DACSEIS

IST-2000-26057

Workpackage 8

Variance Estimation for Estimators subject to Raking Adjustment

Deliverable 8.2

2002

List of contributors:

Yves Berger, Chris Skinner, University of Southampton.

Main responsibility:

Chris Skinner, University of Southampton.

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

© http://www.DACSEIS.de

Preface

Raking ratio estimation is often used with household surveys in government statistics as a method of calibration and to adjust for nonresponse. This document provides S-Plus code to implement a variety of variance estimators for use with raking ratio estimation. Two types of raking ratio estimators are considered, together with the generalized regression estimator as a benchmark for comparison. The main focus is on linearization methods, for which the variance estimator for a raking ratio estimator is similar to that for the generalized regression estimator, subject to come choices of weights. Jackknife variance estimation is also considered. A simulation study of the properties of the alternative point estimators and variance estimators is undertaken using data from both the British Labour Force Survey and the German Sample Survey of Income and Expenditure, allowing not only for sampling but also non-response.

Yves Berger and Chris Skinner — Southampton, October 2002

© http://www.DACSEIS.de

Contents

© http://www.DACSEIS.de

Chapter 1

Introduction

Raking adjustment is a method for constructing survey weights using auxiliary information on population totals.The classical method involves adjusting an initial set of survey weights by an iterative proportional fitting algorithm so that the weighted estimates agree with given marginal population totals for two or more categorical auxiliary variables. For example, in the UK Labour Force Survey three auxiliary variables are used: (a) a geographical variable, with 454 categories (areas), (b) a variable with 22 categories formed by cross-classifying age group and sex, (c) a variable with 578 categories formed by cross-classifying a different age classification by sex and region. At each step of the the algorithm, the weights are 'raked' to achieve agreement with the population totals in the categories of one of the auxiliary variables. The algorithm is typically continued until the differences between the weighted estimates of the auxiliary population totals and the known totals are all within a specified tolerance.

Approaches to raking adjustment are described in Chapter 2. The classical method of iterative proportional fitting was first suggested by Deming and Stephan (1940). Brackstone and Rao (1976, 1979) set out this classical estimator clearly in a sample survey framework, as well as introducing a 'maximum likelihood' version of the estimator, discussed in Section 2.2 Bankier (1986) considered the application of raking to multiple frame sampling. Oh and Scheuren (1987) proposed a modified raking approach, which is a hybrid between classical raking adjustment using two-way margins and poststratification using the cells of the two-way table. Deville *et al.* (1993) defined a generalised class of raking estimators, described in Section 2.3, which includes the classical method as a special case.

In this paper, we consider two approaches to variance estimation: linearization and replication methods. The linearization method depends upon large sample theory for the variances of the raked estimators.

Brackstone and Rao (1976, 1979) set out some basic estimation theory including results on variances for certain designs and for a fixed number of iterations. Konijn (1981) and Choudhry and Lee (1987) presented related theory. Binder and Théberge (1988) proposed a general linearisation approach for the limiting estimator, i.e. the estimator obtained after the iterations have converged. Deville *et al.* (1993) proposed an approach to variance estimation for their generalised class of raking estimators by linearisation, using a result of Deville and Särndal (1992), which showed that under certain conditions the

asymptotic variance of the raking estimator is the same as that of a regression estimator. DEMNATI and RAO (2004) discussed these results and proposed an alternative variance estimator. The linearization method is described in more detail in Chapter 3.

Replication methods, such as the jackknife, the bootstrap and balanced repeated replications and their treatment of raking are reviewed in SHAO and TU (1995),Ch.6 and RUST and RAO (1996). The basic approach is described in Chapter 4.

© http://www.DACSEIS.de

Chapter 2

Estimators Using Raking

2.1 Classical Raking Adjustment

Consider the estimation of a population total τ_Y of a survey variable Y taking values y_i for units i in a population U:

$$\tau_Y = \sum_U y_i$$

Given observations on y_i for units in a sample S, a basic weighted estimator of τ_Y is given by

$$\widehat{\tau}_Y = \sum_S \omega_i y_i$$

where ω_i is a given weight, referred to here as the *initial weight*. This weight may, for example, be the Horvitz-Thompson weight $\omega_i = 1/\pi_i$. In Chapter 4 we shall assume that the weights ω_i are fixed, that is not sample-dependent.

