Journal of Official Statistics, Vol. 23, No. 3, 2007, pp. 371–386

Resampling Variance Estimation in Surveys with Missing Data

A.C. Davison[1] *and S. Sardy*[2]

We discuss variance estimation by resampling in surveys in which data are missing. We derive a formula for linearization in the case of calibrated estimation with deterministic regression imputation, and compare the resulting variance estimates with balanced repeated replication with and without grouping, the bootstrap, the block jackknife, and multiple imputation, for simulated data based on the Swiss Household Budget Survey. We also investigate the number of replications needed for reliable variance estimation under resampling in this context. Linearization, the bootstrap, and multiple imputation perform best in terms of relative bias and mean squared error.

Key words: Balanced repeated replication; bootstrap; calibration; influence function; jackknife; linearization; missing data; multiple imputation.

1. Introduction

Classical variance formulae for sample survey estimators are derived using approximations based on Taylor series expansion of the estimators. When the sample is small or the estimator complex—for instance, because of modifications to account for missing data—it is natural to be concerned about the quality of such approximations, and to consider alternatives such as resampling procedures. The purpose of this article is to give formulae for general variance approximations in the presence of calibration and deterministic imputation, and to compare them numerically with resampling procedures.

Section 2 reviews the classes of estimator that we consider, and Section 3 reviews resampling methods for variance estimation. Section 4 outlines a linearization approach for use when missing data are dealt with by calibration and deterministic regression imputation, and Section 5 contains numerical investigations based on the Swiss Household Budget Survey. The article ends with a brief discussion.

2. Basic Ideas

Consider first complete response for a stratified single stage unequal probability sampling scheme without replacement, with N units divided into H strata, from which a total of

[1] Institute of Mathematics, School of Basic Sciences, Ecole Polytechnique Fédérale de Lausanne, Station 8, CH-1015 Lausanne, Switzerland. Email: Anthony.Davison@epfl.ch (http://stat.epfl.ch)
[2] Section de Mathématiques, Université de Genève, 24, rue du Lièvre, Case postale 64, 1211 Genève 4, Switzerland.
Acknowledgments: This work was largely performed in the context of the European Union project DACSEIS (http://www.dacseis.ch). We thank the other members of the DACSEIS team for their valuable collaboration, Professor J. N. K. Rao for insightful comments, and referees for helpful remarks on an earlier version of the article.

n units are sampled. Let n_h be the number of units sampled from the N_h population units in stratum h, and let π_{hi} be the inclusion probability for unit i of this stratum. In household surveys this unit might consist of a cluster of individuals, in which case the unit response of interest is supposed to be cumulated over the cluster. Let x_{hi} and y_{hi} be variables that have been measured on the units, where y_{hi} is the scalar response of interest and x_{hi} is a $q \times 1$ vector of auxiliary variables, which may be continuous, categorical, or both.

Parameters of the finite population can be classified into two broad groups. The first, largest, and most important group comprises smooth quantities such as the population total $\tau = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}$, the ratio, the correlation, or the change in the ratio between two sampling occasions. The other main group comprises nonsmooth functions of the finite population responses, such as the median, quantiles, and statistics based on them (Berger and Skinner 2003).

Estimation of the finite population parameters is based on the data from the n sampled units and on their inclusion probabilities under the given sampling design. The most important estimator of a total is the Horvitz–Thompson estimator

$$\hat{\tau} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} y_{hi} = \omega^{\mathrm{T}} y \tag{1}$$

where $\omega_{hi} = 1/\pi_{hi}$ are the inverse inclusion probabilities. The variance of $\hat{\tau}$ is readily obtained, but complications arise when the weights themselves are random, or when some of the responses are unavailable.

In many cases population totals are known for some of the auxiliary variables x, and this information can be used to increase precision of estimation by a procedure known as calibration. Suppose that q_C marginals of the q auxiliary variables are known, with $q_C \leq q$, let c be the $q_C \times 1$ vector of known marginals, and let X_C denote the $n \times q$ matrix of auxiliary variables whose marginal total for the entire population is known to equal c. Using the estimation of a total to illustrate calibration, the quality of the Horvitz–Thompson estimator can be improved by choosing the weights w_{hi} to be as close as possible to the original weights ω_{hi} in some metric G, subject to the constraint that the weighted auxiliary variables match the marginals (Deville and Särndal 1992), that is,

$$\min_{w_{hi}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} G(w_{hi}/\omega_{hi}) \quad \text{such that} \quad \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{C_{hi}} = c$$

A widely-used distance measure is the ℓ_2 or squared error metric, $G(x) = (x-1)^2/2$, which results in the calibrated weights

$$w = \omega + \Omega X_C \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} \left(c - X_C^{\mathrm{T}} \omega\right) \tag{2}$$

where Ω denotes the diagonal matrix whose elements are the ω_{hi}, and a calibrated Horvitz–Thompson estimator of the form

$$\hat{\tau} = w^{\mathrm{T}} y = \omega^{\mathrm{T}} y + \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} X_C^{\mathrm{T}} \Omega y \quad = \omega^{\mathrm{T}} y + \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \hat{\gamma} \tag{3}$$

where $\hat{\gamma}$ is the regression estimator when y is regressed on X_C with weight matrix Ω.

Other distance measures have been suggested, but are equivalent asymptotically (Deville and Särndal 1992) and in practice (Deville et al. 1993).

In practice survey data sets are rarely complete, but are subject to unit nonresponse, or item nonresponse, or both. Although calibration is mainly used for variance reduction, it may also be used to allow for unit nonresponse. Item nonresponse, where the covariate x is known but the target variable y is missing, demands a different approach, typically through the use of an imputation model that allows missing values of y to be predicting the data available.

