Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Linearization variance estimation for generalized raking estimators in the presence of nonresponse

by Julia D'Arrigo and Chris Skinner

December 2010

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 181-192
Statistics Canada, Catalogue No. 12-001-X

Linearization variance estimation for generalized raking estimators in the presence of nonresponse

Julia D'Arrigo and Chris Skinner [1]

Abstract

Alternative forms of linearization variance estimators for generalized raking estimators are defined via different choices of the weights applied (a) to residuals and (b) to the estimated regression coefficients used in calculating the residuals. Some theory is presented for three forms of generalized raking estimator, the classical raking ratio estimator, the 'maximum likelihood' raking estimator and the generalized regression estimator, and for associated linearization variance estimators. A simulation study is undertaken, based upon a labour force survey and an income and expenditure survey. Properties of the estimators are assessed with respect to both sampling and nonresponse. The study displays little difference between the properties of the alternative raking estimators for a given sampling scheme and nonresponse model. Amongst the variance estimators, the approach which weights residuals by the design weight can be severely biased in the presence of nonresponse. The approach which weights residuals by the calibrated weight tends to display much less bias. Varying the choice of the weights used to construct the regression coefficients has little impact.

Key Words: Calibration; Nonresponse; Raking; Variance estimation; Weight.

1. Introduction

Survey weighting is widely used to adjust for nonresponse bias. Generalized raking estimation (Deville, Särndal and Sautory 1993) provides a class of weighting methods which may be used when population totals of auxiliary variables are available. These methods can, in principle, remove (large-sample) nonresponse bias when the probability of nonresponse is related to the values of the auxiliary variables via a generalized linear model.

This paper presents some theory for linearization variance estimation for such methods in the presence of nonresponse. It also reports a simulation study of the properties of alternative raking estimators and associated variance estimators in settings designed to mimic two European surveys conducted by national statistical institutes. We consider three forms of raking estimator: the classical raking ratio estimator, the 'maximum likelihood' raking estimator (Brackstone and Rao 1979; Fuller 2002) and the generalized regression estimator (GREG). The first estimator has been used in practice in the British Labour Force Survey (LFS), the first survey upon which our simulation study is based. A version of the second estimator has been used in practice in the German Survey of Income and Expenditure (SIE), the second survey upon which our simulation study is based. The GREG estimator is widely used in many surveys, in particular in the context of nonresponse (Särndal and Lundström 2005).

A number of weighting methods, which do not fall into the class of generalized raking methods considered here, have also been proposed. See Särndal and Lundström (2005) for a historical account and Kott (2006) and Chang and Kott (2008) for some recent developments where the auxiliary variables for which population-level information is available may differ from those variables which are used as covariates in the generalized linear model for the probability of nonresponse.

The primary focus of this paper is on variance estimation and specifically on linearization methods, for which there exist a number of slightly different forms of variance estimator in the literature. In our simulation study we shall compare the properties of alternative raking estimators and associated variance estimators with respect to the effects of both sampling and nonresponse. A previous simulation study by Stukel, Hidiroglou and Särndal (1996) found little difference between two forms of linearization estimator with respect to sampling. However, there are reasons why nonresponse may lead to greater differences. Conditions for unbiasedness of raking estimation methods under nonresponse models vary between estimation methods (*e.g.*, Kalton and Maligalig 1991; Kalton and Flores-Cervantes 2003) and the choice of variance estimator may be more important in the presence of nonresponse (*e.g.*, Fuller 2002, Section 8).

The paper is structured as follows. The generalized raking estimators are defined in section 2 and, after introducing an asymptotic framework, the bias of these estimators is considered in section 3. Linearization variance estimators are defined in section 4. The simulation study is presented in section 5, the results are discussed in section 6 and some concluding remarks are given in section 7.

2. Generalized raking estimation

We consider the class of weighted estimators of a population total $T_y = \sum_U y_i$, which may be expressed as

1. Julia D'Arrigo and Chris Skinner, University of Southampton. E-mail: C.J.Skinner@soton.ac.uk.

$\hat{T}_y = \sum_s w_i\, y_i,$ where y_i is the value of a survey variable for a unit i in a sample s from a population U and w_i is the *survey weight* which may depend on the sample but not on the choice of survey variable. We suppose here that the sample s consists of the set of respondents remaining after sampling and possible unit nonresponse. Generalized raking is a form of weighted estimation which may be employed when auxiliary population information is available in the form of a vector $T_x = \sum_U x_i$ of population totals of values x_i of a vector of auxiliary variables, where x_i is known for all units in s. Following Deville and Särndal (1992), the weights w_i are said to be *calibrated* if they satisfy the *calibration equations* $\sum_s w_i x_i = T_x$. The vector T_x is referred to as the vector of *calibration totals*. The class of generalized raking weights w_i is obtained by minimising the objective function:

$$\sum_s d_i G\,(w_i / d_i), \qquad (2.1)$$

subject to the weights w_i being calibrated, where $G(.)$ is a specified objective function which meets certain criteria (see Deville *et al.* 1993) and d_i is an initial weight. We shall take this to be the design weight, *i.e.*, $d_i = \pi_i^{-1}$, where π_i is the probability that unit i is sampled. Deville and Särndal (1992) show that (subject to $G(.)$ obeying certain conditions), the solution of the above constrained optimisation problem may be expressed as:

$$w_i = d_i\, F\,(x_i' \hat{\lambda}), \qquad (2.2)$$

where $F\,(u) = g^{-1}(u)$ denotes the inverse function of $g\,(u) = d\,G(u)/du$ and $\hat{\lambda}$ is the Lagrange multiplier which solves the calibration equations:

$$\sum_s d_i\, F\,(x_i' \hat{\lambda})\, x_i = T_x. \qquad (2.3)$$

Deville and Särndal (1992) discuss various choices of the $G(.)$ function and associated $F(.)$ function. We consider the following three choices:

linear:

$$G_L(u) = (1/2)(u-1)^2,\ F_L(u) = 1 + u;$$

multiplicative (*raking ratio*):

$$G_M(u) = u \log(u) - u + 1,\ F_M(u) = \exp(u);$$

maximum likelihood raking:

$$G_{ML}(u) = u - 1 - \log(u),\ F_{ML}(u) = (1-u)^{-1}.$$

See also Deville *et al.* (1993) and Fuller (2009, section 2.9) regarding the above terminology for these functions. With the linear choice of $G(.)$, the optimisation problem has a closed form solution and the generalized raking estimator becomes $\hat{T}_y = \hat{T}_{yd} + (T_x - \hat{T}_{xd})' \hat{B}_s,$ the *generalised regression estimator* (GREG), where $\hat{T}_{yd} = \sum_s d_i\, y_i,\ \hat{T}_{xd} = \sum_s d_i\, x_i$ and

$$\hat{B}_s = \left(\sum_s d_i\, x_i\, x_i'\right)^{-1} \sum_s d_i\, x_i\, y_i. \qquad (2.4)$$

With the multiplicative choice of $G(.)$, the calibrated estimator of T_y is the classical raking ratio estimator (Brackstone and Rao 1979) when T_x contains the population counts in the categories of two or more categorical auxiliary variables. For example, in the context of the Britain Labour Force Survey, x_i denotes the vector of indicator variables of three categorical auxiliary variables: $x_i = (\delta_{1..i}, \ldots, \delta_{A..i}, \delta_{.1.i}, \ldots, \delta_{.B.i}, \delta_{..1i}, \ldots, \delta_{..Ci})'$, where $\delta_{a..i} = 1$ if unit i is in category a of the first auxiliary variable and 0 otherwise, $\delta_{.b.i} = 1$ if unit i is in category b of the second auxiliary variable and 0 otherwise and so on. The population total T_x of this vector thus contains the population counts in each of the (marginal) categories of each of the three auxiliary variables. The construction of the weights for classical raking ratio estimation has traditionally involved the use of iterative proportional fitting (Brackstone and Rao 1979). Ireland and Kullback (1968) demonstrate that this method converges to a solution of the above optimisation problem.

The function $G_{ML}(u)$ leads to an alternative 'maximum likelihood' version of raking adjustment, when x_i takes the same form, denoting indicator variables of categorical auxiliary variables. In this case, the objective function in (2.1) may be interpreted as a quantity which is proportional to minus a log likelihood in the case of simple random sampling with replacement (Brackstone and Rao 1979; Fuller 2002).

3. Asymptotic framework and nonresponse bias

We now consider the asymptotic properties of $\hat{T}_y$ with respect to both the sampling design and the nonresponse mechanism. We assume that the latter is such that each unit in the population responds, if sampled, with probability q_i, where this probability is not dependent on the choice of the sample and different units respond independently. We consider an asymptotic framework defined in terms of sequences of finite populations and associated probability sampling designs and response mechanisms (Fuller 2009, section 1.3), with orders of magnitude terms expressed in terms of $n = \sum_U \pi_i\, q_i,$ the expected number of responding units, and $N,$ the population size. We assume there exist positive constants K_1, K_2 and K_3 such that $K_1 < nN^{-1} d_i < K_2$ and $K_3 < q_i$ for all i.

We shall suppose that Horvitz-Thompson estimators of means are consistent for the corresponding finite population means and that central limit theorems hold (as expressed formally in the conditions of Theorem 1.3.9 of Fuller 2009). In particular, we assume that the sequences and the function $F(.)$ are such that there is a unique solution λ of

$$\sum_U q_i F(x_i'\lambda) x_i = T_x, \tag{3.1}$$

with

$$\hat{\lambda} = \lambda + O_p(n^{-0.5}), \tag{3.2}$$

and that

$$\hat{T}_y = \sum_U q_i F(x_i'\lambda) y_i + O_p(Nn^{-0.5}). \tag{3.3}$$

Deville and Särndal (1992) show that $\lambda = 0$ under certain assumptions (their Result 2). However, their assumptions apply just to the distribution induced by the sampling design and include the requirement that $N^{-1}(\hat{T}_{xd} - T_x) \to 0$ in probability. In the case of nonreponse, however, this requirement will often be implausible (*c.f.* Fuller 2002, page 15) and we do not require that λ be the zero vector.

A key assumption which we shall make is:

Condition C: there exists a vector α such that $F(x_i'\alpha) = q_i^{-1}$.

If condition C holds then α solves (3.1) and so $\lambda = \alpha$. It follows from (3.3) that $\hat{T}_y$ is consistent for T_y for any choice of variable y if this condition holds. Thus, we may view condition C as a sufficient condition for the absence of (asymptotic) nonresponse bias. This property of Condition C has been discussed by Fuller, Loughlin and Baker (1994), Fuller (2009, page 284) and Särndal and Lundström (2005, Proposition 9.2) for the case when F is linear. Fuller (2002, page 15), Kott (2006) and Chang and Kott (2008) also consider estimating response probabilities using general models of the form $q_i^{-1} = F(x_i'\alpha)$.

To illustrate what might happen if condition C does not hold, suppose that x_i is just a scalar with $x_i \equiv 1$. Then the unique solution of (3.1) is $\lambda = g(N/\sum_U q_i)$ and $p\lim(\hat{T}_y) = N(\sum_U q_i y_i)/(\sum_U q_i)$. Hence, the asymptotic nonresponse bias will only disappear for those survey variables which are 'uncorrelated' with the response probabilities q_i.