The classical raking adjustment makes use of information on the population counts in the categories of two or more categorical auxiliary variables. Let x_i denote the vector of indicator variables of these categories, for example in the case of three auxiliary variables:

$$x_i = (\delta_{1..i}, \ldots, \delta_{A..i}, \delta_{.1.i}, \ldots, \delta_{.B.i}, \delta_{..1i}, \ldots, \delta_{..\mathrm{Ci}})',$$

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 τ_X of this vector thus contains the population counts in each of the (marginal) categories of each of the three auxiliary variables. It is assumed that τ_X is given and that x_i is known for $i \in s$.

The classical raking adjustment involves iterative modification of the initial weights, ω_i, in a multiplicative way to adjusted weights, w_i, with the aim of satisfying the calibration equations:

$$\sum_S w_i x_i = \tau_X.$$

The resulting raking estimator of τ_Y is:

$$\widehat{\tau}_{YRak} = \sum_S w_i y_i$$

The multiplicative adjustment, depends only upon the cell in the contingency table formed by the auxiliary variables, that is we may write $w_i = \omega_i h(x_i)$, where the multiplicative adjustment factor $h(x_i)$ is fixed for all units with common values of the auxiliary variables. Let $\widehat{N}_\omega[h(x)]$ and $\widehat{N}_w[h(x)]$ denote the weighted estimates of the population counts in the cell of the table defined by x, using the weights ω_i and w_i respectively. Then we may write $\widehat{N}_w[h(x)] = h(x)\widehat{N}_\omega[h(x)]$. The usual iterative modification of the weights involves iterative proportional fitting (BRACKSTONE and RAO, 1979). IRELAND and KULLBACK (1968) demonstrate that this method converges to a solution which minimises:

$$\sum \widehat{N}_w \log(\widehat{N}_w/\widehat{N}_\omega),$$

subject to the calibration equations, where the sum is over all the cells defined by x. The objective function may alternatively be expressed as:

$$\sum_S w_i \log(w_i/\omega_i),$$

that is, under convergence of the iterative algorithm, the w_i minimise the above function, subject to solving the calibration equations. The objective function in this optimisation problem may alternatively be expressed as

$$\sum_S \omega_i G_M(w_i/\omega_i)$$

where $G_M(u) = u\log(u) - u + 1$ is the multiplicative distance measure considered by DEVILLE *et al.* (1993), and it is assumed that the calibration equations imply that $\sum w_i$ is constrained to be a given constant. Using the standard Lagrange multiplier method for constrained minimisation, the multiplicative adjustment factors may be expressed as (DEVILLE *et al.*, 1993):

$$w_i = \omega_i F_M(x_i'\widehat{\lambda})$$

where $F_M(u) = g_M^{-1}(u)$ denotes the inverse function of $g_M(u) = dG_M(u)/du$ and $\widehat{\lambda}$ is the Lagrange multiplier, which solves the calibration equations. It follows from the definition of $G_M(u)$ above that $g_M(u) = \log(u)$ and $F_M(u) = \exp(u)$. Hence $\widehat{\lambda}$ solves

$$\sum_S \omega_i \exp(x_i'\widehat{\lambda})x_i = \sum_U x_i.$$

© http://www.DACSEIS.de

In the discussion in this section, the estimator $\widehat{\tau}_{YRak} = \sum w_i y_i$ may be used if y_i is a scalar or a vector, since the scalar weights, w_i, do not depend upon y_i. More generally, if θ is a population parameter which may be expressed as a function of a vector of population totals $\sum y_i$ then the corresponding function of the weighted sample sum $\sum w_i y_i$ provides a raked estimator of θ.

