A Variance Estimation R-package for Repeated Surveys – Useful for Estimates of Changes in Quarterly and Annual Averages

Øyvind LANGSRUD (Oyvind.Langsrud@ssb.no)
Statistics Norway

ABSTRACT

This paper presents a newly developed R-package for calculation of variances of estimates based on data from several waves of a repeated survey with partly overlapping samples. Development of the package is a part of on-going work on quality improvements of the Labour Force Survey in Norway, which is quarterly and based on a rotating panel. The package can, for example, be used to calculate variances of net changes of annual averages of unemployment rates for persons aged 20-64. The methodology is based on linearly calibrated weights (as calculated by the packages ReGenesees and survey) and residuals from the corresponding regression modelling. These computations may be done separately for each wave. The functionality is generic and the user can specify any calibration model and any linear combination of (quarterly) estimates. Linearization is used to calculate variances of rates. The main method assumes that all relevant population totals can be computed from register data, but situations where totals are unknown for some of the calibration variables are also handled.

Keywords: *Calibration weighting, Rotating panel survey, Nonresponse adjustment, Model-based, Design-based, Multivariate regression, Linearization*

JEL Classification: *C10, C88*

INTRODUCTION

This paper accompanies the R-package, CalibrateSSB (Langsrud, 2016), which calculates variances of (change) estimates based on data from several waves of a repeated survey with partly overlapping samples. Within each wave the data are weighted by linear calibration. Finding the weights can be viewed as the first of three important estimation tasks. The second task is to establish a covariance matrix for the total estimates. The third task is to calculate variances of linear combinations of these estimates, such as mean changes.

The starting point for the package development was to implement generically the variance and covariance formulas in Hamre and Heldal (2013) which are based on Hagesæther and Zhang (2007). These estimates are design based estimates under some asymptotical assumptions and they are calculated from the residuals according to the regression models corresponding to the calibration. A closely related design based approach is recommended in Osier et al. (2013) and described in Berger and Priam (2016).They use residuals from a model which includes covariates which specify the stratification and interactions which specify the rotation of the sampling designs.

However, a main approach of the present paper is not design based and we call it model based. We will calculate a robust or empirical covariance matrix which is also known as a sandwich estimator (Kauermann and Carroll, 2001). Since this deviate from the classical parametric approach, this is also sometimes referred to as not model based. Multivariate linear regression modelling is an important part of the description below. The formulation by matrix notation is very similar to how this is implemented in the R-package.

Ordinary linear calibration and weighted linear regression are equivalent and Section 2 describes calibration as a model based regression technique. Sections 3 and 4 formulate robust model based estimators of the covariance matrix. The corresponding covariance matrix according to Hamre and Heldal (2013) are given in Section 5. Their design based estimate is very similar to the model based estimate. Finding the covariance matrix of linear combinations is described in Section 6. How to handle ratios by linearization is treated in the same section. As written in Section 7, cluster-robust variants of the covariance matrix estimates can be easily obtained. Section 8 treats the situation where some of the population totals of the calibration variables are estimated. This is partly based on the design based methodology of Särndal and Lundström (2005), but a difference is that we assume that all the calibration variables are individually known in the gross sample. A multivariate generalization of their estimate of variance due to non-response is proposed. This is used to adjust the covariance matrix estimates in order to take into account the presence of estimated population totals.

LINEAR CALIBRATION AND MULTIVARIATE REGRESSION

We assume that the population data follow a multivariate multiple regression model defined by

$$\boldsymbol{Y_U} = \boldsymbol{X_U B} + \boldsymbol{E_U} \tag{1}$$

where the $N \times m$ matrix $\boldsymbol{Y_U}$ consists of m outcome variables of interest. The auxiliary matrix $\boldsymbol{X}$ consists of p linearly independent columns and thus the matrix $\boldsymbol{B}$ of regression parameters is of size $p \times m$. The rows of the residual matrix $\boldsymbol{E_U}$ are independent with zero mean and a common covariance matrix, $\boldsymbol{\Sigma}$.