4. Linearization variance estimation

We now proceed to consider the asymptotic variance of $\hat{T}_y$ and its estimation. As in the previous section, the variance is defined with respect to the joint distribution induced by both sampling and nonresponse.

Note first that in general (and in particular for $G_M(.)$ and $G_{ML}(.)$), iteration is needed to solve the calibration equations. There does exist a literature (see Deville *et al.* 1993) which seeks to estimate the variance of $\hat{T}_y$ after a finite number of iterations. We follow instead the approach of Deville *et al.* (1993) and, for example, Binder and Théberge (1988) by approximating the variance of $\hat{T}_y$ by the variance of the 'converged' estimator, *i.e.*, the hypothetical estimator arising from an infinite number of iterations, represented by $\mathrm{var}(\sum_s w_i y_i)$, where the w_i are the 'converged' weights which solve the constrained optimisation problem in section 2.

A linearization variance estimator is obtained by approximating $\mathrm{var}(\sum_s w_i y_i)$ by $\mathrm{var}(\sum_s d_i z_i)$ for a 'linearized variable' z_i (Deville 1999). We now seek to construct this variable using a large sample argument. We first obtain an expression for $\hat{\lambda}$. A Taylor expansion of the left side of the calibration equations in (2.3) gives

$$\sum_s d_i F(x_i'\hat{\lambda}) x_i = \sum_s d_i F_i x_i + \sum_s d_i f(x_i'\lambda^*) x_i x_i'(\hat{\lambda} - \lambda),$$

where $F_i = F(x_i'\lambda)$, λ^* is between $\hat{\lambda}$ and λ and $f(u) = dF(u)/du$ is assumed to exist. Assuming also continuity of $f(.)$, the existence of $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ and using (3.2), we have

$$N^{-1}\sum_s d_i F(x_i'\hat{\lambda}) x_i = N^{-1}\sum_s d_i F_i x_i + N^{-1}\sum_s d_i f_i x_i x_i'(\hat{\lambda} - \lambda) + o_p(n^{-0.5}), \tag{4.1}$$

where $f_i = f(x_i'\lambda)$. Then, assuming $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ is non-singular and using (2.3), we obtain

$$\hat{\lambda} - \lambda = \left[\sum_s d_i f_i x_i x_i'\right]^{-1}\left[T_x - \sum_s d_i F_i x_i\right] + o_p(n^{-0.5}). \tag{4.2}$$

See Fuller (2009, proof of Theorem 1.3.9) for formal details of how (4.1) and (4.2) may be derived and the underlying regularity conditions. Note that to ensure $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ is non-singular may require dropping redundant variables from x_i and possibly (as in Deville and Särndal 1992) modifying the estimator for samples with small probability that result in singularity of this matrix.

A similar argument involving the Taylor expansion of w_i in (2.2) about λ gives:

$$w_i = d_i[F_i + f_i x_i'(\hat{\lambda} - \lambda)] + o_p(Nn^{-1.5}). \tag{4.3}$$

Then, assuming the existence of necessary population moments so that the remainder term in (4.3) holds uniformly across i (Fuller 2009, Corollary 2.7.1.1.), we have

$$\hat{T}_y \equiv \sum_s w_i y_i = \sum_s d_i\left[F_i + f_i x_i'(\hat{\lambda} - \lambda)\right] y_i + o_p(Nn^{-0.5}) \tag{4.4}$$

and hence from (4.2) and (4.4):

$$\hat{T}_y = \sum_s d_i F_i y_i + B\left[T_x - \sum_s d_i F_i x_i\right] + o_p(Nn^{-0.5}), \quad (4.5)$$

where

$$B = \left[\sum_s d_i f_i y_i x_i'\right]\left[\sum_s d_i f_i x_i x_i'\right]^{-1}. \quad (4.6)$$

Note that $F_i = f_i = 1$ under the assumptions of Deville and Särndal (1992) (since in this case $\lambda = 0$ and it follows from the assumptions about $G(.)$ that $F(0) = f(0) = 1$). Hence, under these assumptions, expression (4.5) corresponds to Result 5 of Deville and Särndal (1992), *i.e.*, the generalized raking estimator is asymptotically equivalent to the GREG estimator. Therefore, the asymptotic variance of $\hat{T}_y$ is the same as that of $\sum_s d_i z_i$, where z_i is the linearized variable:

$$z_i = F_i(y_i - \beta x_i), \quad (4.7)$$

and it is assumed that B converges to a finite limit matrix β. An alternative derivation of this expression is given by Demnati and Rao (2004, section 3.4).

For the purpose of linearization variance estimation, $\hat{T}_y$ is treated as the linear estimator $\sum_s d_i \hat{z}_i$, where

$$\hat{z}_i = \hat{F}_i(y_i - \hat{B} x_i) \quad (4.8)$$

is treated as a fixed variable.

A number of choices of $\hat{F}_i$ and $\hat{B}$ have been discussed in the literature. Starting with $\hat{F}_i$, the natural choice implied by the above argument is $\hat{F}_i = F(x_i' \hat{\lambda})$. A simpler choice, however, would be to take $\hat{F}_i = 1$. Deville and Särndal (1992) note that, in their classical theory with $\lambda = 0$, these choices are asymptotically equivalent but they express a preference for the choice $\hat{F}_i = F(x_i' \hat{\lambda})$. In our setting with nonresponse and with $\lambda = 0$ not necessarily holding, the second choice seems preferable and this is emphasized by Fuller (2002, page 15). Note that these two choices imply that $\sum_s d_i \hat{z}_i$ either takes the form $\sum w_i(y_i - \hat{B} x_i)$ when $\hat{F}_i = F(x_i' \hat{\lambda})$ or $\sum d_i(y_i - \hat{B} x_i)$ when $\hat{F}_i = 1$. We shall therefore refer to these choices as either w_i*-weighted residuals* or d_i *- weighted residuals*.