2.2 'Maximum Likelihood' Raking Adjustment

An alternative approach to raking adjustment involves a maximum likelihood argument for the estimation of the population proportions in the cells of the table formed by cross-classifying the auxiliary variables. The calibration equations remain the same but the objective function, summed over the cells, is replaced by (Brackstone and Rao, 1979; Fuller, 2002):

$$-\sum \widehat{N}_\omega \log(\widehat{N}_w),$$

which is proportional to minus the log likelihood in the case of simple random sampling with replacement (Brackstone and Rao, 1979). Equivalently, the objective function may be expressed, summed over sample units, as:

$$\sum_S -\omega_i \log(w_i/\omega_i),$$

(omitting an additional term not involving w_i), or as (Fuller, 2002):

$$\sum_S \omega_i G_{ML}(w_i/\omega_i),$$

where $G_{ML}(u) = u - 1 - \log(u)$. The function G_{ML} is Case 4 of the distance measures considered by Deville and Särndal (1992).

2.3 Generalized Raking Adjustment

Deville *et al.* (1993) discuss other choices G for the functions G_M and G_{ML} considered above. Alternative functions might be chosen, for example, in order to reduce the variation in the resulting adjusted weights. The function G is required to be strictly convex in order for a unique solution to exist for the optimisation problem.

© http://www.DACSEIS.de

Chapter 3

Variance Estimation by Linearization

3.1 The Asymptotic Variance

Following Binder and Théberge (1988) and Deville *et al.* (1993), we consider estimating the asymptotic variance of the 'converged' estimator, i.e. the estimator $\widehat{\tau}_{YRak}$, where the weights w_i solve the constrained optimisation problem. This asymptotic variance is assumed to be a sufficiently close approximation to the asymptotic variance of the estimator obtained after the finite number of iterations used in practice.

We assume that in large samples, $\widehat{\lambda}$ converges to a value λ. Deville and Särndal (1992) assume that $\lambda = 0$, but this property is based upon the assumption that the estimator of τ_X, obtained by applying the basic weights ω_i, is consistent. This assumption will often be false in the case of non-response and we prefer not to make this assumption.

We allow the function $F(.)$ to be general and not necessarily equal to $F_M(.)$. We first expand the adjusted weight $w_i = \omega_i F(x_i'\widehat{\lambda})$ about λ to obtain

$$w_i \simeq \omega_i[F_i + f_i x_i'(\widehat{\lambda} - \lambda)]$$

where $F_i = F(x_i'\lambda)$, $f_i = f(x_i'\lambda)$, $f(u) = dF(u)/du$. Substituting in the calibration equations we obtain:

$$\sum_S \omega_i[F_i + f_i x_i'(\widehat{\lambda} - \lambda)]x_i \simeq \tau_X$$

and hence

$$\widehat{\lambda} - \lambda \simeq [\sum_S \omega_i f_i x_i x_i']^{-1}[\tau_X - \sum_S \omega_i F_i x_i]$$

We are assuming here that the first matrix in this expression is non-singular. It may be necessary to drop redundant variables from x_i to achieve this. For example, in the three-way case above, each of the sums of the indicator variables $\delta_{a..i}$, $\delta_{.b.i}$ and $\delta_{..ci}$ across a, b and c, respectively, equals 1 and it is natural to drop two of these indicators to avoid singularity.

Substituting in the estimator of interest we obtain:

$$\begin{aligned}\widehat{\tau}_{YRak} &\simeq \sum_S \omega_i[F_i + f_i x_i'(\widehat{\lambda} - \lambda)]y_i \\ &\simeq \sum_S \tilde{w}_i y_i + B[\tau_X - \sum_S \tilde{w}_i x_i]\end{aligned}$$

where $\tilde{w}_i = \omega_i F_i$ and $B = [\sum_S \omega_i f_i y_i x_i'][\sum_S \omega_i f_i x_i x_i']^{-1}$.