A common deterministic approach to imputation of missing response values is to use a (generalized) linear model based on the vectors x_{hi} of auxiliary variables. The normal equations for estimating the parameters β of such imputation models across strata may be written in the vector form

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h} x_{hi}\psi(y_{hi}, x_{hi}; \beta) = 0 \tag{4}$$

where ψ, the derivative of the implied loss function with respect to β, is also sometimes known as an influence function. There is a close connection here with M-estimation, commonly used in robust statistics (Huber 1981; Hampel et al. 1986). If the response y is dichotomous it is natural to use logistic regression as the imputation model, and then the y_{hi} are binary indicator variables and $\psi(y, x; \beta) = y - \exp(x^{\mathrm{T}}\beta)/\{1 + \exp(x^{\mathrm{T}}\beta)\}$. If y is continuous then one simple possibility is ratio imputation using a scalar x, for which we take $\psi(y, x; \beta) = y - \beta x$. For a more robust imputation model, one might use Huber's Proposal 2 (Huber 1981), for which $\psi(u) = \mathrm{sign}(u)\min(|u|, \tau)$; here $\tau > 0$ controls the degree of robustness of the fit, with $\tau \to \infty$ recovering the least squares estimator, and $\tau \to 0$ giving higher robustness. Once the linear model M-estimate $\hat{\beta}$ of β has been found, the missing response for an individual with explanatory variable x can be predicted by $x^{\mathrm{T}}\hat{\beta}$, or by a smooth function of this.

For a linear imputation model, the calibrated and imputed Horvitz–Thompson estimator may be written as

$$\begin{aligned}\hat{\tau} = w^{\mathrm{T}}\{Zy + (I - Z)\hat{y}\} &= \omega^{\mathrm{T}}Zy + \left(c - X_C^{\mathrm{T}}\omega\right)^{\mathrm{T}}\left(X_C^{\mathrm{T}}\Omega X_C\right)^{-1}X_C^{\mathrm{T}}\Omega Zy \\ &\quad + \omega^{\mathrm{T}}(I - Z)X\hat{\beta} + \left(c - X_C^{\mathrm{T}}\omega\right)^{\mathrm{T}}\left(X_C^{\mathrm{T}}\Omega X_C\right)^{-1}X_C^{\mathrm{T}}\Omega(I - Z)X\hat{\beta}\end{aligned} \tag{5}$$

where $Z = \mathrm{diag}(z)$ is the $n \times n$ diagonal matrix of indicator variables z_{hi} corresponding to observed response, X is the $n \times q$ matrix that contains the auxiliary variables corresponding to both respondents and nonrespondents, and $\hat{y} = X\hat{\beta}$ represents the $n \times 1$ vector of fitted values from the regression model used for imputation.

3. Resampling Variance Estimation

Modern sample survey estimators often involve calibration and/or imputation, and variance formulae for them cannot be found in classic texts such as Cochran (1977). The simplest approach would be to treat the imputed responses $\hat{y}$ as if they were true responses, but this can lead to considerable underestimation of the true variance. One way to estimate the variance of estimators such as (5) is through resampling. The adaptation of resampling

methods to the survey setting requires special care, because it must take into account the dependence that may be induced by the sampling scheme as well as the effect of possible calibration and imputation. We now briefly outline the main resampling procedures proposed for sample surveys. A more detailed overview may be found in Davison and Sardy (2007).

The jackknife, originally introduced as a method of bias estimation (Quenouille 1949a,b) and subsequently proposed for variance estimation (Tukey 1958), involves the systematic deletion of groups of units at a time, the recomputation of the statistic with each group deleted in turn, and then the combination of all these recalculated statistics. The simplest jackknife entails the deletion of single observations, but this delete-one jackknife is inconsistent for nonsmooth estimators, such as the median and other estimators based on quantiles (Efron 1982). Shao and Wu (1989) and Shao and Tu (1995) have shown that the inconsistency can be repaired by deleting groups of d observations, where $d = o(n) \to \infty$ as $n \to \infty$. Rao and Shao (1992) describe a consistent version of the delete-one jackknife variance estimator using a particular hot deck imputation mechanism to account for nonresponse; see also Fay (1996) for a wider perspective, and Chen and Shao (2001), who show that this approach fails for another popular imputation scheme, nearest-neighbour imputation.

The bootstrap involves recomputing the statistic, now using resampling from an estimated population $\hat{F}$ to obtain bootstrap samples and the corresponding statistics $\hat{\theta}^*$. Repeating this process R times independently yields bootstrap replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_R^*$ of $\hat{\theta}$, and the bootstrap estimate of variance is $v_B = (R-1)^{-1} \sum_r \left(\hat{\theta}_r^* - \overline{\hat{\theta}^*} \right)^2$, where $\overline{\hat{\theta}^*} = R^{-1} \sum_r \hat{\theta}_r^*$. For stratified data, the resampling is performed separately within each stratum. The usual bootstrap uses sampling with replacement, corresponding to independent sampling from an original population, but this does not match the without-replacement sampling generally used in the survey context, so the finite sampling correction is missed, leading to a biased variance estimator. This failure of the usual bootstrap has spurred a good deal of work on modified bootstraps including the without-replacement bootstrap (Gross 1980; Chao and Lo 1985; Bickel and Freedman 1984; Sitter 1992b; Booth et al. 1994; Booth and Hall 1994), the with-replacement bootstrap (McCarthy and Snowden 1985), the rescaling bootstrap (Rao and Wu 1988), and the mirror-match bootstrap (Sitter 1992a). When responses are missing, the imputation mechanism must be applied to each resample (Shao and Sitter 1996): the idea is to reimpute repeatedly using the respondents of the bootstrapped sample to fit the imputation model and then impute the nonrespondents of the bootstrap sample. This is computer intensive, but it is claimed to give consistent variance estimators for medians and other estimators based on quantiles—though the mathematical property of consistency is no quarantee of good performance in practice.