Regarding $\hat{B}$, it follows from our argument on the choices of $\hat{F}_i$ that f_i in (4.2) should be replaced by $\hat{f}_i = f(x_i' \hat{\lambda})$, giving:

(i) $\hat{B} = [\sum_s d_i \hat{f}_i y_i x_i'] [\sum_s d_i \hat{f}_i x_i x_i']^{-1}$, as also proposed by Demnati and Rao (2004).

Other choices are

(ii) $\hat{B} = \hat{B}_s$, as in (2.4), as proposed by Deville *et al.* (1993).

(iii) $\hat{B} = [\sum_s w_i y_i x_i'] [\sum_s w_i x_i x_i']^{-1}$, as proposed by Deville and Särndal (1992, equation 3.4), which might be more practical to compute than $\hat{B}_s$ for users of survey data files which include the w_i weights but not the d_i weights.

The extent to which these choices differ depends on the choice of $G(.)$ function. For the linear case $f(u) = 1$ so that the estimators in (i) and (ii) are identical. In the case of classical raking adjustment, $f(u) = F(u) = \exp(u)$ so that $\hat{f}_i = \hat{F}_i$ and $d_i \hat{f}_i = w_i$ and the estimators (i) and (iii) are identical. For the 'maximum likelihood' raking estimator we have $F(u) = (1-u)^{-1}$ and $f(u) = (1-u)^{-2}$ so that $d_i \hat{f}_i = w_i^2 / d_i$ and the three variance estimators are all distinct.

Having determined the form of $\hat{z}_i$ in (4.8), the linearization variance estimator for $\hat{T}_y$ is obtained by estimating the variance of the linear estimator $\sum_s d_i \hat{z}_i$, treating d_i and $\hat{z}_i$ as fixed. In the case of a stratified multistage sampling design, assuming "with replacement" sampling of primary sampling units (PSUs) within strata, a standard estimator of the variance (*e.g.*, Stukel *et al.* 1996) is:

$$\hat{V}(\hat{T}_y) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (z_{hj} - \bar{z}_h)^2 \quad (4.9)$$

where $z_{hj} = \sum_k d_{hjk} \hat{z}_{hjk}$, $\bar{z}_h = \sum_j z_{hj} / n_h$ and $\hat{z}_{hjk}$ is the value of the variable defined in (4.8) for the k^{th} individual within the j^{th} selected PSU in stratum h. This estimator remains appropriate in the presence of nonresponse if individual response in each PSU is independent of response in all other PSUs and if at least one individual is observed in each selected PSU (Fuller *et al.* 1994, page 78).

5. Simulation studies

In order to compare the performance of the weighted estimators and their corresponding variance estimators, two simulation studies were undertaken by constructing artificial populations using data from the British Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). In each case, $R = 1{,}000$ samples were generated from these populations by first sampling, in a way designed to mimic the real sampling scheme after some simplification, and then removing nonresponding cases according to two nonresponse models. The first assumes multiplicative nonresponse which, from Condition C in section 3, might be expected to lead to least bias for the raking ratio method. The second model assumed additive nonresponse, which might be expected to lead to least bias for the GREG estimator.

For each of the R samples, point estimates of parameters were calculated using the different generalized raking methods presented in section 2 and variance estimates were calculated using the different linearization methods presented in section 4. The properties of the estimators were then summarised.

5.1 Study based on the British Labour Force Survey

The first study was based upon data from the March-May 1998 quarter of the British LFS, a survey of persons living in private households in Britain, designed to provide information on the British labour market and carried out by the Office for National Statistics (ONS). The sample of approximately 58,000 households was treated as an artificial population. Repeated samples were drawn from this population in a way intended to mimic the design used for the LFS (ONS 1998, Section 3). Each sample consisted of 1,211 households selected by stratified simple random sampling with proportional allocation across 19 strata, defined by region of residence. These regions were designed to mimic interviewer areas which defined strata in the LFS. In the LFS all individuals in a sampled household are interviewed if possible. In this simulation study, all the respondents in a sample household were retained, except those aged under 16, who are not relevant for the estimates of interest.

The following two nonresponse models, based upon results of a study of Foster (1998), were used to determine whether sampled individuals responded.

Multiplicative Nonresponse Model:

$$\begin{aligned} q_i^{-1} = 1.15 \; & \times \; 1.17 \text{ (if London)} \\ & \times \; 1.13 \text{ (if aged under 35)} \\ & \times \; 1.1 \text{ (if female)} \end{aligned}$$

Additive Nonresponse Model:

$$\begin{aligned} q_i^{-1} = 1.15 \; & + \; 0.20 \text{ (if London)} \\ & + \; 0.15 \text{ (if aged under 35)} \\ & + \; 0.10 \text{ (if female)} \end{aligned}$$

where q_i is the response probability defined at the beginning of section 3 and the form of the model is chosen to satisfy Condition C.

Three parameters of interest are defined for the artificial population: the total number of persons unemployed, employed or inactive in the workforce. Weights were constructed for responding individuals, with calibration totals consisting of population counts in the categories of three categorical auxiliary variables and with Horvitz-Thompson initial weights d_i, as in section 2. The choice of auxiliary variables was designed to mimic those used in the LFS. However, because of the reduced scale of our artificial population and the consequent smaller numbers of individuals within strata, we simplified the LFS calibration variables to the following three categorical factors, defining 83 control totals:

- area of residence with 23 categories;
- a cross-classification of sex by 10 age groups (consisting of single years for those between 16 and 24 and a separate age group for 25 or older) with 20 categories;
- a cross-classification of region (Northern England; London and South East; Midlands and East Anglia; Scotland) by sex by age in 15-year age groups (16-29, 30-44, 45-59, 60-75 and 75 or older) with 40 categories.