We assume that B converges to a finite limit matrix β in the asymptotic framework. It follows from this and the other approximating assumptions above, in particular that the basic weights ω_i are fixed, that the (normalized) asymptotic distribution of $B[\tau_X - \sum_S \tilde{w}_i x_i]$ is the same as that of $\beta[\tau_X - \sum_S \tilde{w}_i x_i]$. Hence the (normalized) asymptotic variance of $\widehat{\tau}_{YRak}$ is the same as that of $\sum_S z_i$, where z_i is the linearized variate

$$z_i = \tilde{w}_i(y_i - \beta x_i)$$

3.2 Linearisation Variance Estimator

For the purpose of linearization variance estimation, $\widehat{\tau}_{YRak}$ may be treated as the linear estimator $\sum_S \widehat{z}_i$, where $\widehat{z}_i = w_i(y_i - \widehat{B}x_i)$ is treated as a fixed variable and $\widehat{B}$ is defined in the same way as B, with f_i replaced by $\widehat{f}_i = f(x_i'\widehat{\lambda})$. The variance estimator for $\widehat{\tau}_{YRak}$ is then obtained for a given sampling scheme by using a standard variance estimator for that scheme for a linear estimator, applied to $\sum_S \widehat{z}_i$. The same approach applies whether y_i is a scalar or a vector. For an estimator $\widehat{\theta}$ which is a smooth function of a weighted sample sum, the estimator may first be linearized by the usual linearization technique to create a linearized variate u_i for which the variance of $\widehat{\theta}$ may be approximated by the variance of $\sum w_i u_i$,treating the w_i as fixed quantities. To allow for raking adjustment in the weights w_i, the standard variance estimator for the linear estimator $\sum_S \widehat{z}_i$ may again be applied, with u_i replacing y_i in the definition of $\widehat{z}_i$.

The definition of $\widehat{B}$ above in the formula for $\widehat{z}_i$ gives the same variance estimator proposed by DEMNATI and RAO (2004). Following DEVILLE and SÄRNDAL (1992), DEVILLE *et al.* (1993) propose a similar, but not in general identical, variance estimator, where the linearized variate in our notation becomes $z_{DSi} = w_i(y_i - B_{DS}x_i)$, where

$$B_{DS} = \left[\sum_S w_i y_i x_i'\right]\left[\sum_S w_i x_i x_i'\right]^{-1}$$

They note that under the assumptions they make, a consistent variance estimator would also be obtained by replacing w_i by ω_i in B_{DS}. They express a preference for using w_i.

In the case of classical raking adjustment, $F(u) = F_M(u) = \exp(u)$. Hence, $f(u) = f_M(u) = F_M(u)$ and $w_i = \omega_i \widehat{f}_i$ and so $\widehat{B} = B_{DS}$ and the two variance estimators are identical. For the 'maximum likelihood' choice $G_{ML}(u) = u - 1 - \log(u)$, we have $F_{ML}(u) = (1-u)^{-1}$ and $f_{ML}(u) = (1-u)^{-2}$ so that $\omega_i \widehat{f}_i = w_i^2/\omega_i$ and the two variance estimators are not identical.

© http://www.DACSEIS.de

3.3 Computational Simplification of Linearization Variance Estimator

The computation of $\widehat{B}$ (or B_{DS}) may be heavy or prohibitive if both the sample size and the dimension of x_i are large. We consider here how the computational burden may be reduced for the case of raking to two margins.

Let E be the vector of values $y_i - \widehat{B}x_i$ which it is necessary to compute in order to calculate the linearization variance estimator. The computational demands arise in the calculation of $\widehat{B}$. Let Y be the vector of values y_i and X the matrix with rows x_i', so that

$$E = Y - X\widehat{B}'.$$

Let $W = \text{diag}(\omega_i \widehat{f}_i)$ (or $\text{diag}(w_i)$ for B_{DS}), $\breve{Y} = W^{1/2}Y$, $\breve{X} = W^{1/2}X$ so that $\widehat{B} = (\breve{X}'\breve{X})^{-1}\breve{X}'\breve{Y}$.