Balanced half-sampling (McCarthy 1969) is the simplest form of balanced repeated replication. It was originally developed for stratified multistage designs with two primary sampling units drawn with replacement in the first stage. Two main generalizations to surveys with more than $n_h = 2$ observations per stratum have been proposed. The first, investigated by Gurney and Jewett (1975), Gupta and Nigam (1987), Wu (1991) and Sitter (1993), uses orthogonal arrays, but requires a large number of replicates, making it

impractical for many applications. The second generalization, a simpler more pragmatic approach, is to group the primary sampling units in each stratum into two groups, and to apply balanced repeated replication using the groups rather than individual units (Rao and Shao 1996; Wolter 1985, Section 3.7). The balanced repeated replication variance estimator v_{BRR} can be highly variable, and a solution to this suggested by Robert Fay of the U.S. Bureau of the Census (Dippo et al. 1984; Fay 1989) is to use a milder reweighting scheme. Another solution (Rao and Shao 1996) is to repeat the method over differently randomly selected groups to provide several estimates of variance, averaging of which will provide a more stable overall variance estimate. Shao et al. (1998) adjust balanced repeated replication to the presence of nonresponse, by taking into account a deterministic or random imputation mechanism. Under a general stratified multistage sampling design, they establish consistency of the adjusted balanced repeated replication variance estimators for functions of smooth and nonsmooth statistics.

Multiple imputation (Rubin 1987, 1996; Little and Rubin 2002) has also been promoted for variance estimation in complex surveys (Münnich and Rässler 2005)—standard formulae are computed for several datasets for which missing data have been stochastically imputed, and are then combined in such a way as to make proper allowance for the effect of imputation. This approach has been regarded as controversial by certain authors; see, for example Fay (1996).

4. Linearization

Linearization is a general term for the construction of variance estimators based on a linear series expansion of the estimator of interest in terms of underlying quantities; it is sometimes called the delta method. The best-known approach of this type involves Taylor series expansion of the estimator in terms of means or totals around their population values, and the resulting variance is then estimated by replacing population variances and covariances by unbiased estimators based on the sample. Taylor linearization may be applied to estimators that may be written as smooth functions of means, and so is useful for most common survey estimators (Deville 1999).

A more general approach is through a functional expansion of the estimator, sometimes called a von Mises expansion, in which the averages that appear in Taylor series expansion are replaced by sums of directional derivatives that involve individual observations. Depending on the mathematical formalism used, these are Fréchet or Gâteaux derivatives. Although it yields the same results as Taylor series expansion when applied to averages, von Mises expansion may also be applied to quantities not directly expressible as averages, such as sample quantiles. More details may be found in Fernholtz (1983), Hampel et al. (1986), or Davison and Hinkley (1997, §2.7), and Campbell (1980) discusses application to finite population sampling. Here we adopt the functional approach, which deserves to be more widely known.

The main tool in construction of linearization variance estimators using the functional approach is the influence function, whose derivation we now outline. In many cases the estimand θ can be written as a functional $t(F)$ of the underlying distribution function F from which observations $Y_1, \ldots, Y_n$ are supposed independently drawn. A simple estimator of $t(F)$ is then $t(\hat{F})$, where $\hat{F}$ is the empirical distribution function of the data.

For the mean, for instance, $t(F) = \int u dF(u)$ and $t(\hat{F}) = \overline{Y}$ is its empirical analogue. Under some differentiability properties for $t(\cdot)$, the estimate $\hat{\theta} = t(\hat{F})$ can be expanded around $\theta = t(F)$ as $t(\hat{F}) \doteq t(F) + n^{-1}\sum_{i=1}^{n} L_t(Y_i; F)$, where

$$L_t(y; F) = \lim_{\epsilon \to 0} \frac{t\{(1-\epsilon)F + \epsilon\delta_y\} - t(F)}{\epsilon} \tag{6}$$

is the *influence function* for $t(\hat{F})$, δ_y being the distribution function putting a point mass at y. This expansion can be used to establish that the estimator is asymptotically unbiased and Gaussian. Its variance $v_L(F) = n^{-1}\mathrm{var}\{L_t(Y; F)\}$ can be estimated from a sample $y_1, \ldots, y_n$ by

$$\hat{v}_L = n^{-2}\sum_{i=1}^{n} l_i^2 \tag{7}$$

where $l_i = L_t(y_i; \hat{F})$ are the *empirical influence values* for the statistical functional t evaluated at y_i and $\hat{F}$. Here l_i can be thought of as the derivative of t at $\hat{F}$ in the direction of a distribution putting more mass on the ith observation.

There is a close relationship with the jackknife, which can be regarded as providing a numerical approximation to the empirical influence values. Yung and Rao (1996, 2000) have exploited this to produce analytical approximations to jackknife variance estimators that they call jackknife linearization variance estimators; they study theoretical properties of their estimators under various imputation schemes and provide some numerical results.

For stratified random sampling without replacement (7) may be modified to

$$v_L = \sum_{h=1}^{H} (1 - f_h)\frac{1}{(n_h - 1)n_h}\sum_{i=1}^{n_h} l_{hi}^2 \tag{8}$$

where l_{hi} is the empirical influence value corresponding to the ith observation in stratum h.

