PREDICTION OF FINITE POPULATION TOTALS UNDER INFORMATIVE SAMPLING UTILIZING THE SAMPLE DISTRIBUTION

Mikhail Sverchkov and Danny Pfeffermann, Hebrew University
Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem Israel, 91905

Key Words: Cosmetic estimation, Design consistency, Sample-complement distribution

Summary. The sample distribution is the distribution of the response variable for units included in the sample. This distribution is different from the population distribution if the sample selection probabilities depend on the values of the response variable even after conditioning on the model concomitant variables. 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 response variable values (the y-values) are random realizations from some distribution that conditions on known values of concomitant variables (the x-values). The problem considered is the prediction of the population total Y of the response variable values based on the sample y-values, the sampling weights for units in the sample and the population values of the x-values. The use of the sample distribution permits the conditioning on all these values and the prediction of Y is equivalent therefore to the prediction of the y-values for units outside the sample.

The prediction of the non-sampled y-values is achieved by approximating 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. Several predictors obtained by application of this approach are considered and compared to other known methods.

1. RELATIONSHIPS BETWEEN THE POPULATION DISTRIBUTION, THE SAMPLE DISTRIBUTION AND THE SAMPLE-COMPLEMENT DISTRIBUTION.

1.1 The sample distribution

Suppose that the population values $[\mathbf{y}, X] = \{(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)\}$ are random realizations with conditional probability density function (pdf) $f_p(y_i|\mathbf{x}_i)$, which may be either discrete or continuous. We consider single stage sampling with inclusion probabilities $\pi_i = \Pr(i \in s) = g(\mathbf{y}, X, Z)$ for some function g, where Z denotes the population values of design variables (considered as random) used for the sampling process. Let $I_i = 1$ if $i \in s$ and $I_i = 0$, otherwise. The conditional marginal *sample pdf* is defined as,

$$f_s(y_i|\mathbf{x}_i) \overset{def}{=} f(y_i|\mathbf{x}_i, I_i = 1) = \frac{\Pr(I_i = 1|y_i, \mathbf{x}_i) f_p(y_i|\mathbf{x}_i)}{\Pr(I_i = 1|\mathbf{x}_i)} \tag{1.1}$$

with the second equality obtained by application of Bayes theorem. Note that $\Pr(I_i = 1|y_i, \mathbf{x}_i)$ is generally not the same as the sample inclusion probability π_i. It follows from (1.1) that the population and sample pdfs 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 that conditions on the $\mathbf{x}$'s.

In what follows we regard the probabilities π_i as random realizations of the random variable $g(Y, \mathbf{x}, Z)$. Let $w_i = 1/\pi_i$ define the sampling weight of unit i. The following relationships hold for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$, where E_p and E_s denote expectations under the population and sample pdfs respectively. (As a special case, $\mathbf{u}_i = y_i$, $\mathbf{v}_i = \mathbf{x}_i$).

$$f_s(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p(\pi_i|\mathbf{v}_i)} \tag{1.2}$$

$$f_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i|\mathbf{u}_i, \mathbf{v}_i) f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)} \tag{1.3}$$

$$E_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)}. \tag{1.4}$$

It follows from (1.4) that

$$\text{a) } E_s(w_i|\mathbf{v}_i) = \frac{1}{E_p(\pi_i|\mathbf{v}_i)}; \quad \text{b) } E_p(\mathbf{u}_i) = \frac{E_s(w_i\mathbf{u}_i)}{E_s(w_i)};$$
$$\text{c) } E_s(w_i) = \frac{1}{E_p(\pi_i)} \tag{1.5}$$

The proof of these relationships can be found in Pfeffermann and Sverchkov (1999, hereafter PS.) For further discussion of the notion of the sample distribution with illustrations see Pfeffermann et *al.* (1998).

