Journal of Official Statistics, Vol. 11. No. 3, 1995, pp. 277–294

Estimating Distribution Functions with Auxiliary Information using Poststratification

P.L.D. Nascimento Silva[1] *and C.J. Skinner*[1]

The estimation of a finite population distribution function is considered when auxiliary population information is available. A simple poststratification procedure is compared to a number of more sophisticated methods which have been proposed in the literature. Comparisons are made with respect to a number of criteria and via a simulation study based on two real populations. Whereas some more sophisticated procedures can display greater efficiency in certain circumstances, this can be compensated for by lack of robustness or by other practical disadvantages. Poststratification emerges as a simple and practical procedure offering some useful gains in efficiency.

Key words: Regression estimation; finite population; survey data; sampling; linear weighting.

1. Introduction

The distribution function $F(t)$ for a variable y and a population U is the proportion of units in U for which the value of y is less than or equal to t. Such functions may be of considerable interest when, for example, y is a measure of wages or income and the units are individuals or households, or when y is a measure of size, such as number of employees, and the units are establishments.

We consider here the problem of estimating $F(t)$ given sample values of y together with auxiliary population information. One example of this problem which led to our interest in this subject arose in the 1991 Brazilian Population Census, where measures of income are recorded for a sample of households[2] and auxiliary values of a crude measure of income and of several other variables are recorded for 100% of the population enumerated.

Since $F(t)$ is simply a population proportion for any given value of t, conventional methods for estimating means such as ratio and regression estimation may be used to take advantage of the auxiliary information. In many circumstances such standard approaches may be satisfactory. However, sometimes, such as in the Brazilian census

[1] Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, U.K.

[2] 10% of the households are selected in each enumeration district, except for small municipalities, where the sampling fraction is 20%.

Acknowledgements: Pedro Luis do Nascimento Silva is currently on leave from the Instituto Brasileiro de Geografia e Estatística (IBGE). His work was supported by a grant from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). C.J. Skinner's work was supported by grant H519255005 from the Economic and Social Research Council. The authors are very grateful to Ray Chambers for providing the data and, together with two anonymous referees, several useful suggestions which helped improve the paper.

example, the auxiliary information may include the population values of a variable x which is a close proxy for y. In such a situation it seems reasonable to expect that an estimator of $F(t)$ should approach $F(t)$ arbitrarily closely as x approaches y.

This property is not possessed by the conventional ratio or regression estimators, however, essentially because the survey variable $I(y \leqslant t)$ (where the indicator function $I(\cdot)$ takes the value 1 if its argument is true and 0 otherwise) is not in general perfectly correlated with y. This suggests that more efficient use of the auxiliary information is possible.

One approach would simply be to take a ratio or difference estimator based on the binary auxiliary variable $I(x \leqslant t)$, or more generally $I(\hat{R}x \leqslant t)$, as proposed by Rao, Kovar, and Mantel (1990), where $\hat{R}$ estimates the ratio of the y total to the x total in the population. A series of yet more sophisticated estimators has been proposed by Chambers and Dunstan (1986), Kuo (1988), Rao et al. (1990), Chambers, Dorfman, and Wehrly (1993) and Kuk (1993). See also Chambers, Dorfman, and Hall (1992), Rao and Liu (1992), Dorfman (1993) and Chen and Qin (1993).

Here we consider instead a simpler approach, namely a poststratified estimator, with poststrata defined by the intervals of x. Poststratification is an approach familiar to most survey statisticians and has a number of practical advantages discussed in Section 3. Our intention is not to attempt to show that the poststratified estimator markedly outperforms other estimators but rather the converse: if the poststratified estimator is taken as a simple "benchmark" estimator, is there evidence that other more complicated estimators have sufficient advantages to make their use preferable in practice?

This paper has consequently three aims: first, to provide a theoretical comparison of the poststratified estimator with some alternative estimators; second, to provide a corresponding numerical comparison and third, to provide empirical evidence regarding the effect of alternative choices of the poststratum intervals.

The alternative estimators will be defined in Section 2 and compared theoretically in Section 3. The numerical comparison will be presented in Section 4 and our conclusions in Section 5.

2. Alternative Estimators

Let U be a finite population of size N. Let $s \subset U$ be a sample drawn according to a known probability sampling design. Let (x_i, y_i) be pairs of values associated with each unit $i \in U$. Suppose that the values y_i, $i \in s$, and x_i, $i \in U$, are known. The problem is how to use this information to make inference about the finite population distribution function

$$F(t) = N^{-1} \sum_{i \in U} I(y_i \leqslant t). \tag{1}$$

We list below a number of alternative point estimators $\hat{F}(t)$ of $F(t)$. Estimators $s.e.[\hat{F}(t)]$ of the standard error of $\hat{F}(t)$ will be considered briefly in Section 3.7. Inference would usually proceed assuming that $[\hat{F}(t) - F(t)]/s.e.[\hat{F}(t)]$ follows the standard normal distribution. The adequacy of this assumption as well as the

question of how to construct uniform (as opposed to pointwise) confidence regions for the function $F(t)$ will not be considered here.

The simple (ordinary) design-based estimator of $F(t)$ which makes no use of auxiliary information at the estimation stage is

$$\hat{F}_0(t) = \frac{\sum_{i \in s} I(y_i \leqslant t)/\pi_i}{\sum_{i \in s} 1/\pi_i} \tag{2}$$

where π_i is the inclusion probability of unit i.

To define the poststratified estimator, let the G poststrata partitioning U be denoted $U_1, \ldots, U_G$ where $i \in U_g$ if $x_{(g-1)} < x_i \leqslant x_{(g)}$, and where $x_{(0)} = -\infty$, $x_{(1)} < x_{(2)} < \ldots < x_{(G-1)}$ are specified values and $x_{(G)} = \infty$. Let $s_1, \ldots, s_G$ be the corresponding partition of s so that $s_g = s \cap U_g$. Let N_g be the size of U_g and let $\hat{N}_g = \sum_{i \in s_g} 1/\pi_i$, $g = 1, \ldots, G$.

Then the poststratified estimator of $F(t)$ is

$$\hat{F}_{ps}(t) = N^{-1} \sum_{g=1}^{G} \frac{N_g}{\hat{N}_g} \sum_{i \in s_g} \frac{I(y_i \leqslant t)}{\pi_i} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{F}_g(t) \tag{3}$$

where $\hat{F}_g(t) = \hat{N}_g^{-1} \sum_{i \in s_g} I(y_i \leqslant t)/\pi_i$.