We now consider the Horvitz–Thompson estimator and give formulae for its empirical influence functions for stratified sampling in three situations of increasing complexity:

- the standard estimator (1), for which

$$l_{hi} = n_h \omega_{hi} y_{hi} - \omega_h^{\mathrm{T}} y h$$

- the calibrated estimator (3), for which (Canty and Davison 1999)

$$\begin{aligned} l_{hi} =& (n_h \omega_{hi} y_{hi} - \omega_h^{\mathrm{T}} y h) + \left(X_{C_h}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{C_{hi}}\right)^{\mathrm{T}} \hat{\gamma} \\ &+ n_h \omega_{hi} \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} x_{C_{hi}} \left(y_{hi} - x_{C\,hi}^{\mathrm{T}} \hat{\gamma}\right) \end{aligned} \tag{9}$$

  where ω_h and y_h are $n_h \times 1$ vectors of the weights and responses for the h-th stratum, X_{Ch} is the $n_h \times q_C$ matrix of calibration covariates for the h-th stratum, and $\hat{\gamma} = \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} X_C^{\mathrm{T}} \Omega y$; and
- the calibrated estimator (5) with imputation of missing responses. Let

$$\hat{\gamma}_M = \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} X_C^{\mathrm{T}} \Omega (I - Z)\hat{y}$$

correspond to $\hat{\gamma}$, but for those individuals with missing responses, and let $l_i(\hat{\beta})$ be the elements of the $q \times 1$ vector of influence functions for the imputation regression coefficients, corresponding to differentiation with respect to the ith case in stratum h. Then calculations along the lines of those in Rust and Rao (1996) or Canty and Davison (1999) and sketched in the Appendix yield

$$\begin{aligned} l_{hi} =& \left(n_h \omega_{hi} z_{hi} y_{hi} - \omega_h^{\mathrm{T}} Z_h y_h\right) + \left(X_{C_h}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{C_{hi}}\right)^{\mathrm{T}} \hat{\gamma} \\ &+ n_h \omega_{hi} \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} x_{C_{hi}} \left(z_{hi} y_{hi} - x_{C_{hi}}^{\mathrm{T}} \hat{\gamma}\right) \\ &+ \left\{n_h \omega_{hi} (1 - z_{hi}) \hat{y}_{hi} - \omega_h^{\mathrm{T}} (I_h - Z_h) \hat{y}_h\right\} + \omega^{\mathrm{T}} (I - Z) X l_i(\hat{\beta}) \\ &+ \left(X_{C_h}^{\mathrm{T}} \omega_h - n_h \omega_{hi} x_{C_{hi}}\right)^{\mathrm{T}} \hat{\gamma}_M \\ &+ n_h \omega_{hi} \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} x_{C_{hi}} \left\{(1 - z_{hi}) \tilde{y}_{hi} - x_{C_{hi}}^{\mathrm{T}} \hat{\gamma}_M\right\} \\ &+ \left(c - X_C^{\mathrm{T}} \omega\right)^{\mathrm{T}} \left(X_C^{\mathrm{T}} \Omega X_C\right)^{-1} X_{C_h}^{\mathrm{T}} \Omega_h (I_h - Z_h) X_h l_i(\hat{\beta}) \end{aligned} \tag{10}$$

In particular, use of a linear model fitted by least squares for deterministic imputation yields

$$l_i(\hat{\beta}) = n_h z_i (X^{\mathrm{T}} Z X)^{-1} x_i (y_i - x_i^{\mathrm{T}} (X^{\mathrm{T}} Z X)^{-1} X^{\mathrm{T}} Z_y), \quad i = 1, \ldots, \sum_{h=1}^{H} n_h$$

where X is the regression matrix. When the regression coefficients vary among the strata, then the $l_i(\hat{\beta})$ in (10) are taken to be

$$l_i(\hat{\beta}_h) = n_h z_i (X_h^{\mathrm{T}} Z_h X_h)^{-1} x_i \left(y_i - x_i^{\mathrm{T}} \left(X_h^{\mathrm{T}} Z_h X_h\right)^{-1} X_h^{\mathrm{T}} Z_h y_h\right)$$

where X_h, Z_h, and y_h are the covariate matrix, the indicator matrix for observed responses, and the response vector for stratum h.

The advantages of these formulae over resampling techniques are a reduction in computational effort and the possibility of handling massive surveys. However, such formulae entail further assumptions whose validity needs careful investigation in applications: that the models underlying the calibration and imputation schemes and leading to estimators such as (3) and (5) do not introduce serious bias—if so, then (8) may provide an inadequate idea of the uncertainty of the estimator.

5. Numerical Comparison

5.1. Simulation Study

Using a realistic simulation based on the 1998 Swiss Household Budget Survey (Renfer 2001), we consider the calibrated and imputed Horvitz–Thompson estimator of the total expenditure on bread and cereal products, based on complete data from $N = 9,275$ households in $H = 7$ strata of various sizes. Also available on each household is a set of 14

auxiliary variables, of which 10 population margins are known. For the simulation, we consider the $N = 9,275$ households as the whole population, for which we assume we know the total expenditure. We perform stratified random sampling without replacement and with equal inclusion probabilities of 1/8 within 6 strata, and 3/8 in the other stratum, giving a sample size of 1332. Item nonresponse for the response variable is applied using a uniform probability of missingness across the entire sample. On each of the 500 samples simulated, we calculate the calibrated and imputed Horvitz–Thompson estimates, with deterministic imputation performed using linear regression, and use resampling techniques to obtain variances for them. The true variances were computed using estimates from 10,000 samples from the population, with the sampling and nonresponse schemes outlined above.