Partition X, $\breve{X}$ and $\widehat{B}$ according to the constraints defined by the two margins:

$$X = (X_1 X_2), \breve{X} = (\breve{X}_1 \breve{X}_2) \text{ and } \widehat{B} = (\widehat{B}_1 \widehat{B}_2).$$

Let

$$\begin{aligned} R_1 &= I - \breve{X}_1(\breve{X}_1'\breve{X}_1)^{-1}\breve{X}_1' \\ \tilde{Y} &= R_1\breve{Y}, \tilde{X}_2 = R_1\breve{X}_2 \\ \tilde{B}_2' &= (\tilde{X}_2'\tilde{X}_2')^{-1}\tilde{X}_2'\tilde{Y}. \end{aligned}$$

Then it follows from Theorem 3.7 of Seber (1977) that

$$\breve{Y} - \breve{X}\widehat{B}' = \tilde{Y} - \tilde{X}_2\widehat{B}_2',$$

i.e the least squares residual of $\breve{Y}$ on $\breve{X}$ may be obtained by first taking the least squares residuals $\tilde{Y}$ and $\tilde{X}_2$ of $\breve{Y}$ and $\breve{X}_2$ respectively on X_1 and then computing the least squares residual of $\tilde{Y}$ on $\tilde{X}_2$.

It follows that

$$\begin{aligned} E &= Y - X\widehat{B}' \\ &= W^{-1/2}(\tilde{Y} - \tilde{X}_2\widehat{B}_2') \\ &= R_{1W}(Y - X_2\widehat{B}_2') \end{aligned}$$

where $R_{1W} = I - X_1(X_1'WX_1)^{-1}X_1W$.

The computation of E thus involves three main steps: the computation of (i) $R_{1W}Y$, (ii)$R_{1W}X_2$ and (iii) $\widehat{B}_2$. The first two steps are computationally simple, since they only involve subtracting means within categories for the first margin. This is equivalent to the computational step involved in forming the linearization variance estimator for a poststratified estimator. The most burdensome step is the computation of $\widehat{B}_2'$, but this is less burdensome that the computation of $\widehat{B}'$ directly, since the dimension of $\widehat{B}_2'$ is equal

to the number of columns in X_2, less than the dimension of $\widehat{B}'$, which is equal to the number of columns in X. Note that we may write

$$\widehat{B}'_2 = (X'_2 W R_{1W} X_2)^{-1} X'_2 W R_{1W} Y$$

so that the computation of $\widehat{B}'_2$ requires first forming the weighted least squares residuals of Y and of X_2 on X_1 and then computing the least squares regression of the first residuals on the second.

With two margins, the choice of which margins correspond to X_1 and X_2 in the above argument is arbitrary. In practice, it seems natural to select the margin with the greatest number of categories for X_1 in order to minimise the dimension of $\widehat{B}'_2$.

The above discussion could be extended to more than two margins. This requires taking the margins one by one. For each margin it is necessary to form the regression residuals of Y and the corresponding submatrix of X on all margins previously considered. In this way it is possible to reduce the computation of least squares coefficient vectors so that the dimensions of all these vectors are no greater than the largest number of categories for all the margins.

© http://www.DACSEIS.de

Chapter 4

Replication Variance Estimators

Let $\widehat{\theta}$ be an estimator of a scalar population parameter θ, based upon a set of raked weights w_i. The raking method may be any of those considered in Chapter 2. The estimator of the variance of $\widehat{\theta}$ for a broad class of replication methods is obtained by first constructing a set of T replicate weights $w_i^{(t)}$ for $t = 1, \ldots, T$, according to the replication method (Rust and Rao, 1996). Methods of construction are discussed below. For each set of replicate weights, an estimator $\widehat{\theta}^{(t)}$ of θ is then computed in the same way that $\widehat{\theta}$ is computed using the weights w_i. An estimator of the variance of $\widehat{\theta}$ is then given by

$$\widehat{V}(\widehat{\theta}) = \sum_{t=1}^{T} c_t \left(\widehat{\theta}^{(t)} - \widehat{\theta}\right)^2,$$

where c_t is a constant which depends on the replication method. For the bootstrap and balanced repeated replication $c_t = 1/T$. For the jackknife method, c_t is defined below.