1.2 The sample-complement distribution

Similarly to (1.1), we define the conditional pdf for units outside the sample as

$$f_c(y_i|\mathbf{x}_i) \overset{def}{=} f_p(y_i|\mathbf{x}_i, I_i = 0)$$

$$= \frac{\Pr(I_i = 0|y_i, \mathbf{x}_i) f_p(y_i|\mathbf{x}_i)}{\Pr(I_i = 0|\mathbf{x}_i)} \qquad (1.6)$$

The following relationships for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$ follow from (1.2)-(1.5) and the equality, $\Pr(I_i = 0 \mid \mathbf{u}_i, \mathbf{v}_i) = 1 - E_p(\pi_i \mid \mathbf{u}_i, \mathbf{v}_i)$.

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p[(1-\pi_i)|\mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]} \qquad (1.7)$$

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

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

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s[(w_i - 1)|\mathbf{u}_i, \mathbf{v}_i] f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s[(w_i - 1)|\mathbf{v}_i]} \qquad (1.8)$$

(follows by application of (1.5a) to the last expression in (1.7)). Also, by the first equality in (1.7) and (1.8)

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

$$= \frac{E_s[(w_i - 1)\mathbf{u}_i|\mathbf{v}_i]}{E_s[(w_i - 1)|\mathbf{v}_i]}. \qquad (1.9)$$

Remark 1. In practical applications the sampling fraction is ordinarily small and the sample selection probabilities are then likewise small for at least most of the population units. If $\pi_i < \delta$,

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

$$= f_p(\mathbf{u}_i|\mathbf{v}_i) + \frac{E_p\{[E_p(\pi_i|\mathbf{v}_i) - \pi_i]|\mathbf{u}_i, \mathbf{v}_i\} f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]} \qquad (1.10)$$

$$= f_p(\mathbf{u}_i \mid \mathbf{v}_i)(1 + \Delta)$$

where $-\delta < \Delta < \delta/(1-\delta)$, showing that for δ sufficiently small the difference between the population and the sample pdfs is accordingly small.

2. PREDICTION OF FINITE POPULATION TOTALS UNDER INFORMATIVE SAMPLING

Let $Y = \sum_{i=1}^{N} y_i$ define the population total. The problem considered is how to predict Y using the sample data and population values of concomitant variables $\mathbf{x}$ (when available). Denote the design information available for the prediction process by $D_S = \{(y_i, \pi_i) : i \in s, (\mathbf{x}_i, I_i) : i \in p\}$ and let $\hat{Y} = \hat{Y}(D_S)$ define the predictor. The MSE of $\hat{Y}$ given D_S with respect to the population pdf is,

$$MSE(\hat{Y}) = E_p[(\hat{Y} - Y)^2|D_s]$$

$$= E_p\{[\hat{Y} - E_p(Y|D_s)]^2|D_s\} + V_p(Y \mid D_s)$$

$$= [\hat{Y} - E_p(Y|D_s)]^2 + V_p(Y|D_s), \qquad (2.1)$$

since $[\hat{Y} - E_p(Y \mid D_s)]$ is fixed given D_S. It follows that $MSE(\hat{Y})$ is minimized when $\hat{Y} = E_p(Y|D_s)$. Now,

$$E_p(Y|D_s) = E_p(\sum_{i \in p} y_i|D_s) = \sum_{i \in p} E_p(y_i|D_s)$$

$$= \sum_{i \in s} E_p(y_i|D_s, I_i = 1) + \sum_{j \notin s} E_p(y_j|D_s, I_j = 0)$$

$$= \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j|D_s) = \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j|x_j) \qquad (2.2)$$

where in the last equality we assume that y_j for $j \notin s$ and $\{(y_i, \pi_i) : i \in s\}$ are independent given $\mathbf{x}_j$. The prediction problem reduces therefore to the prediction of $E_c(y_j|\mathbf{x}_j)$. In section 3 we consider semi-parametric estimation of the expectations $E_c(y_j|\mathbf{x}_j)$ and hence of Y.

