}}$	1.3 (2.7)	82.7	80.4
	$\hat{Y}_{EI}$	-0.2 (0.6)	18.5	18.8
	$\hat{Y}_{\text{Hajek}}$	0.1 (0.7)	23.5	23.8
2 3rd order polynomial regression	$\hat{Y}_1$	1.3 (0.5)	17.5	17.3
	$\hat{Y}_{2,\text{ Reg}}$	0.6 (0.5)	16.9	16.3
	$\hat{Y}_3$	-0.3 (0.5)	17.1	16.5
	$\hat{Y}_{\text{GREG}}$	0.5 (0.5)	17.9	18.3
3 Simple Regression	$\hat{Y}_{2,\text{ Reg}}$	-0.3 (0.5)	17.3	16.8
	$\hat{Y}_3$	-0.3 (0.5)	17.7	17.3
	$\hat{Y}_{\text{GREG}}$	-0.2 (0.6)	18.8	18.3

True ‘population’ total= 2710.7

Table 2.1
Nominal and Empirical Percentage Coverage of Confidence Intervals, n = 232

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No *x*-values	$\hat{Y}_{\text{H-T}}$	2.5	3.5	5.5	10.0	90.0	97.0	99.0	99.5
	$\hat{Y}_{EI}$	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
	$\hat{Y}_{\text{Hajek}}$	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
2 3rd order polynomial regression	$\hat{Y}_1$	0.0	0.0	1.5	6.5	86.0	90.5	92.5	97.5
	$\hat{Y}_{2,\text{ Reg}}$	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
	$\hat{Y}_3$	0.0	0.5	2.5	6.5	87.5	91.0	95.0	98.5
	$\hat{Y}_{\text{GREG}}$	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
3 Simple Regression	$\hat{Y}_{2,\text{ Reg}}$	0.0	1.0	2.5	7.0	87.0	91.5	97.5	98.0
	$\hat{Y}_3$	0.0	1.0	2.5	7.0	86.0	91.5	96.5	98.0
	$\hat{Y}_{\text{GREG}}$	0.0	1.0	2.5	7.0	86.5	91.5	97.0	98.0

Table 2.2
Nominal and Empirical Percentage Coverage of Confidence Intervals, n = 1,145

Group	Predictor	1	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x-values	$\hat{Y}_{\mathrm{H-T}}$	4.0	7.0	9.0	13.5	95.5	98.0	98.5	99.5
	$\hat{Y}_{EI}$	3.0	5.0	8.0	12.5	92.5	95.5	99.5	100.0
	$\hat{Y}_{\mathrm{Hajek}}$	3.5	5.0	9.5	12.5	92.5	96.0	99.5	100.0
2 3rd order polynomial regression	$\hat{Y}_1$	0.5	2.0	5.0	7.5	86.5	93.5	96.0	97.0
	$\hat{Y}_{2,\mathrm{\ Reg}}$	0.5	3.0	6.0	9.0	86.5	94.5	96.5	97.0
	$\hat{Y}_3$	0.5	2.0	6.0	9.5	88.0	94.0	97.0	98.0
	$\hat{Y}_{\mathrm{GREG}}$	0.5	3.0	5.0	9.0	86.5	94.0	96.5	98.0
3 Simple Regression	$\hat{Y}_{2,\mathrm{\ Reg}}$	0.5	3.0	6.0	11.0	90.0	93.0	97.0	99.5
	$\hat{Y}_3$	0.5	2.5	5.5	10.5	90.0	94.0	97.0	99.5
	$\hat{Y}_{\mathrm{GREG}}$	1.0	3.0	6.0	11.0	90.5	94.0	97.5	99.0

Table 2.3
Nominal and Empirical Percentage Coverage of Confidence Intervals, n = 2,250

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x-values	$\hat{Y}_{\mathrm{H-T}}$	0.5	1.0	5.5	11.0	95.0	97.5	99.0	99.5
	$\hat{Y}_{EI}$	1.0	3.0	5.5	9.0	91.5	96.0	99.0	99.5
	$\hat{Y}_{\mathrm{Hajek}}$	1.0	2.5	5.5	9.0	93.0	97.0	98.5	99.5
2 3rd order polynomial regression	$\hat{Y}_1$	0.5	2.0	5.0	9.0	91.0	94.5	96.5	97.5
	$\hat{Y}_{2,\mathrm{\ Reg}}$	0.5	2.5	6.5	10.5	90.5	94.5	96.5	98.0
	$\hat{Y}_3$	0.5	2.0	7.5	12.5	91.5	95.5	96.5	97.5
	$\hat{Y}_{\mathrm{GREG}}$	0.5	2.0	6.0	11.0	91.0	94.5	96.0	98.0
3 Simple Regression	$\hat{Y}_{2,\mathrm{\ Reg}}$	1.0	3.0	6.0	11.0	91.0	95.0	97.5	99.0
	$\hat{Y}_3$	1.0	2.0	6.0	12.0	90.0	95.0	97.5	98.0
	$\hat{Y}_{\mathrm{GREG}}$	0.0	1.5	5.0	11.5	91.5	95.0	97.5	99.0

As implied by the theoretical developments of this article and illustrated in the empirical study, predicting only the y-values for units outside the sample employing the sample-complement model yields better predictors for the population total than predicting all the population values by use of the population model, as implicitly implemented when using the GREG or Hajek's estimators. Clearly, the differences are only appreciable when the sampling fractions are not negligible.