5.2 Study based on the German sample Survey of Income and Expenditure

Our second study is based on the 1998 German Survey of Income and Expenditure (SIE), a national household survey conducted every 5 years by the Federal Statistical Office, to provide information about the economic and social situation of households, especially regarding the distribution of income and expenditure (Muennich and Schulrle 2003). We used data from a synthetic population of 64,326 households, created to represent 20% of all households from the Bremen region, excluding those with a monthly household net income of DM 35,000 or above (DM denotes the currency of German marks). A quota sampling design was employed for this survey and we have not attempted to mimic this design. Instead, our simulation study employs simple random sampling together with nonresponse. Repeated simple random samples of 1,340 households were drawn from the artificial population, representing a sampling fraction of about 1/48. Nonresponse models were constructed using the results of studies of similar surveys in Great Britain: the Family Expenditure Survey and the National Food Survey (Foster 1998). For each selected sample, the subset of responding households was determined by the following nonresponse models:

Multiplicative Model:

$$\begin{aligned} q_i^{-1} = 1.44 \; & \times \; 1.09 \text{ (if self-employed)} \\ & \times \; 1.03 \text{ (if unemployed)} \\ & \times \; 0.97 \text{ (if employed)} \\ & \times \; 1.16 \text{ (if no children in the household).} \end{aligned}$$

Additive Model:

$$\begin{aligned} q_i^{-1} = 1.44 \; & + \; 0.13 \text{ (if self-employed)} \\ & + \; 0.04 \text{ (if unemployed)} \\ & - \; 0.04 \text{ (if employed)} \\ & + \; 0.23 \text{ (if no children in the household).} \end{aligned}$$

The parameters of interest are the total household net income per quarter and the total household expenditure per quarter, computed from the finite artificial population.

As for the LFS study, each sampled household was assigned a weight. In the actual SIE the weights are constructed using essentially the maximum likelihood raking method by adjusting the sample data simultaneously to the marginal distributions of several characteristics, such as household type, social economic status of the reference person, household net income class and region (land). We try to mimic this adjustment, as far as possible, in our study. However, as for the LFS, because of the problem of strata with small numbers of households we simplify the SIE calibration variables to the following three categorical factors:

- household type with 7 categories
  - mother/father alone + 1child,
  - mother/father alone + 2 or more children,
  - couple with 1 child – spouse employed,
  - couple with 1 child – spouse unemployed,
  - couple with 2 or more children – spouse employed,
  - couple with 2 or more children – spouse unemployed,
  - other.
- social status of the reference person with 5 categories
  - self-employed,
  - civil servant or military,
  - employee,
  - worker,
  - unemployed, pensioner, student or other.
- household net income per quarter with 3 categories
  - 0-5,000 DM,
  - 5-7,000 DM,
  - 7-35,000 DM.

6. Results

6.1 Properties of point estimators

Table 6.1 presents the properties of the point estimators of total unemployed in the LFS study for different calibration methods and alternative assumptions about nonresponse. The properties are assessed following usual practice in simulation studies. For example, the bias in Table 6.1 is obtained from $\hat{B}(\hat{T}_y) = \hat{E}(\hat{T}_y) - T_y$, where $\hat{E}(\hat{T}_y) = 1/R \sum_{r=1}^{R} \hat{T}_{y_r}$, $\hat{T}_{y_r}$ is the value of $\hat{T}_y$ for sample r and R is the number of simulated samples. We observe from this table that the standard error remains virtually constant across alternative raking methods for a given nonresponse model. Nonresponse leads to an increase in the standard error across all estimators as expected (since the sample size is reduced). The table does show evidence of nonresponse bias, which is of a similar order for each of the raking methods. We do not find that this bias is least when the estimator matches the nonresponse model (*i.e.*, the GREG estimator for additive response and the raking estimator for multiplicative response) as we might have expected. Perhaps this is because the covariates used in the nonresponse models (*e.g.*, the aged 35+ variable) are not all included in the calibrating variables. Nevertheless, the nonresponse bias is small in the sense that the root mean square error is very similar to the standard error in each case. Under nonresponse, the GREG calibration method generates some negative weights whereas this is avoided by the two raking methods, as expected. A greater number of very large weights are observed, however, for the 'maximum likelihood' raking estimator.

Corresponding results for the SIE data are presented in Table 6.2. The pattern of results is broadly similar, although there is now no evidence of significant nonresponse bias (*i.e.*, the observed bias could be explained by simulation variation). The standard errors and root mean square errors also remain virtually constant across weighting methods for a given nonresponse model.

Table 6.1
Simulation properties of point estimators of total unemployed using data from LFS with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negative Weights[1]	Number of Very Large Weights[1, 2]
Complete Response:					
GREG	7.6 (14.3)	452.8	452.8	0	0
Classical Raking	8.3 (14.3)	452.8	452.9	0	0
'ML' Raking	9.0 (14.3)	453.3	453.4	0	1
Multiplicative nonresponse:					
GREG	-45.6 (15.8)	498.3	500.3	4	1
Classical Raking	-42.1 (15.8)	498.8	500.6	0	2
'ML' Raking	-39.7 (15.8)	499.4	501.0	0	7
Additive nonresponse:					
GREG	-37.3 (15.7)	497.4	498.8	5	1
Classical Raking	-34.7 (15.7)	497.5	498.7	0	3
'ML' Raking	-32.4 (15.8)	498.1	499.1	0	7

[1] the number of such weights across all sample units and all 1000 samples.
[2] the number of weights more than 10 times the corresponding design weight.