While it is sensible in practice to define the poststrata in such a way that the probability that s_g is empty (and hence that $\hat{N}_g = 0$ and $\hat{F}_{ps}(t)$ is undefined) is very small, the theoretical possibility that this event occurs needs formally to be addressed in a design-based framework. For this purpose we suppose that (as discussed, e.g., by Little 1993) any poststrata with $\hat{N}_g = 0$ are pooled with adjacent poststrata until all $\hat{N}_g$ are positive (see Fuller 1966, for an alternative approach).

Note that $\hat{F}_{ps}(t)$ may alternatively be expressed as a multiple regression estimator (c.f. Särndal, Swensson, and Wretman 1992, p. 264) with G auxiliary variables indicating poststrata membership.

Turning now to alternative estimators proposed in the literature to make efficient use of auxiliary information, we consider first the model-based estimator suggested by Chambers and Dunstan (1986). This estimator is based on a working model

$$y_i = \beta x_i + \sigma(x_i) u_i \tag{4}$$

where β is an unknown parameter, the $\sigma(x_i)$ are known and the u_i are independent and identically distributed outcomes of a random variable with zero mean. The estimator proposed by Chambers and Dunstan is

$$\hat{F}_{cd}(t) = N^{-1} \left[\sum_{i \in s} I(y_i \leqslant t) + n^{-1} \sum_{j \in U-s} \sum_{i \in s} I\left(u_{ni} \leqslant \frac{t - b_n x_j}{\sigma(x_j)} \right) \right] \tag{5}$$

where

$$b_n = \left[\sum_{i \in s} \frac{y_i x_i}{\sigma^2(x_i)} \right] \left[\sum_{i \in s} \frac{x_i^2}{\sigma^2(x_i)} \right]^{-1}, \qquad u_{ni} = \frac{y_i - b_n x_i}{\sigma(x_i)}$$

and $U - s$ denotes the set of nonsampled units.

Following instead a design-based approach, Rao et al. (1990) proposed three estimators. The first are the ratio estimator $\hat{F}_r(t)$ and the difference estimator $\hat{F}_d(t)$, obtained by treating $I(y_i \leqslant t)$ as the survey variable and $I(\hat{R}x_i \leqslant t)$ as an auxiliary variable, where

$$\hat{R} = \left[\sum_{i \in s} \frac{y_i}{\pi_i}\right] \Big/ \left[\sum_{i \in s} \frac{x_i}{\pi_i}\right]$$

estimates the ratio $R = Y/X$ of the population totals.

The third is the model-assisted (modified) difference estimator

$$\hat{F}_{dm}(t) = N^{-1}\left[\sum_{i \in s} \frac{I(y_i \leqslant t)}{\pi_i} - \sum_{i \in s} \frac{\hat{G}_{ic}}{\pi_i} + \sum_{i \in U} \hat{G}_i\right] \tag{6}$$

where

$$\hat{G}_i = \left[\sum_{j \in s} I\left(\hat{u}_j \leqslant \frac{t - \hat{R}x_i}{\sqrt{x_i}}\right) \Big/ \pi_j\right] \left[\sum_{j \in s} 1/\pi_j\right]^{-1}$$

$$\hat{G}_{ic} = \left[\sum_{j \in s} I\left(\hat{u}_j \leqslant \frac{t - \hat{R}x_i}{\sqrt{x_i}}\right) \pi_i \Big/ \pi_{ij}\right] \left[\sum_{j \in s} \pi_i/\pi_{ij}\right]^{-1}, \qquad \hat{u}_j = \frac{y_j - \hat{R}x_j}{\sqrt{x_j}}$$

and π_{ij} is the joint inclusion probability of units i and j.

The last two estimators to be considered here are the nonparametric kernel estimators proposed by Kuo (1988) and Kuk (1993), given respectively by

$$\hat{F}_{ko}(t) = N^{-1}\left[\sum_{i \in s} I(y_i \leqslant t) + \sum_{j \in U-s} \sum_{i \in s} w_{ij} I(y_i \leqslant t)\right] \tag{7}$$

$$\hat{F}_{kk}(t) = N^{-1} \sum_{j \in U} \hat{R}_j \tag{8}$$

where

$$w_{ij} = \frac{K[(x_j - x_i)/b]}{\sum_{i \in s} K[(x_j - x_i)/b]}$$

are weights for Kuo's estimator, $K(z) = e^{-z^2/2}$ is the standard normal density (kernel),

$$\hat{R}_j = \frac{\sum_{i \in s} w[(x_j - x_i)/b] W[(t - y_i)/b]/\pi_i}{\sum_{i \in s} w[(x_j - x_i)/b]/\pi_i},$$

where

$W(z) = e^z/(1 + e^z)$ is the standard logistic distribution function with density $w(z) = e^z/(1 + e^z)^2$ and b is the bandwidth parameter used to control the amount of smoothing.

3. Theoretical Comparison of Estimators

We now compare the estimators listed in Section 2 with respect to a number of criteria which may be important in practice.

3.1. *Is $\hat{F}(t)$ a genuine distribution function?*

For an estimator $\hat{F}(t)$ of $F(t)$ to be a genuine distribution function it should be monotonic increasing and such that $\hat{F}(-\infty) = 0$, $\hat{F}(\infty) = 1$. These properties may easily be verified for $\hat{F}_o(t)$, $\hat{F}_{ps}(t)$, $\hat{F}_{cd}(t)$, $\hat{F}_{ko}(t)$ and $\hat{F}_{kk}(t)$. However, none of the three estimators of Rao et al. (1990) is monotonic increasing in general, as noted by Kuk (1993), although this property could be achieved by suitable transformation as pointed out by the authors.