3. SEMI-PARAMETRIC PREDICTION OF FINITE POPULATION TOTALS

Suppose that the sample-complement model takes the form

$$y_j = C_\beta(\mathbf{x}_j) + \varepsilon_j, \quad E_c(\varepsilon_j|\mathbf{x}_j) = 0,$$

$$E_c(\varepsilon_j^2|\mathbf{x}_j) = \sigma^2 v(\mathbf{x}_j), \quad j \notin s \qquad (3.1)$$

where $C_\beta(\mathbf{x})$ is a known (possibly nonlinear) function of $\mathbf{x}$ that depends on an unknown vector parameter β, and $v(\mathbf{x})$ is known but σ^2 unknown.

Remark 2. In actual applications the form of the model can be identified by a two-step procedure. First, estimate $E_s(w_i|\mathbf{x}_i)$ and $r_i = \dfrac{w_i - 1}{E_s[(w_i - 1)|\mathbf{x}_i]}$ by

regressing w_i against $\mathbf{x}_i$ using the sample data. Denote the resulting estimate by $\hat{r}_i = \frac{w_i - 1}{\hat{E}_s(w_i|\mathbf{x}_i) - 1}$ and let $y_i^* = \hat{r}_i y_i$. Second, analyze the relationship between y_i^* with $\mathbf{x}_i$ for identifying the form of $C_\beta(\mathbf{x}_i)$ utilizing the equality $E_c(y_i|\mathbf{x}_i) = E_s(r_i y_i|\mathbf{x}_i)$ (equation 1.9). A similar procedure can be used for identifying the function $v(\mathbf{x}_i)$ based on the empirical residuals $\hat{\varepsilon}_i = y_i - \hat{E}_s(\hat{r}_i y_i|\mathbf{x}_i)$.

For given forms of the functions $C_\beta(\mathbf{x}_i)$ and $v(\mathbf{x}_i)$, the vector β satisfies,

$$\beta = \arg\min_{\tilde{\beta}} E_c\left(\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Big|\mathbf{x}_j\right)$$

$$= \arg\min_{\tilde{\beta}} E_s\left(\left[\frac{w_j - 1}{E_s(w_j|\mathbf{x}_j) - 1}\right]\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Big|\mathbf{x}_j\right). \quad (3.2)$$

Hence, if $w(\mathbf{x}) = E_s(w|\mathbf{x})$ can be identified and estimated properly, the vector β can be estimated as

$$\hat{\beta}_1 = \arg\min_{\tilde{\beta}} \sum_{i\in s}\left(\hat{r}_i \frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)}\right) \quad (3.3)$$

where $\hat{r}_i = (w_i - 1)/(\hat{E}_s(w_i|\mathbf{x}_i) - 1)$. The predictor of the population total takes the form

$$\hat{Y}_1 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_1}(\mathbf{x}_j). \quad (3.4)$$

On the other hand, it follows from (3.1) that

$$E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Big|\mathbf{x}_j\right) = E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right). \quad (3.5)$$

Thus, application of (1.9) to the right hand side of (3.5) but without the conditioning on $\boldsymbol{x}$ implies that,

$$\beta = \arg\min_{\tilde{\beta}} E_c\left(\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right)$$

$$= \arg\min_{\tilde{\beta}} E_s\left(\left[\frac{w_j - 1}{E_s(w_j) - 1}\right]\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right) \quad (3.6)$$

and β can be estimated as,

$$\hat{\beta}_2 = \arg\min_{\tilde{\beta}} \sum_{i\in s}(w_i - 1)\frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \quad (3.7)$$

since $E_s(w_i) = constant$. The predictor of Y is now

$$\hat{Y}_2 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_2}(\mathbf{x}_j). \quad (3.8)$$

Remark 3. The prominent 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 the expectation $w(\mathbf{x})$ can be estimated properly, the predictor $\hat{Y}_1$ is expected to be more accurate since the weights $r_i = (w_i - 1)/(E_s(w_i|\mathbf{x}_i) - 1)$ are less variable than the weights $(w_i - 1)$.