We assume that the first column of $\boldsymbol{X_U}$ is the constant vector of 1's. Below such a constant vector is written as **1**. The population totals can then be written as $\boldsymbol{T} = \boldsymbol{Y_U}^T\mathbf{1}$. Now, assume that the outcome variables are observed in a response set, r. According to the ordinary regression estimates the transposed total vector based on predictions are

$$\hat{T}^T = \mathbf{1}^T\widehat{\boldsymbol{Y_U}} = \underbrace{\overbrace{\mathbf{1}^T\boldsymbol{X_U}}^{\text{x-totals}}\ \overbrace{(\boldsymbol{X_r}^T\boldsymbol{X_r})^{-1}\boldsymbol{X_r}^T}^{\text{Estimate of } \boldsymbol{B}}}_{\boldsymbol{w}}\boldsymbol{Y_r} \tag{2}$$

These totals are weighted sums of the observations in the response set and $\boldsymbol{w}$ are these linearly calibrated weights. The calibration equation is satisfied: $\boldsymbol{w}^T\boldsymbol{X_r} = \mathbf{1}^T\boldsymbol{X_U}$. Above (2) the observed part of $\boldsymbol{Y_U}$ was replaced by predicted values. The totals will, however, be unaffected by such a replacement when a constant term is included in the model. The residuals sum to zero.

The standard theory of calibration involves sampling weights and the calibrated weights can then be written as

$$\boldsymbol{w}^T = \mathbf{1}^T\boldsymbol{X_U}(\boldsymbol{X_r}^T\boldsymbol{D}\boldsymbol{X_r})^{-1}\boldsymbol{X_r}^T\boldsymbol{D} \tag{3}$$

where $\boldsymbol{D}$ is a diagonal matrix of sampling weights. This means that ordinary regression is replaced by weighted regression. To justify weighted regression under the population model (1) , the covariance matrix assumption needs to include the inverse sampling weights as proportionality constants. This time the weighted (sampling weights) sum of the residuals is zero.

In practice, (near) collinearity may need to be handled and the calibrated weights can then be computed as

$$\boldsymbol{w}^T = \mathbf{1}^T\boldsymbol{X_U}(\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{X_r})^{-}\boldsymbol{D}^{\frac{1}{2}} \tag{4}$$

where "-" denotes a generalized inverse.

THE COVARIANCE MATRIX OF TOTAL ESTIMATES

We consider the total estimate

$$\hat{T} = \boldsymbol{Y_r}^T\boldsymbol{w} \tag{5}$$

Under the finite sampling approach the relevant covariance matrix estimate is

$$\mathrm{Cov}\left(\hat{T}-T\right)=\mathrm{Cov}\left(\hat{T}-\boldsymbol{Y_r}^T\mathbf{1}\right)+\mathrm{Cov}\left(T-\boldsymbol{Y_r}^T\mathbf{1}\right) \tag{6}$$

Under the above model (1) it turns out that

$$\mathrm{Cov}\left(\hat{T}-T\right)=\left(\boldsymbol{w}^T(\boldsymbol{w}-\mathbf{1})\right)\cdot\boldsymbol{\Sigma} \tag{7}$$

where "·" is included to indicate scalar multiplication. By plugging in an estimate of **Σ** we obtain

$$\widehat{\mathrm{Cov}}\left(\hat{T}-T\right)=\frac{1}{n_r-p}\left(\boldsymbol{w}^T(\boldsymbol{w}-\mathbf{1})\right)\cdot\widehat{\boldsymbol{E_r}}^T\widehat{\boldsymbol{E_r}} \tag{8}$$

where $\widehat{\boldsymbol{E_r}}=\boldsymbol{Y_r}-\widehat{\boldsymbol{Y_r}}$ is the matrix of residuals from the regression modelling underlying the calibration. To relax the assumption of a common covariance matrix we will consider a robust or empirical covariance matrix estimate which can be said to be a sandwich type estimator:

$$\widehat{\mathrm{Cov}}\left(\hat{T}\right)=\left(\boldsymbol{W}\odot\widehat{\boldsymbol{E_r}}\right)^T\left(\boldsymbol{W}\odot\widehat{\boldsymbol{E_r}}\right) \tag{9}$$

Here ⊙ denotes element-wise multiplication and $\boldsymbol{W}=\boldsymbol{w}\mathbf{1}^T$. That is all columns of $\boldsymbol{W}$ are equal and $\hat{\boldsymbol{T}}$ can be expressed as (11) below. This equation means that a individual covariance matrix estimate is based on each row of $(\boldsymbol{W}\odot\widehat{\boldsymbol{E_r}})$. The final estimate is obtained by simply summing these individual estimates together. This is consequence of independent rows. The case of dependence within clusters is treated in Section 7.