The bootstrap of the calibrated and imputed Horwitz–Thompson estimator involved the procedure of Shao and Sitter (1996): missing responses were imputed deterministically using a linear model fitted to the bootstrapped full respondents, and with the imputed dataset calibrated to the weights by linear regression. The sampling fraction here is sufficiently small to apply a standard, with-replacement, bootstrap. In order to keep the computational effort to a reasonable level, we took $R = 100$ bootstrap replicates, see Section 5.2. To match the computational burden of the bootstrap, we used roughly the same number of block deletions when applying the block jackknife with replacement. This was applied with 13 randomly selected blocks in each stratum, leading to about 91 computations in all for each jackknife variance estimate. Two forms of balanced repeated replication were applied, the first using a single random split of each stratum into two halves for each replication; no Fay factor was used but the weights for those observations included in the replicate were multiplied by a factor of two before calibration. The second form, repeatedly grouped balanced repeated replication, averages over variance estimates from 13 such splits. The linearization estimators were those given by (8) and (10).

The standard formulae for multiple imputation were applied, using 30 random imputations from a linear model fitted to the complete data; for parametric imputation we used a homoscedastic normal error model, with the values of the regression parameters and variance changing randomly and independently according to the fitted normal and chi-squared distributions between simulations; for nonparametric imputation errors were simulated according to a model-based residual bootstrap (Davison and Hinkley 1997, p. 262).

Table 1 and Figures 1 and 2 compare the performances of these variance estimation techniques for missingness rates of 0%, 20%, 40%, and 60%. Linearization and both multiple imputation methods give the same results when no data are missing. The block jackknife underestimates the true variances, which are systematically overestimated by repeatedly grouped balanced repeated replication. Ungrouped balanced repeated replication is highly variable by comparison, in agreement with results of Rao and Shao (1996), but grouping reduces its variance appreciably. Linearization works well for low levels of missingness, and overall produces variances that are slightly low but quite stable. For higher levels of missingness the bootstrap performs best. Nonparametric multiple imputation also performs well.

Figure 2 shows how the variance estimates for the 500 simulated data sets are correlated with the bootstrap variance estimates. Linearization, repeatedly grouped balanced

Table 1. Relative bias and root mean squared error (%) for the different resampling plans applied to simulated data based on the 1998 Swiss Household Budget Survey, for different proportions of missing data. The relative bias and root mean squared error are obtained by dividing the bias and root mean squared error by the target 'true' variance of the estimator

Proportion missing (%)	Relative bias (%)				Relative RMSE (%)			
	0	20	40	60	0	20	40	60
Block jackknife	−10	−11	−12	−15	15	15	17	19
Balanced repeated replication (BRR)	3	6	2	4	31	34	30	32
Randomly grouped BRR	8	8	8	7	14	14	14	15
Bootstrap	7	6	4	1	12	12	12	12
Linearization	−0.3	−2	−5	−9	6	7	9	13
Multiple imputation, parametric	−0.3	1	−2	−14	6	8	9	17
Multiple imputation, nonparametric	−0.3	12	16	9	6	14	18	13

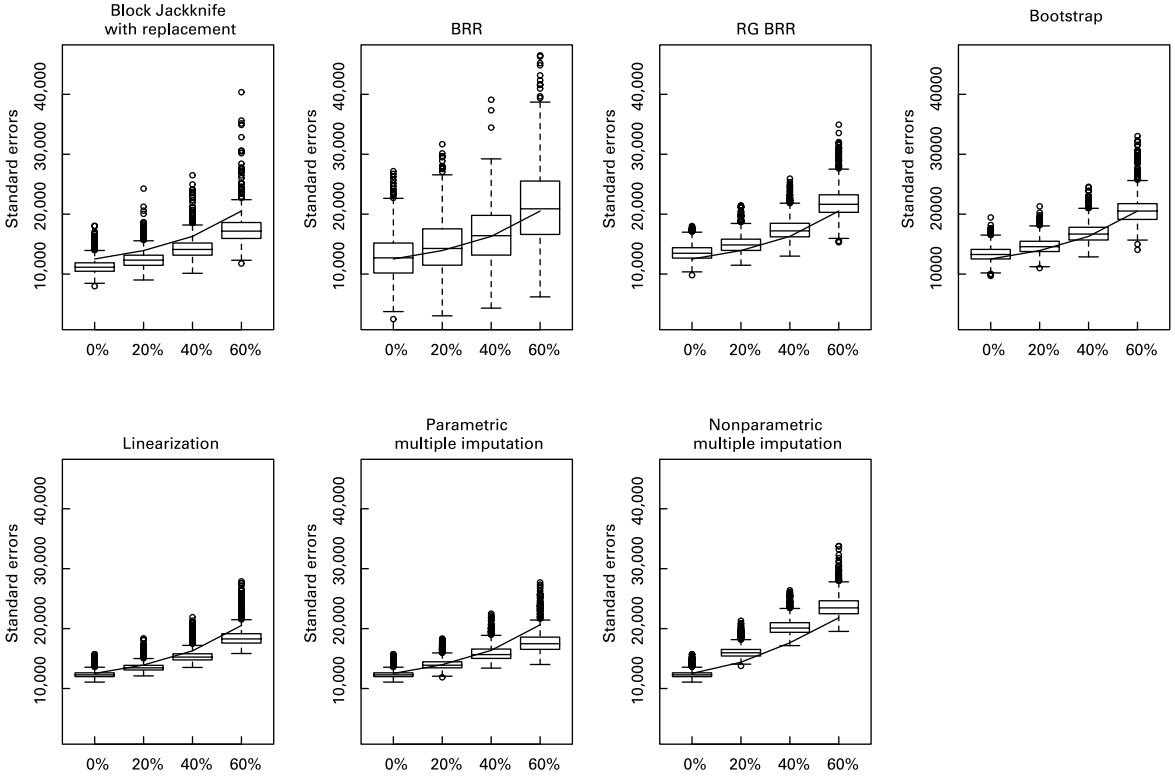

Fig. 1. Comparison of resampling estimators of variance in the presence of calibration and imputation, as a function of the proportion of missing data. Simulation based on the 1998 Swiss Household Budget Survey. The solid line shows the true variances, estimated from 10,000 simulations, and the boxplots show the variance estimates computed for 500 samples. RG and BRR indicate repeatedly grouped and balanced repeated replication, respectively