3.2. *Does $\hat{F}(t) = F(t)$ when $y_i = x_i$?*

As discussed in the introduction, it seems desirable that an estimator $\hat{F}(t)$ approaches $F(t)$ as x approaches y. This property clearly does not hold for the simple estimator $\hat{F}_o(t)$ since it makes no use of x information. For the poststratified estimator, if $y_i = x_i$ then it can be verified that $\hat{F}_{ps}(t) = F(t)$ for $t = x_{(g)}$ for $g = 1, \ldots, G-1$. For other values of t, equality will not hold in general, although we may expect deviations to be small if the poststrata are reasonably fine. The property does hold for each of the estimators $\hat{F}_{cd}(t)$, $\hat{F}_r(t)$, $\hat{F}_d(t)$ and $\hat{F}_{dm}(t)$ but not in general for $\hat{F}_{ko}(t)$ and $\hat{F}_{kk}(t)$.

3.3. *Is the use of auxiliary information flexible?*

So far we have assumed that the auxiliary information consists of the population values of x_i. In practice, the available information will often be either less or more than this and it is therefore important that the estimation procedure be sufficiently flexible to adapt to such circumstances.

Sometimes the available information on a continuous variable such as age will only consist of the numbers N_g in certain intervals of values of the variable. In such circumstances the poststratified estimator is evidently still calculable whereas the other estimators which require individual x_i values are not.

On the other hand, often not only are the x_i values available but so too are the values of other auxiliary variables. In this situation the poststratified estimator may be extended by treating it as a multiple regression estimator. The estimators $\hat{F}_{cd}(t)$, $\hat{F}_d(t)$ and $\hat{F}_{dm}(t)$ may similarly be extended by replacing $b_n x_i$ and $\hat{R}x_i$, respectively, by the appropriate estimated value of a linear predictor including the values of the additional variables. The extension of the estimators $\hat{F}_r(t)$, $\hat{F}_{ko}(t)$ and $\hat{F}_{kk}(t)$ seems less straightforward, however.

3.4. *Is computation simple?*

Computation of an estimator $\hat{F}(t)$ is particularly simple if it can be expressed in the usual weighted form

$$\hat{F}(t) = \sum_{i \in s} w_i I(y_i \leqslant t)$$

where the weights w_i depend neither on the y_i values nor on t. Of the estimators

considered in Section 2, only the simple estimator $\hat{F}_o(t)$, the poststratified estimator $\hat{F}_{ps}(t)$ and the kernel estimator $\hat{F}_{ko}(t)$ possess this property.

Because of this property, the poststratified estimator can easily be calculated using the same weights as are used for the estimation of any mean. This property is particularly useful when there are many survey variables for which either distribution functions or means are to be estimated.

The kernel estimator $\hat{F}_{ko}(t)$ also has this property but the weights w_i would be more costly to compute. The computation of the estimators $\hat{F}_{cd}(t)$, $\hat{F}_{dm}(t)$ and $\hat{F}_{kk}(t)$ is even more intensive, requiring double summations over both s and U.

3.5. *Is definition of $\hat{F}(t)$ automatic?*

A further consideration related to simplicity of computation is whether the definition of the estimator is automatic, in the sense that no choices are required. The estimators $\hat{F}_o(t)$, $\hat{F}_r(t)$ and $\hat{F}_d(t)$ are automatic in this sense. The poststratified estimator is not in general automatic unless the auxiliary information is already grouped into prespecified categories such as age groups. Otherwise the cut-off values $x_{(g)}$ must be chosen.

The model-based estimator $\hat{F}_{cd}(t)$ similarly requires the initial specification of a model of the form (4) and, in particular, of a functional form for $\sigma(x)$. Similarly, the estimator $\hat{F}_{dm}(t)$ of Rao et al. (1990) is model-dependent in the sense that they suggest replacing $\sqrt{x}$ in $\hat{F}_{dm}(t)$ by $\sigma(x)$ if a model of form (4) is supposed to hold with $\sigma(x) \neq \sqrt{x}$. Finally both the nonparametric estimators $\hat{F}_{ko}(t)$ and $\hat{F}_{kk}(t)$ require the specification of the bandwidth b, with Kuk's estimator also requiring appropriate scaling of the response variable.

3.6. *Bias*

The moments of each estimator may be evaluated either with respect to a model, such as (4), or with respect to the sampling design. From a model-based point of view, Chambers and Dunstan (1986) demonstrate the asymptotic unbiasedness of $\hat{F}_{cd}(t)$. Similarly Rao et al. (1990) note the asymptotic model-unbiasedness of $\hat{F}_{dm}(t)$. The poststratified estimator is exactly model-unbiased under a model for which y_i has a common mean within each poststratum (e.g., Valliant 1993). There may, however, be some model-bias under a model such as (4).

From a design-based point of view, the poststratified estimator takes the form of a separate ratio estimator and hence will be asymptotically unbiased under standard sampling designs. The exact bias will depend on the rule for collapsing poststrata when $\hat{N}_g = 0$ for any g. For some sampling designs, such as simple random sampling, $\hat{F}_{ps}(t)$ will be exactly unbiased, conditional on values of $\hat{N}_g > 0$ for all g. Fuller (1966) indicates how poststratified estimators can be modified to make them unconditionally unbiased in such cases. The conditional properties of the poststratified estimator under general sampling designs are considered by Rao (1985) and Casady and Valliant (1993).

3.7. *Variance and its estimation*

A further desirable property of an estimator of $F(t)$ is that both a variance expression and an estimator of this variance are available. Rao et al. (1990) present expressions for the asymptotic design-based variances of the estimators $\hat{F}_r(t)$, $\hat{F}_d(t)$ and $\hat{F}_{dm}(t)$, as well as for estimators of these design-based asymptotic variances. However, they recognize that estimating the variance of $\hat{F}_{dm}(t)$ may be cumbersome under unequal probability sampling designs, since it involves computations with third order inclusion probabilities.

An approximate expression for the variance of the poststratified estimator, derived analogously to the results in Rao et al. (1990), is

$$\mathrm{Var}[\hat{F}_{ps}(t)] \doteq V[I(y_i \leqslant t) - F_{g(i)}(t)] \tag{9}$$

where the operator notation V is defined by

$$V(a_i) = N^{-2} \sum_{i<j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{a_i}{\pi_i} - \frac{a_j}{\pi_j} \right)^2 \tag{10}$$

for an argument a_i, where $g(i)$ is the poststratum to which unit i belongs and where $F_g(t) = N_g^{-1} \sum_{i \in U_g} I(y_i \leqslant t)$ is the population distribution function of y in poststratum g.