An estimate of Cov($\hat{\boldsymbol{T}}$-T) can be obtained by replacing $\boldsymbol{W}$ by ($\boldsymbol{W}$ - **1**) where **1** is a matrix of ones. Then, (9) is a multivariate generalization of the robust variance estimate described in Valliant et al. (2000). The last component of (6) is then considered as negligible. But instead of neglecting this component we instead suggest the estimate

$$\widehat{\mathrm{Cov}}\left(\hat{T}-T\right)=\left(\boldsymbol{W}\odot\widehat{\boldsymbol{E_r}}\right)^T\left((\boldsymbol{W}-\mathbf{1})\odot\widehat{\boldsymbol{E_r}}\right) \tag{10}$$

which turns out when the last component of (6) is estimated from the observed residuals by using weights (more details below).

The residuals may be replaced by adjusted residuals as described below.

THE COVARIANCE MATRIX IN PANEL SURVEYS

Now we will generalize the above methodology to panel surveys and the total estimate is written as

$$\hat{\boldsymbol{T}}=(\boldsymbol{W}\odot\boldsymbol{Y_r})^T\,\mathbf{1} \tag{11}$$

Each row of $\boldsymbol{Y_r}$ represents a sampling unit and each column is a variable from a specific wave. Thus, two columns of $\boldsymbol{Y_r}$ may be the same variable from two different waves. These are calibrated separately and the weights are different. Unlike above, all columns of $\boldsymbol{W}$ are not equal. Furthermore, the samples may not be completely overlapping and therefore many elements of $\boldsymbol{Y_r}$ are not observed. We solve this by setting the corresponding elements of $\boldsymbol{W}$ to zero. For the computations below we also set the corresponding elements of $\widehat{\boldsymbol{E_r}}$ to zero.

Since missing residuals have zero weights we can still make a robust estimate of $\hat{\boldsymbol{T}}$ according to (9). Even if the data set is complicated, this estimate is very simple. The problem of partly overlapping samples is handled indirectly.

Again a possible estimate of Cov($\hat{\boldsymbol{T}}$-T) can be obtained by replacing $\boldsymbol{W}$ by ($\boldsymbol{W}$ - $\mathbf{1}$). Equation (10) cannot be used when the columns of $\boldsymbol{W}$ differ. Instead we propose an estimate which is consistent with (10):

$$\begin{aligned}\widehat{\text{Cov}} &= \widehat{\text{Cov}}\left(\hat{T}-T\right) = F\left(\boldsymbol{W},\widehat{\boldsymbol{E_r}}\right)\\ &= \left((\boldsymbol{W}-\mathbf{1})\odot\widehat{\boldsymbol{E_r}}\right)^T\left((\boldsymbol{W}-\mathbf{1})\odot\widehat{\boldsymbol{E_r}}\right)\\ &+ \left(\sqrt{\boldsymbol{W}-\mathbf{1}}\odot\widehat{\boldsymbol{E_r}}\right)^T\left(\sqrt{\boldsymbol{W}-\mathbf{1}}\odot\widehat{\boldsymbol{E_r}}\right)\end{aligned} \tag{12}$$

Negative elements (caused by zero weights) are set to zero before the element-wise square root operation. The last row act as an estimate of the last component of (6). To calculate the covariance between two variables (numbered 1 and 2), we here use $(w_{i1}-1)(w_{i2}-1)+\sqrt{(w_{i1}-1)(w_{i2}-1)}$ as a substitute for $w_i(w_i-1)$ which is used when the weights are equal (10). More specifically, this means that outside the sample (last component of (6)), the covariance between variables 1 and 2 are estimated as a weighted mean of the corresponding residual products with weights $\sqrt{(w_{i1}-1)(w_{i2}-1)}$. Furthermore, the size of the overlapping population outside the sample is assumed to be the sum of these weights. Unless the weights are equal, this is in practice a conservative (towards zero) estimate of the covariance. Have in mind that we are now only discussing a component of the variance which is negligible when the sampling fraction is small. The notation, $\widehat{\text{Cov}}$, is introduced since the estimate is an alternative to similar design based estimates below. The function name, F, is introduced since it is needed later in this paper.