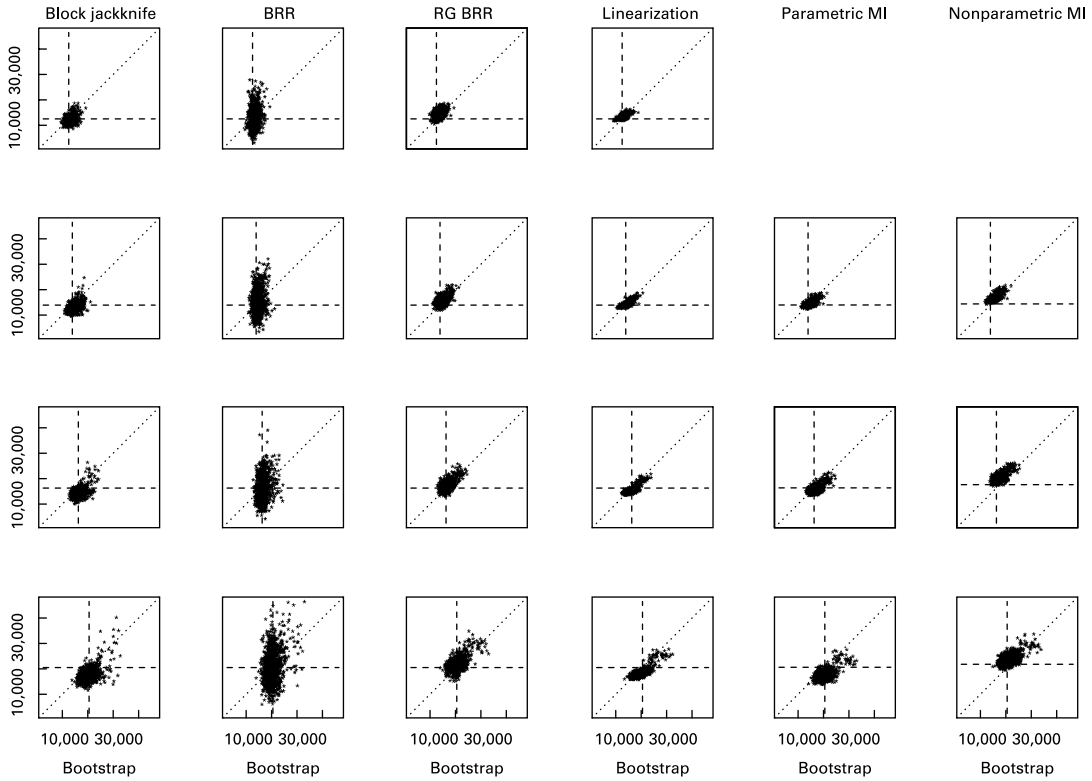

Fig. 2. Comparison of resampling standard errors in the presence of calibration and imputation, as a function of the proportion of missing data; from top to bottom 0%, 20%, 40%, 60% item nonresponse. The dashed lines are the 'true' sampling standard errors, and the dotted line shows $x = y$. Simulation based on the 1998 Swiss Household Budget Survey. RG, BRR and MI indicate repeatedly grouped, balanced repeated replication and multiple imputation, respectively

repeated replication and multiple imputation variance estimates are fairly closely correlated with bootstrap variance estimates. The added variability of the variances from balanced repeated replication shows clearly.

Overall the bootstrap approach of Shao and Sitter (1996), the linearization method of Section 4, and nonparametric multiple imputation seem best in terms of bias and stability. As far as computation time is concerned, the advantage goes to linearization, which is up to fifty times faster than the other methods included in the study.

5.2. *Number of Resamples*

The use of resampling entails choosing the number R of resamples. The comparisons described in Section 5.1 were made for roughly equal computational effort, with around $R = 100$ replications for each method. However, they tend to overestimate the variances, which consist of components of variation between samples and between resamples.

More explicitly, consider the bootstrap estimate of variance v_B for an estimator $\hat{\theta}$. As $R \to \infty$, we have $v_B \to \sigma_B^2(Y)$, say, where $\sigma_B^2(Y)$ is the 'ideal' but unattainable variance that would be obtained from an infinite bootstrap. If we suppose that v_B has approximately a scaled χ_R^2 distribution, as would be the case in large samples, and let $\mathrm{E}^*(\cdot)$ and $\mathrm{var}^*(\cdot)$ denote expectation and variance over possible bootstrap resamples, conditional on the underlying sample Y, then $\mathrm{E}^*(v_B) = \sigma_B^2(Y)$ and $\mathrm{var}^*(v_B) \doteq 2\sigma_B^4(Y)/R$. Hence $\mathrm{var}(v_B) \doteq \mathrm{var}\{\sigma_B^2(Y)\} + 2\mathrm{E}\{\sigma_B^4(Y)\}/R$, where $\mathrm{E}(\cdot)$ and $\mathrm{var}(\cdot)$ are taken with respect to the distribution of samples Y. A similar argument should hold for other resampling variance estimators, such as the grouped jackknife or balanced repeated replication.