This follows from the fact that 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 effects 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 ignorable and $f_c(y_i|\mathbf{x}_i) = f_s(y_i|\mathbf{x}_i)$. In this case the estimator $\hat{\beta}_1$ coincides with the optimal Generalized Least Square (GLS) estimator of β since $r_i = 1$ and the model (3.1) holds for the sample data.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and hence the predictors $\hat{Y}_1$ and $\hat{Y}_2$ coincide when the w_i are independent of $\mathbf{x}_i$ since in this case $w(\mathbf{x}_i) = constant$.

The use of the predictors $\hat{Y}_1$ and $\hat{Y}_2$ requires the identification of the sample-complement model. Next we develop another predictor that only requires the identification of the sample model. The approach leading to this predictor is a 'sample-complement analogue' of the "bias correction method" proposed in Chambers, Dorfman and Sverchkov (2001). The proposed predictor bases on the following relationship,

$$\sum_{j\notin s} E_c(y_j|\mathbf{x}_j) =$$

$$= \sum_{j\notin s} E_s(y_j|\mathbf{x}_j) + (N-n)\frac{1}{N-n}\sum_{j\notin s} E_c\{[y_j - E_s(y_j|\mathbf{x}_j)]|\mathbf{x}_j\}$$

$$\approx \sum_{j\notin s} E_s(y_j|\mathbf{x}_j) + (N-n)\frac{1}{N-n}\sum_{j\notin s} E_c[y_j - E_s(y_j|\mathbf{x}_j)], \quad (3.9)$$

where in the last row we replaced the mean of the conditional expectations $E_c\{[y_j - E_s(y_j|\mathbf{x}_j)]|\mathbf{x}_j\}$ by the mean of the unconditional expectations $E_c[y_j - E_s(y_j|\mathbf{x}_j)]$. By (1.9),

$$E_c[y_j - E_s(y_j|\mathbf{x}_j)] = E_s\left(\frac{w_j - 1}{E_s(w_j) - 1}[y_j - E_s(y_j|\mathbf{x}_j)]\right),$$

so that the sample-complement mean in the last row of (3.9) can be estimated by

$$\hat{M}_c = \frac{1}{n}\sum_{i\in s}\left(\frac{w_i - 1}{\bar{w}_s - 1}[y_i - \hat{E}_s(y_i|\mathbf{x}_i)]\right);\ \bar{w}_s = \Sigma_{i\in s} w_i / n.$$

The proposed predictor takes therefore the form

$$\hat{Y}_3 = \sum_{i\in s} y_i + \sum_{j\notin s}\hat{E}_s(y_j|\mathbf{x}_j) + (N-n)\hat{M}_c \qquad (3.10)$$

with $\hat{E}_s(y_j|\mathbf{x}_j)$ estimated from the sample data.

The use of $\hat{Y}_3$ only requires the identification and estimation of the sample regression $E_s(y_j|\mathbf{x}_j)$, which can be carried out using conventional regression techniques. Moreover, under general conditions this predictor is 'design consistent' even if the expectation $\hat{E}_s(y_j|\mathbf{x}_j)$ is misspecified. Many analysts view the design consistency property as an essential requirement from any predictor, see the discussions in Hansen *et al.* (1983) and Sarndal (1980). For sampling designs such that $\sum_s w_i = N$ for all s, or if one estimates $\hat{E}_s(w_i) = N/n$, $\hat{Y}_3$ has the form of the Generalized Regression estimator (GREG), Sarndal (1980).