The construction of the replicate weights $w_i^{(t)}$ involves first taking the initial weights ω_i. From these a set of initial replication weights $\omega_i^{(t)}, t = 1, \ldots, T$, is constructed according to the replication method and the sampling scheme. Next the raking adjustment method is applied to each of these T sets of weights separately. This generates the required weights $w_i^{(t)}$. This approach can be applied validly to a wide class of adjustment methods including classical raking, the 'maximum likelihood' raking, generalized raking and the modified method of Oh and Scheuren (1987) (see Rust and Rao, 1996).

To illustrate the construction of the initial replicate weights $\omega_i^{(t)}$, consider the jackknife method with stratified multistage sampling. The number of replicates in this case is $T = \sum n_h$, where n_h is the number of primary sampling units (PSUs) in stratum $h = 1, \ldots, H$ and H is the number of strata. Let replicate t correspond to deleting PSU j in stratum h and let the label i for the weight w_i correspond to element m in PSU l in stratum k. Then the replicate weights are given by

$$\begin{aligned} w_i^{(t)} &= w_i && \text{if } k \neq h \\ &= c_t w_i && \text{if } k = h, l \neq j \\ &= 0 && \text{if } k = h, l = j \end{aligned}$$

where $c_t = (n_h - 1)/n_h$, for $t = 1, \ldots, T$.

© http://www.DACSEIS.de

Appendix A

S-Plus ® Functions

- RAKING.WEIGHTS: This function computes the usual raking weights based on two margin variables.
- CALIBRATION.WEIGHTS: This function computes the usual OLS Calibration weights.
- CALIBRATION.RESIDUALS: This function computes the weighted residuals that are used for variance estimation. A variance estimator is a variance estimator of the sample total of these weighted residuals.
- REPLICATION.VARIANCE: This function computes a variance estimator based on a "general" replication variance estimator. This function needs the pseudo-values.

```
# FUNCTION RAKING.WEIGHTS
# ========================
#
# This function computes the usual raking weights.
#
# INPUT:
#       * Initial.Weights: the initial weights (a nx1 vector).
#       * Var.Margin.1: a vector describing the first margin
#         (a nx1 vector). This variable contains the labels
#         specifying the first margin.
#       * Var.Margin.2: a vector describing the second margin
#         (a nx1 vector). This variable contains the labels
#         specifying the second margin.
#       * Vect.Tot.Margin.1: the population totals of the first margin
#         (a nx1 vector).
#       * Vect.Tot.Margin.2: the population totals of the second margin
#         (a nx1 vector).
#       * Tolerance: the maximum value between the difference between
#         weights between two iterations.
#       * Max.Iter: the maximal number of iteration.
#         The iterative method stop after ''Max.Iter'' iterations.
#
# OUTPUT:
```

```
#       * The vector of final weights.
#

RAKING.WEIGHTS <-function(Initial.Weights,Var.Margin.1,Var.Margin.2,
                  Vect.Tot.Margin.1,Vect.Tot.Margin.2,Tolerance,Max.Iter) {

    Sample.Size <- length(Initial.Weights)


# FIND THE LABEL USED FOR THE FIRST MARGIN
    Table <- table(Var.Margin.1)
    Modalities.Margin.1 <- dimnames(Table)[[1]]
    mode(Modalities.Margin.1) <- ''numeric''
    NB.Modalities.Margin.1 <-length(Modalities.Margin.1)


# FIND THE LABEL USED FOR THE SECOND MARGIN
    Table <- table(Var.Margin.2)
    Modalities.Margin.2 <- dimnames(Table)[[1]]
    mode(Modalities.Margin.2) <- ''numeric''
    NB.Modalities.Margin.2 <-length(Modalities.Margin.2)


# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE FIRST MARGIN
    Mat.Margin.1 <-matrix(rep(0,times=NB.Modalities.Margin.1*Sample.Size),
                          ncol=NB.Modalities.Margin.1,nrow=Sample.Size)

    for(j in (1:NB.Modalities.Margin.1)) {
        Modalities <- Modalities.Margin.1[j]
        Mat.Margin.1[,j] <- as.numeric(Var.Margin.1 == Modalities)
    }


# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE SECOND MARGIN
    Mat.Margin.2 <-matrix(rep(0,times=NB.Modalities.Margin.2*Sample.Size),
                          ncol=NB.Modalities.Margin.2,nrow=Sample.Size)

    for(j in (1:NB.Modalities.Margin.2)) {
        Modalities <- Modalities.Margin.2[j]
        Mat.Margin.2[,j] <- as.numeric(Var.Margin.2 == Modalities)
    }


# COMPUTE THE RAKING WEIGHT BY THE ITERATIVE METHOD
    Final.Weights <- Initial.Weights
    Last.Weights <- Initial.Weights
    Converge <- FALSE
    k <- 1

    while(!Converge){


    # POSTSTRATIFICATION ON THE FIRST MARGIN
        Final.Weights <-CALIBRATION.WEIGHTS(Final.Weights,
        Mat.Margin.1,Vect.Tot.Margin.1)
```