In order to highlight this point further, we present in Table 3 the mean prediction error (mpe) in the original scale (grams) over the 1,000 samples when predicting the sample-complement values;

$$\mathrm{mpe} = \sum\nolimits_{r=1}^{1,000} \left[\sum\nolimits_{j \notin S_k} (\hat{y}_j - y_j) / (N - n) \right] \Big/ 1,000$$

where S_r defines the r^{th} selected sample. The mpe's are shown for three predictors, all utilizing the working model (7.1) and thus having the general form, $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j + \hat{\beta}_2 x_j^2 + \hat{\beta}_3 x_j^3,\ j \notin s.$ For the first predictor the vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ is estimated by OLS, which corresponds to the use of the sample model; for the second predictor β is estimated by the probability weighted estimator $\hat{B}_{pw}$, that corresponds to the use of the population model whereas for the third predictor β is estimated by the estimator $\hat{B}_c$ which is computed similarly to $\hat{B}_{pw}$ but with weights ($w_i - 1$), that corresponds to the use of the sample-complement model.

Table 3
Mean Prediction Errors and Std of Means (in brackets) Under Three Prediction Models

Sample size	Sample Model	Population model	Sample-Complement model
232	329.0 (2.2)	10.3 (2.3)	4.3 (2.3)
1,145	375.0 (0.9)	37.7 (1.1)	2.4 (1.1)
2,250	387.5 (0.6)	85.8 (0.7)	0.9 (0.8)

The clear conclusion emerging from Table 3 is that the use of either the population model or the model holding for units in the sample for the prediction of *y*-values of units outside the sample can result in appreciable biases. Notice that the bias induced by use of the population model increases as the sampling fraction increases, which agrees with the previous discussion asserting that the difference between the sample and sample-complement models only shows up with relatively large sample sizes (see Comment 2).

8. CONCLUDING REMARKS

In this article we use the sample and sample-complement distributions for developing *design consistent* predictors of finite population totals. Known predictors in common use are shown to be special cases of the present theory. The MSEs of the new predictors are estimated by a combination of an inverse sampling algorithm and a resampling method. As supported by theory and illustrated in the empirical study, predictors of finite population totals that only require the prediction of the outcome values for units outside the sample perform better than predictors in common use even under a design based framework, unless the sampling fractions are very small. The MSE estimators are shown to perform well both in terms of bias and when used for the computation of confidence intervals for the population totals. Further experimentation with this kind of predictors and MSE estimation is therefore highly recommended.

ACKNOWLEDGEMENT

The authors would like to thank the associate editor and two referees for very constructive comments.

REFERENCES

BREWER, K.R.W. (1963). Ratio estimationand finite populations: some results deducible from the assumptions of an underlying stochastic process. *Australian Journal of Statistics.* 5, 93-105.

BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*. 25, 205-212.

CHAMBERS, R.L., DORFMAN, A. and SVERCHKOV, M. (2003). Nonparametric regression with complex survey data. In, *Analysis of Survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 151-174.

FULLER, W. (2003). Statistical analysis from complex survey data. Tutorial presented at the International Statistical Institute meeting, Berlin, Germany. Slides of the Tutorial appear in http://cssm.iastate.edu/academic/ staff/fuller.html.

HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association.*. 78, 776-807.

KIM, D.H. (2002). Bayesian and empirical Bayesian analysis under informative sampling. *Sankhyā B*. 64, 267-288.

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.

PATAK, Z., HIDIROGLOU, M. and LAVALLÉE, P. (2000). The methodology of the Workplace and Employee Survey. *Proceedings of the Second International Conference on Establishment Surveys*, June 17-21, 2000, Buffalo, New York, American Statistical Association. 223-232.

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

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

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

PFEFFERMANN, D., and SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā,* Series B. 61. 166-186.

PFEFFERMANN, D., and SVERCHKOV, M. (2003a). Fitting generalized linear models under informative probability sampling. In *Analysis of survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 175-195.

PFEFFERMANN, D., and SVERCHKOV, M. (2003b). Small area estimation under informative sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association (to appear).

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*. 57, 377-387.

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