4. EXAMPLES

4.1 Prediction with no concomitant variables

Let $\mathbf{x}_i = 1$ for all i. Then,

$$\hat{Y} = \sum_{i\in s} y_i + \sum_{j\notin s}\hat{E}_c(y_j)$$

$$= \sum_{i\in s} y_i + (N-n)\hat{E}_s\left(\frac{w_j - 1}{\hat{E}_s(w_j) - 1}y_j\right). \qquad (4.1)$$

Estimating the two unknown expectations in the second row of (4.1) by the respective sample means yields,

$$\hat{Y}_{EI} = \sum_{i\in s} y_i + (N-n)\frac{1}{n}\sum_{i\in s}\frac{w_i - 1}{\bar{w}_s - 1}y_i$$

$$= \sum_{i\in s} y_i + \frac{(N-n)}{\sum_{i\in s}(w_i - 1)}\sum_{i\in s}(w_i - 1)y_i. \qquad (4.2)$$

In (4.2) $\sum_{i\in s}(w_i - 1)y_i$ is the Horvitz-Thompson estimator of $\sum_{j\notin s} y_j$. The multiplier $\frac{(N-n)}{\sum_{i\in s}(w_i - 1)}$ is a "Hajek type correction" for controlling the variability of the sampling weights. 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 standard Horvitz-Thompson estimator $\hat{Y}_{H-T} = \sum_{i\in s} w_i y_i$.

Remark 4. The predictor $\hat{Y}_{EI}$ can be obtained as a special case of the Cosmetic predictors proposed by Brewer (1999). It should be emphasized, however, that the development of these predictors and the derivation of their prediction MSE assumes explicitly noninformative sampling. In particular, if the population model is the linear regression model with an intercept and constant variances and the sampling design is such that $\sum_{i\in s} w_i = N$ for all s, the recommended cosmetic predictor coincides with the optimal predictor under the population model.

Rather than only 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 (1.4), $E_p(y_i) = E_s(w_i y_i)/E_s(w_i)$ and estimating again the sample expectations by the corresponding sample means yields the familiar Hajek estimator,

$$\hat{Y}_{Hajek} = \sum_{k=1}^{N}\hat{E}_p(y_k) = N\hat{E}_s\left(\frac{w_i y_i}{\hat{E}_s(w_i)}\right) = \frac{N}{\sum_{i\in s} w_i}\sum_{i\in s} w_i y_i \qquad (4.3)$$

The predictors in (4.2) and (4.3) are the same, and they coincide with the Horvitz-Thompson estimator for sampling designs such that $\sum_{i\in s} w_i = N$. Notice, on the other hand, that by considering the estimation of Y as a prediction problem, one has to predict $(N\text{-}n)$ values under the approach leading to the use of $\hat{Y}_{EI}$ and N values under the approach leading to $\hat{Y}_{Hajek}$. For n/N sufficiently large, one can expect the predictor $\hat{Y}_{EI}$ to be superior (see Section 5).

4.2 Prediction with concomitant variables

Let the population model be,

$$y_i = H_\beta(\mathbf{x}_i) + \varepsilon_i,\ E_p(\varepsilon_i|\mathbf{x}_i) = 0,$$
$$E_p(\varepsilon_i^2|\mathbf{x}_i) = v(\mathbf{x}_i),\ E_p(\varepsilon_i\varepsilon_j|\mathbf{x}_i, \mathbf{x}_j) = 0 \ \text{for}\ i \neq j \qquad (4.4)$$

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

$$\pi_i = K\times[y_i g(\mathbf{x}_i) + \delta_i],\ E_p(\delta_i|\mathbf{x}_i, y_i) = 0 \qquad (4.5)$$

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'.)

Under (4.4), $\pi(x_i) = E_p(\pi_i | \mathbf{x}_i) = KH_\beta(\mathbf{x}_i)g(\mathbf{x}_i)$. Hence, by (1.9) and (4.5),

$$E_c(y_i|\mathbf{x}_i) = E_p\left(\frac{1-\pi_i}{1-\pi(\mathbf{x}_i)} y_i|\mathbf{x}_i\right)$$
$$= E_p\left(\frac{1-\pi(\mathbf{x}_i) - K\varepsilon_i g(\mathbf{x}_i) - K\delta_i}{1-\pi(\mathbf{x}_i)} y_i|\mathbf{x}_i\right)$$
$$= E_p(y_i|\mathbf{x}_i) - \frac{Kg(\mathbf{x}_i)v(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}. \quad (4.6)$$

The last expression in (4.6) shows that $E_c(y_i|\mathbf{x}_i) < E_p(y_i|\mathbf{x}_i) = H_\beta(\mathbf{x}_i)$ which is clear since for the inclusion probabilities defined by (4.5), the sample-complement tends to include the units with the smaller y- values. Note, however, that as $K \to 0$, $(n/N \to 0)$, $E_p(y_i|\mathbf{x}_i) - E_c(y_i|\mathbf{x}_i) \to 0$.