© http://www.DACSEIS.de

```
    # POSTSTRATIFICATION ON THE SECOND MARGIN
        Final.Weights <-CALIBRATION.WEIGHTS(Final.Weights,
                        Mat.Margin.2,Vect.Tot.Margin.2)


    # TEST TO STOP THE ITERATION
        Converge <- (max(abs(Last.Weights-Final.Weights)) < Tolerance)
        Last.Weights <- Final.Weights
        k <- k + 1

        if(k>Max.Iter) {
            cat(''ERROR: The raking iteration do not converge after '',
                        Max.Iter,'' iterations.'')
            Converge <- TRUE
        }
    }


# OUTPUT
    Final.Weights
}
```

```
# FUNCTION CALIBRATION.WEIGHTS
# ============================
#
# This function computes the usual OLS Calibration weights.
#
# INPUT:
#       * Initial.Weights: the initial weights (a nx1 vector).
#       * Mat.Cal.Var: the sample value of the calibration
#         variables (a nxp matrix).
#       * Tot.Cal.Var: the population total of the calibration
#         variables (a nx1 vector).
#
# OUTPUT:
#       * The vector of final weights.
#

CALIBRATION.WEIGHTS <-function(Initial.Weights,Mat.Cal.Var,Tot.Cal.Var) {

    Sample.Size <- length(Initial.Weights)


# COMPUTE THE g WEIGHTS
    Diag.Pi <- diag(1/Initial.Weights)
    T.Mat.Cal.Var <- t(Mat.Cal.Var)
    Est.Tot.Cal.Var <- T.Mat.Cal.Var %*% Initial.Weights
    T.Vect.1 <-t(as.vector(rep(1,times=Sample.Size)))
    Calibration.Weights <- T.Vect.1 + t(Tot.Cal.Var-Est.Tot.Cal.Var) %*%
        GINVERSE(T.Mat.Cal.Var %*% Mat.Cal.Var) %*% T.Mat.Cal.Var %*
            % Diag.Pi


# COMPUTE THE FINAL WEIGHTS
    Final.Weights <- as.vector(Calibration.Weights * Initial.Weights)
```

```
# OUTPUT
    Final.Weights
}
```

```
# FUNCTION CALIBRATION.RESIDUALS
# ==============================
#
# This function computes the weighted residuals that are used for
# variance estimation by linearization. A variance estimator is
# a variance estimator of the sample total of these weighted residuals.
#
# INPUT:
#       * Vect.Y: the value of the study variable (a nx1 vector).
#       * Weights.of.Residuals: the weights given
#         to the residuals (a nx1 vector).
#       * Weights.for.Coef.Regression: the weights
#         used for the coeficient of regression (a nx1 vector).
#       * Mat.Cal.Var: the sample value of
#         the calibration variables (a nxp matrix).
#
# OUTPUT:
#       * A vector of weighted residuals.
#
# REMARK:
#       * The ''Weights.of.Residuals'' and the
#         ''Weights.for.Coef.Regression'' can be (i) the initial
#         weights of function ''CALIBRATION.WEIGHTS'' or (ii)
#         the final weights of function ''CALIBRATION.WEIGHTS''.
#       * The weights for ''Weights.for.Coef.Regression'' must
#         be positive.
#

CALIBRATION.RESIDUALS <-function(Vect.Y,Weights.of.Residuals,
    Weights.for.Coef.Regression,Mat.Cal.Var) {


# PUT THE DATA IN A SINGLE DATA FRAME
    Data <-data.frame(cbind(Vect.Y,Mat.Cal.Var))

# CREATE THE NAME AND THE FORMULA FOR THE WEIGHTED LEAST SQUARE FIT
    NB.Cal.Var <-as.numeric(dim(Mat.Cal.Var)[2])\qquad
    List.Name.Variables <- c(''Y'')
    Formula <- ''Y ~-1''

    for(i in (1:NB.Cal.Var)) {
        Name.New.Variable <-paste(''X'',i,sep='''')
        Formula <- paste(Formula,'' + '',Name.New.Variable,sep='''')
        List.Name.Variables <-c(List.Name.Variables,Name.New.Variable)
    }

    Formula <- as.formula(Formula)
    List.Blank <- dimnames(Data)[[1]]
    Names.List <-list(List.Blank,List.Name.Variables)
```