To avoid downward bias caused by model fitting one can replace the residuals by adjusted residuals as described in Valliant et al. (2000):

$$\widehat{\boldsymbol{E_r}} \rightarrow \widehat{\boldsymbol{E_r}} \oslash \sqrt{\mathbf{1}-\boldsymbol{H_{ii}}} \tag{13}$$

where $\oslash$ denotes element-wise division and again the square root is also element-wise. The matrix $\boldsymbol{H_{ii}}$ consists of the diagonal elements of the so-called hat matrix from the regression calculation underlying the calibration and they are also known as the leverages. In the situation of Section 3 all columns of $\boldsymbol{H_{ii}}$ are equal (similar to $\boldsymbol{W}$). To be more robust against model misspecification one can drop the square root:

$$\widehat{\boldsymbol{E_r}} \rightarrow \widehat{\boldsymbol{E_r}} \oslash (\mathbf{1} - \boldsymbol{H_{ii}}) \tag{14}$$

The resulting residuals are exactly those obtained by leave-one-out cross-validation (Shao, 1993) and as written by Valliant et al. (2000) it is guaranteed that the corresponding robust variance estimate is not biased downwards.

A DESIGN BASED COVARIANCE MATRIX

First we define an observed unobserved indicator matrix, $\boldsymbol{J}$, with the same dimensions as $\boldsymbol{Y_r}$. Then, the matrix $\boldsymbol{n} = \boldsymbol{J^T J}$ has ordinary sample sizes on the diagonal and otherwise sample sizes of overlaps. The row vector of mean weighted residuals can be written as

$$\overline{\mathit{WE}} = \mathbf{1}^T \left(\boldsymbol{W} \odot \widehat{\boldsymbol{E_r}}\right) \oslash \mathrm{diag}(\boldsymbol{n}) \tag{15}$$

where diag($\boldsymbol{n}$) consists of the diagonal elements of $\boldsymbol{n}$.

The formulas of variances and covariances in Hamre and Heldal (2013) can now be summarized as

$$\widehat{\mathrm{Cov}} = \left(\boldsymbol{n} \oslash (\boldsymbol{n} - \mathbf{1})\right) \odot \left(\left(\boldsymbol{W} \odot \widehat{\boldsymbol{E_r}}\right)^T \left(\boldsymbol{W} \odot \widehat{\boldsymbol{E_r}}\right) - \boldsymbol{n} \odot \overline{\mathit{WE}}^T \, \overline{\mathit{WE}}\right) \tag{16}$$

The variances and covariances of differences and ratios which are described in Hamre and Heldal (2013) are in accordance with the description in the section below.

LINEAR COMBINATIONS AND RATIOS

To calculate the covariance matrix of linear combinations we make use of the property

$$\mathrm{Cov}(\boldsymbol{M}Z) = \boldsymbol{M}\mathrm{Cov}(Z)\,\boldsymbol{M^T} \tag{17}$$

where Z is a random vector and where $\boldsymbol{M}$ is a matrix of coefficients for the linear combinations. In the four-element case we choose $\boldsymbol{M} = [-1,-1, 1, 1]/2$ to calculate the difference between the mean of the two first and the two last values. Several linear combinations are obtained when $\boldsymbol{M}$ has several rows.

In order to calculate variances of ratios and covariances between ratios we use a first order Taylor series (delta method) substitute:

$$\frac{A}{C} \rightarrow \left(\frac{A}{c} - \frac{aC}{c^2}\right) \tag{18}$$

where uppercase means stochastic variables and lower case means observed values. Hence when going from a vector of ordinary totals to a vector ratios, $1/c$ and $-a/c^2$ are the coefficients needed to define the linear combination used to create the covariance matrix.

When $Z = [Z_1,Z_2,Z_3,Z_4]^T$ we obtain the covariance of $[Z_1/Z_2,Z_3/Z_4]^T$ by using

$$\boldsymbol{M} = \begin{bmatrix} 1/z_2 & -z_1/z_2^2 & 0 & 0 \\ 0 & 0 & 1/z_4 & -z_3/z_4^2 \end{bmatrix} \tag{19}$$

In practice we may compute the covariance matrix by utilizing (17) in three steps starting with $Z = \hat{\boldsymbol{T}}$:

1. Linear combinations of totals.
2. Ratios of these linear combinations.
3. Linear combinations of these ratios.

CLUSTER-ROBUST ESTIMATION

The above discussion assumes independent individuals or random sampling of individuals (not families). However, clusters (families) might have been neglected and we may calculate the variance in a way that is robust against model misspecification of this kind (Cameron and Miller, 2015).