Figures 1 and 2 show this variability based on $R = 100$, but the question arises whether the results might be different with larger R. To investigate this we performed a further simulation without missing data, and computed the variances for a variety of values of R,

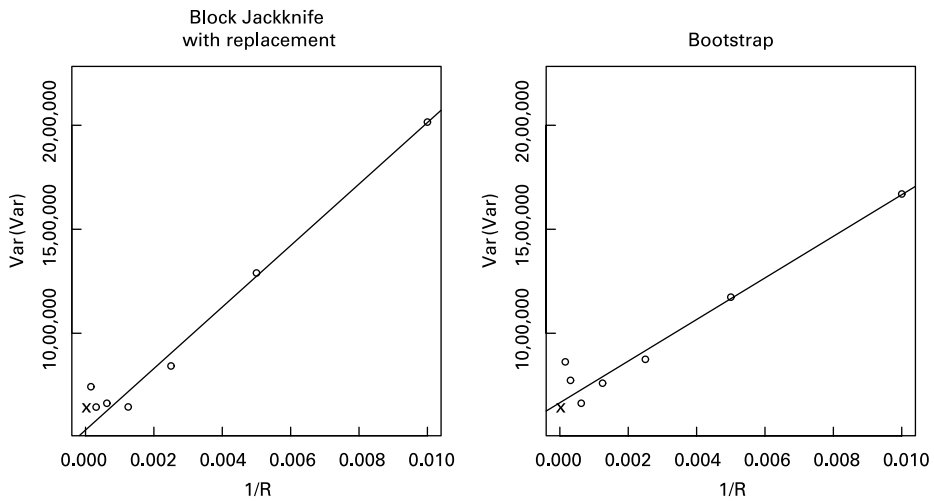

Fig. 3. Variance of bootstrap and block jackknife variance estimators as functions of the inverse number of resamples R^{-1}. The × shows the estimated variance of the linearisation variance estimator across samples, and the lines show weighted linear regression fits

for the bootstrap and the grouped jackknife, which seem to be the most promising of the resampling methods. The results, plotted in Figure 3, show the anticipated linear reduction in $\mathrm{var}(v_B)$ with R^{-1}. The relative advantage of the bootstrap is retained for values of $R \leq 400$, beyond which the grouped jackknife appears less variable. For very large R, the bootstrap variance estimator behaves similarly to the linearisation estimator, while the grouped jackknife varies by less, coherent with the downward bias it exhibits in Figure 1.

The number of resamples used to estimate a variance in practice is typically of the order of a few hundred, so this small study suggests that the bootstrap variance estimator is to be preferred to the grouped jackknife, on grounds of its lower variability for given computational effort and its lower bias.

6. Discussion

The broad conclusions of the numerical study above support those of Canty and Davison (1999), who concluded that linearization and the bootstrap were the simplest and most accurate methods of variance estimation in their study. They did not consider imputation, but found similar conclusions for a variety of smooth estimators and for differences of them between two sampling occasions. It seems reasonable to suppose that the general results seen above would also extend to a broader context.

Both calibration and imputation entail assumptions about how missing responses are related to known explanatory variables. It is interesting to note that the relatively simple form of calibration and the linear imputation scheme adopted in Section 5.1 seem to work fairly well, in the sense of producing fairly unbiased variance estimators, even when large proportions of the data are missing. This suggests that even simple approaches to allowing for missing data, coupled to appropriate resampling schemes, may produce reasonable results, though it would be essential to have a clearer idea when this conclusion holds and reliable diagnostics of its failure.

Appendix

For a linear functional $t(F) = \int a(u)dF(u)$, the definition of the influence function (6) yields $L_t(y; F) = t(\delta_y) - t(F) = a(y) - \int a(u)dF(u)$, with corresponding empirical influence values $l_i = L_t(y_i, \hat{F}) = a(y_i) - n^{-1}\sum a(y_j)$. The computation of the empirical influence values for (1), (3) and (5) uses the fact that these estimators are linear or are products of (almost) linear estimators. For example, the empirical influence function for $\sum_{i=1}^n \omega_i y_i = n \times n^{-1} \sum_{i=1}^n \omega_i y_i$ is

$$l_i = n\omega_i y_i - \sum_{j=1}^{n} \omega_j y_j$$

For a parameter $t(F)$ determined by the estimating equation

$$\int \psi\{u; t(F)\} dF(u) = 0$$

for all F in a suitable space of distributions, we see on replacing F by $(1 - \varepsilon)F + \varepsilon\delta_y$, differentiating with respect to ε, and setting $\varepsilon = 0$, that the influence function $L_t(v; F) =$

$\psi\{v; t(F)\} / \int -\psi_t\{u; t(F)\} dF(u)$, where ψ_t represents the derivative of $\psi(\cdot, \cdot)$ with respect to its second argument. Hence the empirical influence function for $\hat{\beta}$, a solution to the estimating equation $\sum_{j=1}^{n} \psi(y_j, x_j; \beta) = 0$, may be written as

$$l_i = \left\{ -n^{-1} \sum_{j=1}^{n} \frac{\partial \psi(y_j, x_j; \hat{\beta})}{\partial \beta^{\mathrm{T}}} \right\}^{-1} \psi(y_i, x_i; \hat{\beta})$$

For example, the defining equation for a least squares estimator may be written

$$X^{\mathrm{T}}(y - X\beta) = \sum_j x_j \left(y_j - x_j^{\mathrm{T}} \beta \right) = 0$$

so $\psi(y_i, x_i; \beta) = x_i \left(y_i - x_i^{\mathrm{T}} \beta \right)$, where x_i^{T} is the ith row of the regression matrix X, and so $l_i = n(X^{\mathrm{T}}X)^{-1} x_i \left(y_i - x_i^{\mathrm{T}} \hat{\beta} \right)$.