As a special case of (4.4), suppose that there is a single auxiliary variable x and that $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$. As well known, for noninformative sampling and with unknown β, the optimal predictor of Y, (minimizing $E_p[(\hat{Y}-Y)^2|D_s]$ is in this case the familiar Ratio estimator $\hat{Y}_R = N\frac{\overline{X}}{\overline{x}}\overline{y}$ with $\overline{y}$ and $\overline{x}$ denoting the respective sample means and $\overline{X}$ defining the population mean; see, e.g., Royall (1970).

Let in (4.5) $g(x) = 1$ for all x so that $\pi_i = \frac{n(y_i+\delta_i)}{\sum^N (y_i+\delta_i)}$, which for sufficiently large N and under some regularity conditions can be approximated as $\pi_i \approx \frac{n(y_i+\delta_i)}{N\beta\overline{X}}$, implying that $E_p(\pi_i|x_i) \approx \frac{nx_i}{N\overline{X}}$. By (4.6) $E_c(y_j|x_j) = x_j\beta - \sigma^2 x_j/[\beta(f^{-1}\overline{X} - x_j)]$ where $f = n/N$ is the sampling fraction, so that for known β and σ^2 Y is predicted as

$$\hat{Y}_{\text{EII}} = \sum_{i\in s} y_i + \beta \sum_{j\notin s} x_j - \frac{\sigma^2}{\beta} \sum_{j\notin s} \frac{x_j}{f^{-1}\overline{X} - x_j}. \quad (4.7)$$

Remark 5. Under noninformative sampling and with β known, the optimal predictor of Y is, $\hat{Y} = \sum_{i\in s} y_i + \beta \sum_{j\notin s} x_j$ and the prediction MSE is $E_p[(\hat{Y}-Y)^2|D_s] = \sigma^2 \sum_{j\notin s} x_j$. As easily verified, the prediction MSE of $\hat{Y}_{\text{EII}}$ under the same population model but with the sample selection probabilities defined by (4.5) with $g(x) = 1$ is,

$$MSE_p(\hat{Y}_{\text{EII}}) = \sigma^2 \sum_{j\notin s} x_j + (\sigma^2/\beta)^2 \sum_{j\notin s} [x_j^2/(f^{-1}\overline{X} - x_j)^2]$$

4.3 Design consistent regression predictors

The predictor defined by (4.7) is strictly 'model dependent' and as illustrated by Hansen *et. al* (1983), small deviations from this model, not easily detected from the sample data (even under noninformative sampling) may yield poor predictors. Consider therefore the following alternative family of regression predictors which is based on similar principles underlying the use of the classical regression estimator in the case of noninformative sampling.

$$Y_{\text{I,Reg}} = \sum_{i\in s} y_i + \hat{Y}_c + B_c(\hat{X}_c - X_c) \quad (4.8)$$

where $(Y_c, X_c) = \sum_{j\notin s}(y_j, x_j)$ and $(\hat{Y}_c, \hat{X}_c)$ are design consistent predictors of (Y_c, X_c). For a fixed coefficient B_c, $Y_{\text{I,Reg}}$ is design consistent for Y irrespective of the true population model. In practice, B_c can be replaced by a consistent estimator of the regression coefficient indexing the linear regression of y on x in the sample-complement. For example, with a single concomitant variable x, replace B_c by $\hat{B}_c$ where

$$\hat{B}_c = \frac{\hat{E}_c(y_j x_j) - \hat{E}_c(y_j)\hat{E}_c(x_j)}{\hat{E}_c(x_j^2) - \hat{E}_c^2(x_j)}, \quad (4.9)$$

with $\hat{E}_c(f_j) = \sum_{i\in s} \frac{w_i - 1}{\overline{w}_s - 1} f_i$; $f_j = y_j x_j, y_j, x_j, x_j^2$.