Expression (9) enables us to judge when the poststratified estimator will be most efficient. In particular, its asymptotic variance will be zero if the y values within each poststratum are either all above t ($F_g(t) = 1$) or else all below t ($F_g(t) = 0$). This also implies that no single poststratification could be the best for all t.

The asymptotic variance of $\hat{F}_{ps}(t)$ may be estimated by replacing $F_{g(i)}(t)$ in (9) by $\hat{F}_{g(i)}(t)$ and by replacing V in (9) by its sample analogue v

$$v(a_i) = N^{-2} \sum_{i<j \in s} (\pi_i \pi_j - \pi_{ij}) \pi_{ij}^{-1} \left(\frac{a_i}{\pi_i} - \frac{a_j}{\pi_j} \right)^2. \tag{11}$$

A variance estimator with possibly superior conditional properties, following Rao (1985) and Särndal, Swensson, and Wretman (1989), is obtained by replacing a_i in (11) by $N_{g(i)} a_i / \hat{N}_{g(i)}$.

Kuk (1993) presents an expression for the design-based variance of $\hat{F}_{kk}(t)$, which depends on the variances and covariances of the $\hat{R}_j$, as well as a corresponding variance estimator. The estimator can, however, be computationally intensive even for small sample sizes, since it depends on calculating variances and covariances between the N values $\hat{R}_j$.

4. Numerical Comparison of the Estimators

In this section we present the results of a Monte Carlo comparison of the various estimators of $F(t)$. The simulation study consisted of selecting 1,000 samples of sizes 30 and 50, by simple random sampling without replacement, from each of two

populations. From each sample and for each estimator in Section 2, estimates of the distribution function $F(t)$ were calculated for 11 different values of t, namely the quantiles t_q of the population distribution function such that $q = 1/12, \ldots, 11/12$, $F(t_q) = q$.

The bias (BIAS) and root mean squared error (RMSE) of each estimator for each quantile t_q were estimated by

$$\text{BIAS}(t_q) = \frac{1}{1000}\sum_{s}^{1000}[\hat{F}^s(t_q) - F(t_q)] = \frac{1}{1000}\sum_{s}^{1000}[\hat{F}^s(t_q) - q]$$

$$\text{RMSE}(t_q) = \sqrt{\frac{1}{1000}\sum_{s}^{1000}[\hat{F}^s(t_q) - F(t_q)]^2}$$

where $\hat{F}^s(t_q)$ is the value of a given estimator at the quantile t_q computed from sample s.

Corresponding aggregated measures of performance used to summarize the results for the various quantiles are the average absolute bias (AVAB) and the average root mean squared error (AVRMSE), given respectively by

$$\text{AVAB} = \frac{1}{11}\sum_{q}|\text{BIAS}(t_q)|$$

$$\text{AVRMSE} = \sqrt{\frac{1}{11}\sum_{q}[\text{RMSE}(t_q)]^2}.$$

We also consider a global measure of performance of the estimators across all 11 quantiles for each sample s, by computing the maximum absolute deviation (MAD) statistic

$$\text{MAD}(s) = \max_{q}|\hat{F}^s(t_q) - F(t_q)|.$$

4.1. *The simulation populations*

The first population comprises 338 sugar cane farms surveyed in 1982 in Queensland, Australia, as used originally by Chambers and Dunstan (1986), and later by Rao et al. (1990) and Kuk (1993). We took income from cane as y and area assigned for growing cane as x. Figure 1 displays a scatterplot of the population data we used.

The second population consists of 430 farms with 50 or more beef cattle surveyed in the 1988 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics. This population was originally used by Chambers et al. (1993) and subsequently by Kuk (1993) for their respective simulation studies. In this case, y is income from beef and x is the number of beef cattle in each farm. Chambers et al. (1993) present a scatterplot of the beef farms population data.

The two different populations enable us to investigate the relative efficiencies of the poststratified estimator versus the other estimators in both a situation where (4)

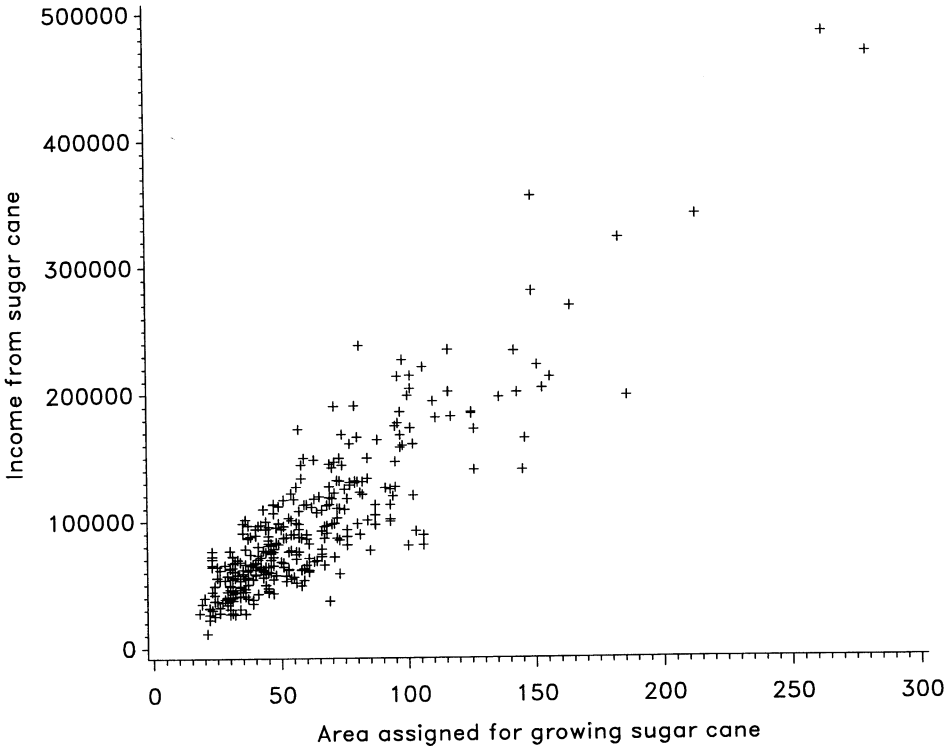

Fig. 1. Scatter plot of sugar cane farms data

provides a good working model of the relationship between y and x (the sugar cane farms) and also a situation where (4) is not a good model for that relationship (the beef farms). This second population should provide some indication of the "robustness" of the alternative estimators. Another reason for using both populations in our study was the availability of similar simulation results for some of the estimators considered here which could serve as a benchmark against which to compare our results.