The derivation of empirical influence values (9) and (10) for the estimators (3) and (5) is a messy but straightforward application of the formulae above and the rule for differentation of a product.

7. References

Berger, Y.G. and Skinner, C.J. (2003). Variance Estimation for a Low Income Proportion. Applied Statistics, 52, 457–468.

Bickel, P.J. and Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. Annals of Statistics, 12, 470–482.

Booth, J.G., Butler, R.W., and Hall, P. (1994). Bootstrap Methods for Finite Populations. Journal of the American Statistical Association, 89, 1282–1289.

Booth, J.G. and Hall, P. (1994). Monte Carlo Approximation and the Iterated Bootstrap. Biometrika, 81, 331–340.

Campbell, C. (1980). A Different View of Finite Population Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 319–324.

Canty, A.J. and Davison, A.C. (1999). Resampling-based Variance Estimation for Labour Force Surveys. The Statistician, 48, 379–391.

Chao, M.T. and Lo, S.H. (1985). A Bootstrap Method for Finite Populations. Sankyā A, 47, 399–405.

Chen, J. and Shao, J. (2001). Jackknife Variance Estimation for Nearest-neighbor Imputation. Journal of the American Statistical Association, 96, 260–269.

Cochran, W.G. (1977). Sampling Techniques. Third edition. New York: Wiley.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge: Cambridge University Press.

Davison, A.C. and Sardy, S. (2007). Méthodes de rééchantillonnage pour l'estimation de variance en sondage. Journal de la Société Française de Statistique, 147, 3–32. [In French]

Deville, J.C. (1999). Variance Estimation for Complex Statistics, and Estimators: Linearization and Residual Techniques. Survey Methodology, 25, 193–203.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Proceedings of the American Statistical Association, Section on Survey Research Methods, 489–494.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Philadephia: SIAM.

Fay, R.E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. Proceedings of the American Statistical Association, Social Statistics Section, 212–217.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91, 490–498.

Fernholtz, L.T. (1983). Von Mises Calculus for Statistical Functionals. Lecture Notes in Statistics, Vol. 19. New York: Springer.

Gross, S. (1980). Median Estimation in Sample Surveys. Proceedings of the American Statistical Association, Section on Survey Research Methods, 181–184.

Gupta, V.K. and Nigam, A.K. (1987). Mixed Orthogonal Arrays for Variance Estimation with Unequal Numbers of Primary Selections per Stratum. Biometrika, 74, 735–742.

Gurney, M. and Jewett, R.S. (1975). Constructing Orthogonal Replications for Standard Errors. Journal of the American Statistical Association, 70, 819–821.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. New York: Wiley.

Huber, P.J. (1981). Robust Statistics. New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. Second edition. New York: Wiley.

McCarthy, P.J. (1969). Pseudo-replication: Half samples. Review of the Interational Statistical Institute, 37, 239–264.

McCarthy, P.J. and Snowden, C.B. (1985). The Bootstrap and Finite Population Sampling. Vital and Health Statistics, 2, 2–95.

Münnich, R. and Rässler, S. (2005). PRIMA: A New Multiple Imputation Procedure for Binary Variables. Journal of Official Statistics, 21, 325–341.

Quenouille, M.H. (1949a). Approximate Tests of Correlation in Time-Series. Journal of the Royal Statistical Society, Series B, 11, 68–84.

Quenouille, M.H. (1949b). Notes on Bias in Estimation. Biometrika, 43, 353–360.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. Biometrika, 79, 811–822.

Rao, J.N.K. and Shao, J. (1996). On Balanced Half-sample Variance Estimation in Stratified Random Sampling. Journal of the American Statistical Association, 91, 343–348.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241.

Renfer, J.-P. (2001). Description and Process of the Household and Budget Survey of 1998 (HBS 1998). Swiss Federal Statistical Office. 1–19.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. (1996). Multiple Imputation After 18 + Years (with Discussion). Journal of the American Statistical Association, 91, 473–520.

Rust, K.F. and Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. Statistical Methods in Medical Research, 5, 283–310.

Shao, J., Chen, Y., and Chen, Y. (1998). Balanced Repeated Replication for Stratified Multistage Survey Data Under Imputation. Journal of the American Statistical Association, 93, 819–831.

Shao, J. and Sitter, R.R. (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278–1288.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag.

Shao, J. and Wu, C.F.J. (1989). A General Theory for Jackknife Variance Estimation. Annals of Statistics, 17, 1176–1197.

Sitter, R.R. (1992a). A Resampling Procedure for Complex Survey Data. Journal of the American Statistical Association, 87, 755–765.

Sitter, R.R. (1992b). Comparing Three Bootstrap Methods for Survey Data. Canadian Journal of Statistics, 20, 135–154.

Sitter, R.R. (1993). Balanced Repeated Replications Based on Orthogonal Multi-arrays. Biometrika, 80, 211–221.

Tukey, J.W. (1958). Bias and Confidence in Not Quite Large Samples (Abstract). Annals of Mathematical Statistics, 29, 614.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Wu, C.F.J. (1991). Balanced Repeated Replications Based on Mixed Orthogonal Arrays. Biometrika, 78, 181–188.

Yung, W. and Rao, J.N.K. (1996). Jackknife Linearization Variance Estimators Under Stratified Multistage Sampling. Survey Methodology, 22, 23–31.

Yung, W. and Rao, J.N.K. (2000). Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information. Journal of the American Statistical Association, 95, 903–915.

Received February 2006
Revised March 2007