The replacement of B_c by $\hat{B}_c$ (and hence of $Y_{\text{I,Reg}}$ by $\hat{Y}_{\text{I,Reg}}$) preserves the design consistency property since

$$Y - \hat{Y}_{\text{I,Reg}} = (Y - Y_{\text{I,Reg}}) + (\hat{B}_c - B_c)(X_c - \hat{X}_c). \quad (4.10)$$

An example for a predictor in this family is obtained by estimating

$$\hat{Y}_c = \frac{(N-n)}{\sum_{i\in s}(w_i - 1)} \sum_{i\in s} (w_i - 1) y_i \quad (4.11)$$

(second component of (4.2)) and similarly for $\hat{X}_c$.

5. EMPIRICAL ILLUSTRATIONS

In order to illustrate the performance of the predictors proposed 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 birthweight,* see Korn and Graubard (1995) for more details. For the present illustrations 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 population mean of *birthweight*, using *gestational age* as the concomitant variable. The working regression model postulated for the sample-complement is the third order polynomial regression,

$$y_j = \beta_0 x_j + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3 + \varepsilon_j = C_\beta(x_j) + \varepsilon_j , \quad (5.1)$$

with the residuals assumed to be independent and with constant variances. The computation of the predictor $\hat{Y}_1$ defined by (3.4) requires also the identification and estimation of the expectation $w(x) = E_s(w \mid x)$ and for this we used the procedure described in Pfeffermann and Sverchkov (1999). (The latter article uses the same data for illustrating the performance of regression estimators derived from the sample distribution.)

Table 1 shows the empirical bias and Root Mean Square Error (RMSE) of the various predictors as obtained when drawing 1000 samples of size n=1726. (Unconditional bias and RMSE over all samples.) The 'population' size is N=9948. The first estimator, $\hat{Y}_{\mathrm{Re}\,g}$ is the unweighted regression estimator (based on the three powers of the concomitant variable) and its relative large bias indicates the high degree of informativeness of the sample selection. The other predictors are defined in the previous sections. As can be seen, all these predictors are virtually unbiased although the last four predictors are statistically biased based on the conventional t-statistics.

The other notable result emerging from the table is the very large variance of $\hat{Y}_{H-T}$. The predictor $\hat{Y}_{Hajek}$ is much more stable but as suggested in Section 4, the predictor $\hat{Y}_{EI}$ which only predicts the values for the sample- complement has an even smaller variance. The use of the concomitant variable in the last four predictors further reduces the variance and they all perform equally well in the present study. In particular, estimating the expectation $w(x)$ for calculating the predictor $\hat{Y}_1$ does not reduce the variance compared to the use of the predictor $\hat{Y}_2$.

REFERENCES

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

Chambers, R. L., Dorfman, A., and Sverchkov, M. (2001). Nonparametric regression with complex survey data. In, *Analysis of Complex surveys*, Ed. C. Skinner and R. Chambers, Wiley (forthcoming)

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.

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.

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. *Sankhya*, Series B, **61**, 166-186.

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

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

Table 1. Bias and RMSE of Alternative Predictors; True mean = 2871
Population size = 9948, Sample size = 1726, Number of samples = 1000

Predictor	$\hat{Y}_{\mathrm{Re}\,g}$	$\hat{Y}_{H-T}$	$\hat{Y}_{Hajek}$	$\hat{Y}_{EI}$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{1,\mathrm{Re}\,g}$
Bias	326.2	0.13	0.85	0.61	4.26	4.05	2.63	2.32
RMSE	326.7	130.2	34.1	28.5	24.0	23.9	23.7	24.5