4.2. *Alternative poststratification schemes*

To obtain evidence on the effect of choice of poststrata we considered three alternative schemes:

E. Equal numbers of units in the poststrata – this corresponds to choosing the values of $x_{(1)} < x_{(2)} < \ldots < x_{(G-1)}$ such that $N_g = N/G$ for all $g = 1, \ldots, G$;
R. Equal aggregate square root size in the poststrata – this corresponds to defining the values of $x_{(1)} < x_{(2)} < \ldots < x_{(G-1)}$ such that $\sum_{i \in U_g} \sqrt{x_i} = \sum_{i \in U} \sqrt{x_i}/G$ for all $g = 1, \ldots, G$;
T. Equal aggregate (Total) size in the various poststrata – this implies taking the values of $x_{(1)} < x_{(2)} < \ldots < x_{(G-1)}$ such that $\sum_{i \in U_g} x_i = \sum_{i \in U} x_i/G$ for all $g = 1, \ldots, G$.

Table 1. Average Root Mean Squared Error (AVRMSE) (*) for some versions of the poststratified estimator over 1,000 samples

Estimators	Sugar cane farms		Beef farms	
	$n = 30$	$n = 50$	$n = 30$	$n = 50$
Simple	756	548	748	574
Poststratified-4E	594	435	576	419
Poststratified-4R	601	435	597	444
Poststratified-4T	602	435	687	510

(*)Values multiplied by 10,000.

For each of the schemes E, R, and T, which correspond to equations (12.5.7), (12.5.8) and (12.5.9) of Särndal et al. (1992, p. 461–462), respectively, the population was partitioned into 2, 3, 4, and 5 poststrata, producing a total of 12 alternative poststratification versions, which we shall denote by codes such as 4E, for the poststratification using scheme E with 4 poststrata, and so on.

In Table 1 we present the results of the AVRMSE for the three versions of the poststratified estimator with four poststrata. The simple estimator is also included for reference. These results indicate that there is a reasonable gain of efficiency for the poststratified estimators compared to the simple estimator.

The same is true for other poststratification schemes and number of partitions. We note that the biggest percentage gains are achieved going from no poststrata (simple estimator) to two poststrata, for these populations and sample sizes. The best choice seems to be four poststrata, with little further gain if five poststrata are used, but little loss with only three poststrata.

We note that, because the beef farms population is very skewed, scheme T favours leaving the largest poststrata with only a few units. This increases the chances of poststrata having no sample units. When an empty poststratum was found in any sample, it was collapsed with the nearest non-empty poststratum, so that the poststratified estimator could be computed. In the extreme case, pooling corresponds to using the simple estimator.

We also observe that scheme T always produces the largest values of AVRMSE compared to schemes E and R for partitions of the population with the same number of poststrata. The "best" versions of the poststratified estimator are all obtained with scheme E, although scheme R produces only slightly worse results. Scheme E minimizes the probability of "empty poststrata" under the sampling design adopted and this may account for this scheme's superior properties for samples of these sizes.

We also used the simulation results to estimate the bias of the poststratification estimator. However, in our study we found no evidence of any significant bias, under all simulation conditions and for all the 11 quantiles considered. The estimates for the bias of the poststratified estimator were usually no larger than those for the simple estimator, which is theoretically unbiased under the simple random sampling without replacement design used to select the samples. The estimated bias was also generally much smaller than the root mean squared error.

The simulation standard error of each bias estimate is roughly 20 ($\doteq 600/\sqrt{1,000}$) and it seems that the deviations of the estimates of bias from zero for both post-

stratified and simple estimators can largely be accounted for by simulation error.

We also note that the RMSE figures for the various versions of the poststratified estimator are reasonably stable over the whole range of the quantiles studied, with a slight reduction at the extremes, due perhaps to the fact that $\hat{F}(t)$ and $F(t)$ are bounded above and below. It is also apparent that the "optimal" poststratification, in a minimum RMSE sense, varies from quantile to quantile (as predicted earlier), although in this example the optimal choice usually falls within versions of the partitioning scheme E for different numbers of poststrata.

We conclude this section by selecting the version 4E of the poststratified estimator which appears to give the "best" results over the set of simulation conditions, to be compared in the next section with the other estimators from the literature.

4.3. Poststratified estimator versus competitors

In this section we compare the poststratified estimator (version 4E) with the other competing estimators.

Note that to compute the model-based estimator $\hat{F}_{cd}(t)$ the variance function used was $\sigma(x_i) = \sqrt{x_i}$. The bandwidth used for the computation of Kuo's estimator $\hat{F}_{ko}(t)$ was $b = 1.06\sigma_x n^{-1/5}$.

Note also that for the computation of Kuk's estimator $\hat{F}_{kk}(t)$ the values of the response variable had to be scaled, since the bandwidth used to control the smoothing is the same for both variables x and y. In the case of the sugar cane data set, we divided the values of the response variable by 1,000.

For the beef farms data set, Kuk proposed not only scaling the y values by dividing them by 100, but he also performed a 1/4 power transformation on both variables. In order to study the sensitivity of Kuk's estimator to the choice of transformation we decided for this data set to compute two versions of Kuk's estimator: one using the "raw" data, only with the response variable "properly" scaled (divided by 100), and another following Kuk's suggestion and taking the 1/4 power transformation of both scaled response and auxiliary variables. For Kuk's estimator $\hat{F}_{kk}(t)$, the bandwidth used was calculated as the rounded value of the range of the

Table 2. Average Root Mean Squared Error (AVRMSE) (*) for several estimators over 1,000 samples

Estimators	Sugar cane farms		Beef farms	
	$n = 30$	$n = 50$	$n = 30$	$n = 50$
Simple	756	548	748	574
Kuo – nonparametric	633	456	720	551
Poststratified-4E	594	435	576	419
Modified difference	535	406	533	396
Kuk – transformed	(**)	(**)	428	337
Kuk – raw data	474	350	1193	903
Chambers–Dunstan	351	299	373	299

(*)Values multiplied by 10,000.
(**)Not available for the sugar cane farms.

auxiliary variable (x) divided by the sample size (n), after scaling and transforming the data.