Table 6.2
Simulation properties of point estimators of total income using data from SIE with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negatives Weights	Number of Very Large Weights
Complete Response:					
GREG	-172.2 (331.3)	10,477.3	10,478.7	0	0
Classical Raking	-170.6 (331.5)	10,484.1	10,485.8	0	0
‘ML’ Raking	-169.8 (331.8)	10,491.5	10,492.9	0	0
Multiplicative nonresponse:					
GREG	-495.7 (429.7)	13,586.8	13,595.8	0	0
Classical Raking	-493.8 (429.6)	13,584.6	13,593.5	0	0
‘ML’ Raking	-463.5 (429.5)	13,582.8	13,590.7	0	0
Additive nonresponse:					
GREG	-473.2 (430.5)	13,614.8	13,623.0	0	0
Classical Raking	-469.4 (430.5)	13,612.9	13,621.0	0	0
‘ML’ Raking	-439.5 (430.5)	13,613.5	13,620.6	0	0

6.2 Properties of variance estimators

The properties of the different estimators of the variances of the point estimators of the total unemployed from the LFS are shown in the Table 6.3 (the ‘standard error estimate’ in the table refers to the square root of the variance estimate). We make a number of observations:

- weighting the residuals by w_i rather than by d_i reduces the bias and root mean squared error of the standard error estimator. The bias arising from the use of d_i weighted residuals in the case of nonresponse is particularly important (as noted by Fuller 2002) but there are also non-negligible reductions of bias even in the complete response case.
- The choice of weight used in $\hat{B}$ for the calculation of residuals seems to have little impact.
- For a given nonresponse setting and choice of weighting the residuals, there is little difference in the results for the different choices of point estimator.

The results in Table 6.3 are extended in Table 6.4 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider two additional parameters: total numbers employed and inactive. We see again that the relative bias arising from using d_i weighted residuals can be substantial in the presence of nonresponse, over 20% in several cases, and that this is reduced using the w_i weighted residuals. Again, little change is observed in the percent relative bias of the standard error estimators when different choices of weights are used in the calculation of $\hat{B}$ for the residuals.

Corresponding results for the SIE data when estimating total income are shown in Table 6.5. Again, the pattern of results is broadly similar to that for the LFS data in Table 6.3. For the complete response case, the use of w_i weighted residuals rather than d_i weighted residuals leads to modest improvement in bias and RMSE of the standard error estimators. For the nonresponse cases the improvements are considerable. Little change in the standard error estimators is observed when modifying the choice of weight used to compute the estimated regression coefficients. The results in Table 6.5 are extended in Table 6.6 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider one additional parameter: total expenditure per quarter. We see again that the relative bias arising from using d_i weighted residuals can be substantial in the presence of nonresponse, over 35% in all cases, and that this is reduced using the w_i weighted residuals, for which the relative bias never exceeds about 3%.

Table 6.3
Properties of variance estimators when estimating total unemployed from the LFS (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	*d*	*d*	433.9	-18.8 (0.9)	33.4	93.5
	d	*w*	434.3	-18.5 (0.9)	33.3	93.5
	w	*d*	442.8	-10.0 (1.0)	31.9	93.8
	w	*w*	441.9	-10.8 (1.0)	32.0	93.7
Classical Raking	*d*	*d*	433.9	-18.8 (0.9)	33.4	93.5
	d	*w*	434.2	-18.5 (0.9)	33.3	93.5
	w	*d*	443.0	-9.8 (1.0)	32.0	93.8
	w	*w*	442.0	-10.7 (1.0)	32.0	93.8
'ML' Raking	*d*	*d*	433.9	-19.4 (0.9)	33.7	93.5
	d	*w*	434.3	-19.1 (0.9)	33.6	93.5
	d	*df*	435.4	-17.9 (0.9)	33.0	93.5
	w	*d*	443.7	-9.6 (1.0)	32.5	93.7
	w	*w*	442.3	-11.1 (1.0)	32.4	93.7
	w	*df*	441.6	-11.8 (1.0)	32.3	93.7
Multiplicative nonresponse:						
GREG	*d*	*d*	385.7	-112.6 (0.9)	116.0	85.8
	d	*w*	386.1	-112.1 (0.9)	115.5	85.8
	w	*d*	489.5	-8.8 (1.2)	39.2	94.2
	w	*w*	487.8	-10.4 (1.2)	39.2	94.2
Classical Raking	*d*	*d*	385.7	-113.1 (0.9)	116.5	85.7
	d	*w*	386.1	-112.7 (0.9)	116.1	85.7
	w	*d*	490.3	-8.5 (1.2)	39.6	94.3
	w	*w*	488.4	-10.4 (1.2)	39.5	94.1
'ML' Raking	*d*	*d*	385.7	-113.7 (0.9)	117.1	85.4
	d	*w*	386.2	-113.2 (0.9)	116.6	85.6
	d	*df*	387.8	-111.6 (0.9)	115.0	85.8
	w	*d*	491.9	-7.5 (1.3)	40.4	94.2
	w	*w*	488.9	-10.5 (1.2)	39.9	94.0
	w	*df*	487.5	-11.9 (1.2)	39.8	94.0
Additive nonresponse:						
GREG	*d*	*d*	386.5	-110.9 (0.9)	114.4	86.0
	d	*w*	387.0	-110.5 (0.9)	113.9	86.0
	w	*d*	489.3	-8.2 (1.2)	39.0	94.6
	w	*w*	487.6	-9.8 (1.2)	39.0	94.6
Classical Raking	*d*	*d*	386.5	-111.0 (0.9)	114.4	85.8
	d	*w*	387.0	-110.6 (0.9)	114.0	85.8
	w	*d*	490.1	-7.4 (1.2)	39.2	94.7
	w	*w*	488.1	-9.4 (1.2)	39.1	94.6
'ML' Raking	*d*	*d*	386.5	-111.6 (0.9)	115.0	85.6
	d	*w*	387.0	-111.1 (0.9)	114.6	85.6
	d	*df*	388.6	-109.5 (0.9)	113.0	85.9
	w	*d*	491.6	-6.5 (1.3)	40.0	94.7
	w	*w*	488.6	-9.5 (1.2)	39.5	94.6
	w	*df*	487.3	-10.8 (1.2)	39.4	94.6

[1] see text following equation (4.8), where choices df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.
[2] percentage of 95% normal-theory confidence intervals containing true value.