© http://www.DACSEIS.de

```
    dimnames(Data) <- Names.List


# THE WEIGHTED LEAST SQUARE FIT
    Model.Fit <- lm(formula=Formula,data=Data,
                    weights=Weights.for.Coef.Regression)


# THE WEIGHTED RESIDUALS
    Residuals <- residuals(Model.Fit) * Weights.of.Residuals

# OUTPUT
Residuals
}
```

```
# FUNCTION REPLICATION.VARIANCE
# ============================
#
# This function computes a variance estimator based on a ''general''
# replication variance estimator. This function needs the pseudo-values.
#
# INPUT:
#       * Pseudo.values: the Pseudo-values (a kx1 vector).
#       * Scale.Factor.for.Variance: a vector containing the
#         scale factor.for the variance (a kx1 vector).
#
# OUTPUT:
#       * The variance estimate.
#
# REMARK:
#       * The i-th pseudo-values are defined by T(i) - T;
#         where T is the point estimate and T(i) is the
#         point estimate based upon the i-th replicate
#         i = 1,... ,k.
#       * The scale factors are constants used tom scale
#         the variance estimator. They depend on the
#         replication method.
#

REPLICATION.VARIANCE <-function(Pseudo.values,Scale.Factor.for.Variance) {

    Variance <-sum(Scale.Factor.for.Variance*(Pseudo.values)^2)

    Variance

}
```

© http://www.DACSEIS.de

References

Bankier, M. D. (1986): Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association* **81**, 1074–1079.

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

Brackstone, G. J. and Rao, J. N. K. (1976): Raking ratio estimators. *Survey Methodology* **2**, 63–69.

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

Choudhry, G. H. and Lee, H. (1987): Variance estimation for the Canadian Labour Force Survey. *Survey Methodology* **13**, 147–161.

Deming, W. E. and Stephan, F. F. (1940): On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**, 427–444.

Demnati, A. and Rao, J. N. K. (2004): Linearization variance estimators for survey data. to appear in: Survey Methodology.

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.

Fuller, W. A. (2002): Regression estimating for survey samples. *Survey Methodology* **28**, 5–23.

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

Konijn, H. S. (1981): Biases, variances and covariances of raking ratio estimators for marginal and cell totals and averages of observed characteristics. *Metrika* **28**, 109–121.

Oh, H. L. and Scheuren, F. (1987): Modified raking ratio estimation. *Survey Methodology* **13**, 209–219.

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.

Seber, G. A. F. (1977): *Linear Regression Analysis*. New York: Wiley.

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