Table 2 presents the AVRMSE statistics for all the estimators, except the ratio and difference estimators. These were excluded since they were clearly outperformed by all other estimators considered here, except the simple one, as expected in view of the results already obtained by Rao et al. (1990).

We see that there is considerable variation in the performance of the estimators in terms of average root mean squared error. The best overall performance is achieved by Chambers and Dunstan's model-based estimator. This was expected in the case of the sugar cane farms population, where model (4) describes the data very well. However, in the case of the beef farms population where the model does not fit, this result is more impressive, although the improvement over its closest competitors is smaller than for the sugar cane farms. We shall see later that this overall performance measure conceals some problems of bias for this estimator especially for the lower quantiles of the distribution, something already noted by Chambers et al. (1993) and Kuk (1993).

Kuk's estimator, computed from the raw data in the sugar cane population and from the transformed data in the beef farms population, shows the second best performance. However, the raw data version of Kuk's estimator in the beef farms population presents the worst performance among all the estimators considered. This is due to the huge asymmetry of the data and to the fact that a fixed bandwidth is used for the smoothing over the whole range of values of both x and y. This suggests that the choice of an adequate transformation of the data prior to calculating the estimator $\hat{F}_{kk}(t)$ is very important.

The modified difference estimator proposed by Rao et al. (1990) follows next, with values for AVRMSE which still provide a great improvement over the simple estimator. It is closely followed by the poststratified estimator, which presents only slightly worse results.

The poststratified estimator shows results which are always better than those for the ratio and difference estimators, the simplest competitors. It also performs better than Kuo's nonparametric estimator $\hat{F}_{ko}(t)$.

An analysis of the distribution of the maximum absolute deviations (MAD) computed for each estimator for each population and sample size confirms the ranking above. As an illustration, Figure 2 presents box-plots of the distributions of the MAD obtained for the beef farms population and samples of size 50.

We now turn our attention to the analysis of the estimated biases. Table 3 presents the measures of average absolute bias (AVAB) for the estimators considered.

These results suggest that the differences of the estimated biases from zero are largely accounted for by simulation error (remember the simulation standard error is roughly 20) in the case of the simple estimator, the modified difference estimator and the poststratified estimator (as already noted in Section 4.2). However, there are some noticeable biases for Chambers and Dunstan's model-based estimator and also for both nonparametric estimators. The estimated biases deviate significantly from zero for most quantiles, in both populations and for both sample sizes. The estimated biases for Kuk's estimator are usually smaller

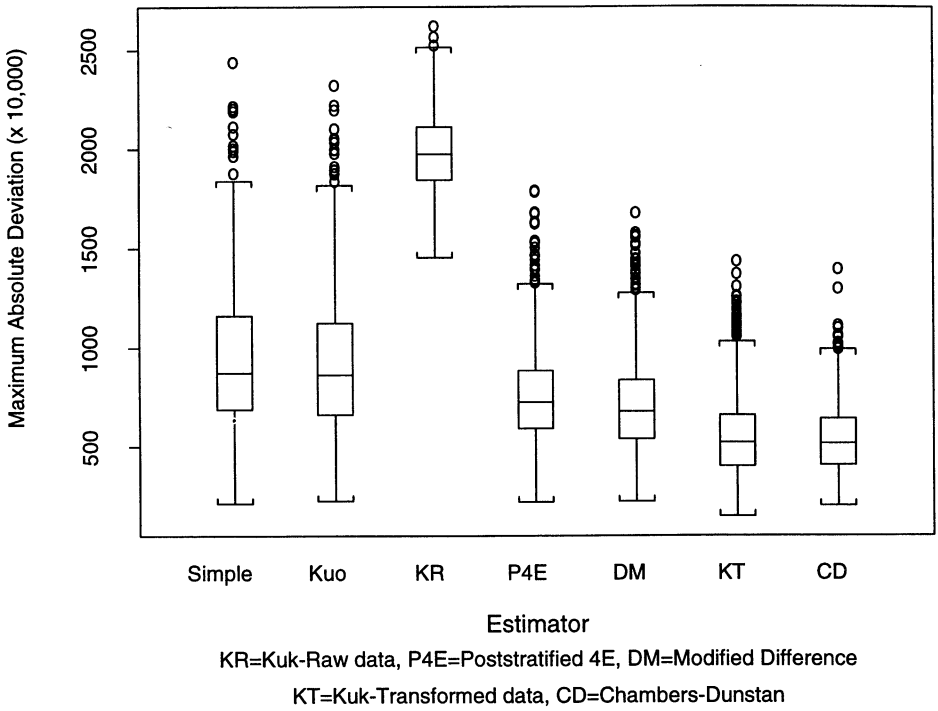

Fig. 2. Boxplots of maximum absolute deviation, beef farms data set, sample size = 50

(in absolute terms) than those for Chambers and Dunstan's estimator. For the beef farms population though, the raw data version of Kuk's estimator has some very large estimated biases. This is not surprising given the extreme skewness in the data.

Even though Chambers and Dunstan's estimator was often substantially biased, especially for the lower quantiles, this did not generally lead to a poor root mean squared error performance, as the results in Table 2 show. These results apply, however, only to sample sizes of 30 and 50, which are small in practice. On theoretical grounds one might expect the bias of Chambers and Dunstan's estimator to make a

Table 3. Average Absolute Bias (AVAB) (*) for several estimators over 1,000 samples

Estimators	Sugar cane farms		Beef farms	
	$n = 30$	$n = 50$	$n = 30$	$n = 50$
Simple	18	10	16	10
Kuo – nonparametric	111	95	165	146
Poststratified-4E	9	4	10	8
Modified difference	12	7	10	9
Kuk – transformed	(**)	(**)	91	46
Kuk – raw data	137	77	945	672
Chambers–Dunstan	189	181	152	135

(*)Values multiplied by 10,000.
(**)Not available for the sugar cane farms.

relatively greater contribution to the RMSE as the sample size increases. This was certainly the case when moving from $n = 30$ to $n = 50$. It is perhaps also relevant that $n = 200$ for the example presented by Rao et al. (1990) where Chambers and Dunstan's estimator has greater RMSE than their modified difference estimator for $q = 1/4$.