Table 6.4
Relative bias (%) of standard error estimators of unemployed, employed and inactive totals from LFS (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Relative Bias of Standard Error Estimator		
			Unemployed	Employed	Inactive
Complete Response:					
GREG	*d*	*d*	-4.2	-3.4	0.5
	d	*w*	-4.1	-3.3	0.6
	w	*d*	-2.2	-2.2	1.9
	w	*w*	-2.4	-2.3	1.7
Classical Raking	*d*	*d*	-4.2	-3.3	0.7
	d	*w*	-4.1	-3.2	0.8
	w	*d*	-2.2	-2.1	2.1
	w	*w*	-2.4	-2.2	1.9
‘ML’ Raking	*d*	*d*	-4.3	-3.3	0.7
	d	*w*	-4.2	-3.3	0.8
	d	*df*	-4.0	-3.1	1.1
	w	*d*	-2.1	-2.0	2.3
	w	*w*	-2.4	-2.2	1.9
	w	*df*	-2.6	-2.3	1.8
Multiplicative nonresponse:					
GREG	*d*	*d*	-22.6	-22.3	-18.2
	d	*w*	-22.5	-22.2	-18.1
	w	*d*	-1.8	-3.3	1.8
	w	*w*	-2.1	-3.5	1.5
Classical Raking	*d*	*d*	-22.7	-30.6	-18.4
	d	*w*	-22.6	-30.5	-18.3
	w	*d*	-1.7	-13.5	1.7
	w	*w*	-2.1	-13.7	1.3
‘ML’ Raking	*d*	*d*	-22.8	-22.0	-18.4
	d	*w*	-22.7	-21.9	-18.3
	d	*df*	-22.3	-21.7	-17.9
	w	*d*	-1.5	-2.7	1.9
	w	*w*	-2.1	-3.1	1.3
	w	*df*	-2.4	-3.3	1.1
Additive nonresponse:					
GREG	*d*	*d*	-22.3	-21.8	-18.5
	d	*w*	-22.2	-21.7	-18.4
	w	*d*	-1.6	-2.9	1.1
	w	*w*	-2.0	-3.1	0.8
Classical Raking	*d*	*d*	-22.3	-30.2	-18.0
	d	*w*	-22.2	-30.1	-17.9
	w	*d*	-1.5	-13.3	1.8
	w	*w*	-1.9	-13.5	1.4
‘ML’ Raking	*d*	*d*	-22.4	-21.6	-18.0
	d	*w*	-22.3	-21.5	-17.9
	d	*df*	-22.0	-21.3	-17.6
	w	*d*	-1.3	-2.4	2.0
	w	*w*	-1.9	-2.8	1.5
	w	*df*	-2.2	-3.0	1.3

[1] see text following equation (4.8), where df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.

Table 6.5
Properties of variance estimators when estimating total income from the SIE (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	*d*	*d*	10,338.8	-138.5 (6.9)	259.0	93.8
	d	*w*	10,339.2	-138.2 (6.9)	258.8	93.8
	w	*d*	10,377.9	-99.5 (6.9)	240.0	94.1
	w	*w*	10,376.8	-100.5 (6.9)	240.3	94.1
Classical Raking	*d*	*d*	10,338.8	-145.3 (6.9)	262.7	93.8
	d	*w*	10,339.2	-144.9 (6.9)	262.5	93.8
	w	*d*	10,370.0	-106.1 (6.9)	243.1	94.0
	w	*w*	10,376.9	-107.2 (6.9)	243.5	94.0
'ML' Raking	*d*	*d*	10,338.8	-152.7 (6.9)	266.9	93.9
	d	*w*	10,339.2	-152.4 (6.9)	266.7	93.9
	d	*df*	10,340.3	-151.3 (6.9)	266.1	94.0
	w	*d*	10,378.3	-113.2 (6.9)	246.5	94.0
	w	*w*	10,377.1	-114.4 (6.9)	247.0	94.0
	w	*df*	10,376.7	-114.8 (6.9)	247.2	94.0
Multiplicative nonresponse:						
GREG	*d*	*d*	8,104.7	-5,482.1 (7.4)	5,487.1	75.8
	d	*w*	8,105.5	-5,481.3 (7.4)	5,486.3	75.8
	w	*d*	13,214.5	-372.3 (12.8)	549.7	94.5
	w	*w*	13,210.9	-375.9 (12.8)	551.7	94.5
Classical Raking	*d*	*d*	8,104.7	-5,479.8 (7.4)	5,484.9	75.8
	d	*w*	8,105.5	-5,479.1 (7.4)	5,484.1	75.8
	w	*d*	13,214.1	-370.4 (12.8)	549.4	94.5
	w	*w*	13,210.4	-374.2 (12.8)	551.5	94.5
'ML' Raking	*d*	*d*	8,104.7	-5,478.1 (7.4)	5,483.1	75.8
	d	*w*	8,105.5	-5,477.3 (7.4)	5,482.3	75.8
	d	*df*	8,108.1	-5,474.7 (7.4)	5,479.7	75.9
	w	*d*	13,215.2	-367.6 (12.9)	549.4	94.5
	w	*w*	13,210.6	-372.2 (12.9)	551.6	94.5
	w	*df*	13,208.9	-373.9 (12.9)	552.3	94.5
Additive nonresponse:						
GREG	*d*	*d*	8,106.3	-5,508.5 (7.4)	5,513.5	75.6
	d	*w*	8,107.1	-5,507.7 (7.4)	5,512.7	75.6
	w	*d*	13,207.9	-407.0 (12.8)	573.8	94.3
	w	*w*	13,204.3	-410.5 (12.8)	575.9	94.3
Classical Raking	*d*	*d*	8,106.3	-5,506.6 (7.4)	5,511.6	75.7
	d	*w*	8,107.1	-5,505.9 (7.4)	5,510.9	75.7
	w	*d*	13,207.7	-405.3 (12.8)	573.6	94.1
	w	*w*	13,203.9	-409.0 (12.8)	575.8	94.1
'ML' Raking	*d*	*d*	8,106.3	-5,507.2 (7.4)	5,512.2	75.9
	d	*w*	8,107.1	-5,506.4 (7.4)	5,511.4	75.9
	d	*df*	8,109.7	-5,503.8 (7.4)	5,508.8	75.9
	w	*d*	13,208.9	-404.6 (12.9)	574.8	94.1
	w	*w*	13,204.2	-409.2 (12.9)	577.3	94.1
	w	*df*	13,202.5	-411.0 (12.9)	578.1	94.1

[1]see text following equation (4.8), where choices df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.
[2] percentage of 95% normal-theory confidence intervals containing true value.

Table 6.6
Relative bias (%) of variance estimators of expenditure and income totals from SIE (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Relative Bias of Standard Error Estimator	
			Expenditure	Income
Complete Response:				
GREG	*d*	*d*	0.7	-1.3
	d	*w*	0.7	-1.3
	w	*d*	1.3	-1.0
	w	*w*	1.3	-1.0
Classical Raking	*d*	*d*	0.7	-1.4
	d	*w*	0.7	-1.4
	w	*d*	1.2	-1.0
	w	*w*	1.2	-1.0
'ML' Raking	*d*	*d*	0.6	-1.5
	d	*w*	0.6	-1.5
	d	*df*	0.6	-1.4
	w	*d*	1.2	-1.1
	w	*w*	1.2	-1.1
	w	*df*	1.2	-1.1
Multiplicative nonresponse:				
GREG	*d*	*d*	-38.2	-40.4
	d	*w*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.8
Classical Raking	*d*	*d*	-38.2	-40.3
	d	*w*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.8
'ML' Raking	*d*	*d*	-38.2	-40.3
	d	*w*	-38.2	-40.3
	d	*df*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.7
	w	*df*	-0.4	-2.8
Additive nonresponse:				
GREG	*d*	*d*	-38.1	-40.5
	d	*w*	-38.1	-40.5
	w	*d*	-0.2	-3.0
	w	*w*	-0.2	-3.0
Classical Raking	*d*	*d*	-38.1	-40.5
	d	*w*	-38.1	-40.5
	w	*d*	-0.2	-3.0
	w	*w*	-0.2	-3.0
'ML' Raking	*d*	*d*	-38.2	-40.5
	d	*w*	-38.2	-40.5
	d	*df*	-38.1	-40.4
	w	*d*	-0.2	-3.0
	w	*w*	-0.3	-3.0
	w	*df*	-0.3	-3.0

[1] see text following equation (4.8), where df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.

7. Conclusions

The simulation study showed little difference between the bias or variance properties of the three calibration estimators considered: the GREG estimator, the classical raking estimator and the maximum likelihood raking estimator. Some small differences in the distribution of extreme weights were observed: the maximum likelihood raking estimator had the most very large weights and the GREG estimator was the only one with a few negative weights.

Amongst the variance estimators, the main finding was the contrast between the approach which weights residuals by the design weight and that which weights them by the calibrated weight. It was found that the latter variance estimator always had smaller bias and that this effect was very marked in the presence of nonresponse, when the former estimator could be severely biased. The bias of the latter estimator was generally small and the coverage level of the associated confidence intervals was generally close to the nominal coverage.

Alternative ways of weighting the observations in constructing the regression coefficients, when calculating the residuals in the linearization variance estimator, were considered but little effect was observed and there was no evidence that this choice is important in practice.

In general, the findings for the categorical variables in the British Labour Force Survey were remarkably similar to the findings for the continuous variables in the German Income and Expenditure survey.

Acknowledgements

Comments from two referees helped improve this paper significantly. We are grateful to the Office for National Statistics for making the Labour Force Survey data available and to Ralf Münnich and colleagues on the DACSEIS project (http://www.dacseis.de/) for providing the synthetic population based on the German Survey of Income and Expenditure. This research was supported by the Economic and Social Research Council.

References

Binder, D.A., and Théberge, A. (1988). Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, 16, Supp. 47-55.

Brackstone, G.J., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyã*, Series C, 41, 97-114.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.

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, 376-82.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-20.

Foster, K. (1998). Evaluating nonresponse on household surveys. *GSS Methodology Series*, 8, Office for National Statistics. London.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken: Wiley.

Fuller, W.A., Loughlin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., and Maligalig, D.S. (1991). A comparison of methods for weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Muennich, R., and Schulrle, J. (2003). Monte Carlo simulation study of European surveys, Workpackage 3, Deliverables 3.1 and 3.2. DACSEIS project. Available at http://www.uni-trier.de/index.php?id=29730.

Office for National Statistics (1998). *Labour Force Survey User Guide, Volume* 1: *Background and Methodology*. London.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, England.

Stukel, D.M., Hidiroglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.