To investigate this further we performed a limited simulation exercise by selecting 200 simple random samples without replacement of size 300 from the beef farms population. The bias of Chambers and Dunstan's estimator became smaller, but it still tended to be much larger than for the poststratified estimator. For an unbiased estimator, such as the simple estimator, we would expect the RMSE to reduce by a factor of $\sqrt{300(1 - 30/430)/30(1 - 300/430)} \doteq 5.5$ as n increases from 30 to 300 ($N = 430$). This was indeed roughly the case for the simple and poststratified estimators. However, the RMSE of Chambers and Dunstan's estimator was reduced only by a factor of 3.9, since the relative contribution of the bias increased. The consequence was that, for $n = 300$, the poststratified estimator was roughly equally efficient to Chambers and Dunstan's estimator.

We next compare conditional biases, as discussed, for example, by Chambers and Dunstan (1986). Figure 3 presents a plot of the estimated conditional bias for several estimators of the distribution function for the first quartile ($q = 1/4$), for samples of size 30 from the sugar cane population. The conditional biases were estimated from 20 groups of 50 samples each, formed using the ordered values of

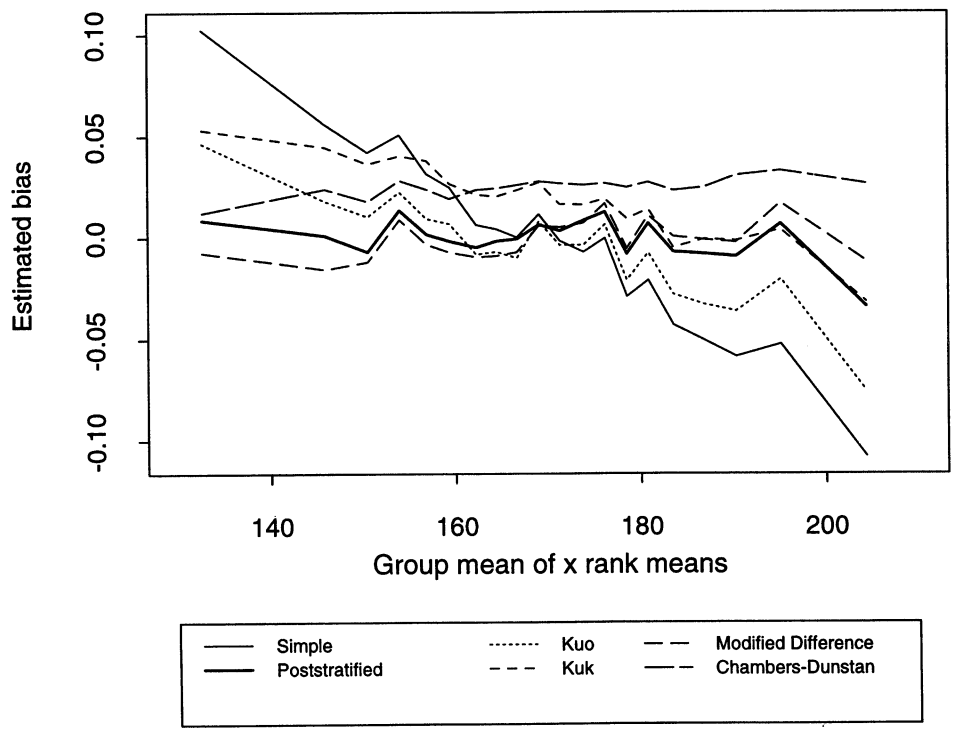

Fig. 3. Estimated conditional bias, $q = 1/4$, sugar cane farms, samples of size 30

Table 4. Ratio of root mean squared error (RMSE) (in %) for poststratification estimator (4E) over simple estimator for selected population quantiles

Quantile	Sugar cane		Beef farms	
	$n = 30$	$n = 50$	$n = 30$	$n = 50$
$t_{.10}$	93.7	93.7	90.0	96.1
$t_{.25}$	84.0	84.0	78.3	80.4
$t_{.50}$	76.1	79.1	73.3	70.4
$t_{.75}$	77.3	76.0	79.7	64.3
$t_{.90}$	96.1	90.9	96.6	90.2

means of population x ranks of the sample x values, following the approach of Chambers et al. (1993).

From Figure 3 we see that both the modified difference and the poststratified estimators display good conditional behaviour. Kuk's estimator presents a downward trend in the conditional bias, although with reasonably small values. Kuo's estimator displays a poor conditional performance, second only to that of the simple estimator, which uses no auxiliary information. The model-based estimator of Chambers and Dunstan presents a small conditional bias with little variation over the whole range of samples. Similar results were obtained for other values of q and for the beef farms population.

Next we briefly consider the estimation of quantiles t_q by inverting the distribution function estimators to give $\hat{F}^{-1}(q)$. For linear estimators, such as the simple and poststratified estimators, this inversion is straightforward. In Table 4 we compared the poststratified estimator with the simple estimator in terms of the ratio of the corresponding root mean squared errors for the quantiles t_q such that $q = 0.10$, 0.25, 0.5, 0.75 and 0.90. The poststratified (4E) quantile estimator provides some useful gains in precision over the simple estimator. We note that other poststratification schemes yielded even better precision for these quantiles.

For the sugar cane population ($n = 30$), these results are similar to those reported by Rao et al. (1990) for their ratio and difference quantile estimators. Other quantile estimators based on inverting nonlinear estimators of the distribution function were not considered here because of greater computational difficulties.

In terms of bias, the poststratified estimator performed better than the simple estimator overall. For the beef farms population, some nonnegligible biases were observed, especially for the tail quantiles. For the sugar cane farms, relative bias was not important, except perhaps for the lower quantile $t_{.10}$.

Finally we also briefly considered variance estimation for the poststratified estimator (scheme 4E), by computing two alternative variance estimates for each sample estimate $\hat{F}_{ps}(t_q)$. Both variance estimators are based on expression (11), the first with $a_i = I(y_i \leqslant t) - \hat{F}_{g(i)}(t)$, which we call v_a, and the second with $a_i = [I(y_i \leqslant t) - \hat{F}_{g(i)}(t)]N_{g(i)}/\hat{N}_{g(i)}$ which we call v_b. Note that further collapsing of poststrata to achieve a minimum of two sample elements was needed in order for the variance estimates to be computed.

Table 5 displays simulation estimates of average absolute bias and average root mean squared error of these variance estimators computed relative to the simulation

Table 5. Average absolute bias (AVAB) and average root mean squared error (AVRMSE) (*) for alternative variance estimators under poststratification scheme 4E

Simulation population and sample size	AVAB		AVRMSE	
	v_a	v_b	v_a	v_b
Sugar cane, $n = 30$	7.4	5.6	12.4	13.9
Sugar cane, $n = 50$	2.6	1.8	4.8	5.2
Beef farms, $n = 30$	7.5	5.9	12.8	14.7
Beef farms, $n = 50$	2.2	1.5	5.1	5.5

(*)Values multiplied by 10,000.

variances of the point estimators. These estimates were obtained by averaging over the 11 variance estimates for the quantiles considered.

The variance estimator v_a performed slightly better than v_b in terms of root mean squared error, although the reverse was true in terms of bias. Some bias was observed for samples of size 30 in both populations, possibly due to the higher rates of collapsed poststrata. Bias was substantially smaller for samples of size 50.

The main finding is that both variance estimators perform satisfactorily. The sizes of their biases relative to their standard deviations decline as n increases. The sizes of their standard deviations correspond roughly to conventional variance estimation. For example, the relative standard deviation of the usual variance estimator of the sample mean in a random sample from a normal population is $\sqrt{2/(n-1)}$ or about 0.20 when $n = 50$. For the AVRMSEs in the sugar cane population the corresponding figure for v_a is $4.8 \times 10,000/435^2 = 0.25$. This would be reduced somewhat if the bias was excluded. The variance estimators are also convenient in practical terms, for being simple to implement, especially under simple random sampling.

5. Conclusions and Discussion

In this paper we have compared a poststratified estimator with several other estimators of the finite population distribution function. The poststratified estimator possesses a number of desirable properties, such as yielding a genuine distribution function, asymptotic unbiasedness, availability of variance estimator, simplicity of computation, etc. Three estimators – the simple, ratio and difference estimators – possess some of these properties, but appear to be inferior in terms of efficiency. This leaves four other estimators that we have considered – Chambers and Dunstan's, the modified difference estimator and both nonparametric estimators of Kuo and Kuk – as potentially serious competitors.

Chambers and Dunstan's estimator can be very efficient when the model upon which it is based is appropriate. However, as noted by Rao et al. (1990), Chambers et al. (1993) and Dorfman (1993), this estimator can perform poorly under model misspecification. It thus appears to be an estimator which may be more suited to applications where detailed model checking can be performed, rather than to routine use in surveys, where a poststratified estimator may be preferable.

Kuk's estimator also performed well, although the results appeared to be somewhat sensitive to the choice of transformation and there were some biases which might be

more important for larger sample sizes. It would be useful to obtain further empirical evidence on whether the gain in efficiency displayed by Kuk's estimator compared to the modified difference estimator holds up in other settings. Some evidence which suggests that this might not be the case is provided by Chambers et al. (1993); they show that some nonparametric estimators, which bear some resemblance to Kuk's perform fairly similar to the modified difference estimator.

Kuo's estimator was clearly outperformed by all the other estimators that use auxiliary information. It seems that such nonparametric estimators may be worthy of further research to clarify the choice of transformations and bandwidth, but they are likely to remain fairly complicated estimators, suited mainly to special applications.

Finally, there is the modified difference estimator of Rao et al. (1990). This estimator seems to offer a fairly consistent slight improvement in efficiency compared to poststratification. For example, poststratification offers a 21% reduction in average root mean squared error for the sugar cane farms with $n = 50$ compared to the simple estimator which uses no auxiliary information (Table 2), whereas the modified difference estimator offers a 26% reduction. Against this slight gain in efficiency stands a number of practical disadvantages. The modified difference estimator is not necessarily monotonic; it requires fairly heavy computation with algorithms which differ from the ones usually adopted in survey estimation and variance estimation can be complicated.

In conclusion, we suggest that in many standard survey settings poststratification and more generally regression estimation provide a simple and practical approach to incorporating auxiliary information into the estimation of distribution functions which can offer some useful gains in efficiency. We found no evidence, for simple random sampling, to indicate that greater efficiency can be obtained by creating the poststrata by a scheme other than that for which equal sample sizes are expected within poststrata. Our (limited) evidence also suggests that little efficiency can be gained by having more than three or four poststrata.

6. References

Casady, R.J. and Valliant, R. (1993). Conditional Properties of Poststratified Estimators Under Normal Theory. Technical report, U.S. Bureau of Labor Statistics.

Chambers, R.L., Dorfman, A.H., and Hall, P. (1992). Properties of Estimators of the Finite Population Distribution Function. Biometrika, 79, 577–582.

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. Journal of the American Statistical Association, 88, 268–277.

Chambers, R.L. and Dunstan, R. (1986). Estimating Distribution Function from Survey Data. Biometrika, 73, 597–604.

Chen, J. and Qin, J. (1993). Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information. Biometrika, 80, 107–116.

Dorfman, A.H. (1993). A Comparison of Design-Based and Model-Based Estimators of the Finite Population Distribution Function. Australian Journal of Statistics, 35, 29–41.

Fuller, W.A. (1966). Estimation Employing Poststrata. Journal of the American Statistical Association, 61, 1172–1183.

Kuk, A.C. (1993). A Kernel Method for Estimating Finite Population Distribution Functions Using Auxiliary Information. Biometrika, 80, 385–392.

Kuo, L. (1988). Classical and Prediction Approaches to Estimating Distribution Functions from Survey Data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 280–285.

Little, R.J.A. (1993). Poststratification: A Modeler's Perspective. Journal of the American Statistical Association, 88, 1001–1012.

Rao, J.N.K. (1985). Conditional Inference in Survey Sampling. Survey Methodology, 11, 15–31.

Rao, J.N.K., Kovar, J.G., and Mantel, H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. Biometrika, 77, 365–375.

Rao, J.N.K. and Liu, J. (1992). On Estimating Distribution Functions from Sample Survey Data Using Supplementary Information at the Estimation Stage. In Nonparametric Statistics and Related Topics (A.K.Md.E. Saleh ed.), Amsterdam: Elsevier Science Publishers, 399–407.

Särndal, C.E., Swensson, B., and Wretman, J. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527–537.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Valliant, R. (1993). Poststratification and Conditional Variance Estimation. Journal of the American Statistical Association, 88, 89–96.

Received November 1993
Revised July 1995