If e_1 and e_2 are two independent residuals we estimate

$$\widehat{\text{Var}}(w_1e_1 + w_2e_2) = w_1^2\hat{e}_1^2 + w_2^2\hat{e}_2^2 \tag{20}$$

If we cannot assume independence we can instead estimate this as

$$\widehat{\text{Var}}(w_1e_1 + w_2e_2) = (w_1\hat{e}_1 + w_2\hat{e}_2)^2 \tag{21}$$

In other words we can say that (21) is a cluster-robust alternative to (20). This is similar for several variables and covariance matrices. We can use this to make a cluster-robust alternative to (9). We simply replace the matrix $(\boldsymbol{W} \odot \boldsymbol{E_r})$ by a matrix with fewer rows. In the new matrix the rows are summed within clusters. We can use a similar technique to construct a cluster-robust variant of (12).

ALLOWING ESTIMATED POPULATION TOTALS

Above non-response was not treated explicitly. Calibration was performed based on the net sample only. Now we will consider the situation were some population totals are unavailable, but the corresponding x-variables are individually available in the gross sample. In this case we suggest to calculate the weights, $\boldsymbol{w}$, according to this expression of $\hat{\boldsymbol{T}}^{\boldsymbol{T}}$:

$$\mathbf{1}^T\widehat{\boldsymbol{Y}_{\boldsymbol{U}}} = \boldsymbol{w}^T\boldsymbol{Y}_{\boldsymbol{r}} = \overbrace{\underbrace{\mathbf{1}^T\widetilde{\boldsymbol{X}}_{\boldsymbol{U}}(\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}{}^T\widetilde{\boldsymbol{X}}_{\boldsymbol{s}})^{-1}\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}{}^T}_{\widetilde{\boldsymbol{w}}}\hphantom{\boldsymbol{X}_{\boldsymbol{s}}}}^{\text{Estimated x-totals}}\underbrace{\mathllap{\boldsymbol{X}_{\boldsymbol{s}}}\,\overbrace{(\boldsymbol{X}_{\boldsymbol{r}}{}^T\boldsymbol{X}_{\boldsymbol{r}})^{-1}\boldsymbol{X}_{\boldsymbol{r}}{}^T}^{\text{Estimate of }\boldsymbol{B}}\,\boldsymbol{Y}_{\boldsymbol{r}}}_{\widehat{\boldsymbol{Y}_{\boldsymbol{s}}}} \qquad (22)$$

Here $\boldsymbol{X}_{\boldsymbol{s}}$ contains data for the whole gross sample and $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}$ is the same matrix except that the last *k* columns are omitted. These omitted columns correspond to the unavailable population totals. The matrix $\widetilde{\boldsymbol{X}}_{\boldsymbol{U}}$ is the corresponding population matrix. One way of viewing the difference between this expression and expression (2) is that ordinary x-totals are replaced by estimates. These estimates are found from population predictions based on regressing ***Xs*** on $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}$. Then, for the first *p-k* variables, the estimated and observed totals are identical. A second viewpoint is to say that the estimated x-totals are obtained by using the weights, $\widetilde{\boldsymbol{w}}$, which are calculated by calibration from the gross sample to the population by using $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}$. Equation (22) can also be interpreted a third way. The y-totals are found by weighting the predicted gross sample ($\widehat{\boldsymbol{Y}_{\boldsymbol{s}}}$) by the weights $\widetilde{\boldsymbol{w}}$. We can generalize (22) to include diagonal matrices of design weights so that ordinary regressions are replaced by weighed regressions.

This procedure is closely related to the recommended procedure in Särndal and Lundström (2005). The difference is that they estimate the unavailable population totals by using design weights instead of $\widetilde{\boldsymbol{w}}$. Then all the x-variables do not need to be individually known in the gross sample. However they mention: "*If available, a better estimate (but unbiased or nearly so) is allowed to take place of* [the design weighted total]." Also note that calibration according to (22) generalized to incorporate design weights corresponds to the method in Estevao and Särndal (2002) referred to as the one that "*use the complete auxiliary information*".

In order to estimate variances in this case we can calculate two matrices of residuals.

$$\widehat{\boldsymbol{E}_{\boldsymbol{r}}^{*}} \sim \text{from regressing } \boldsymbol{Y}_{\boldsymbol{r}} \text{ on } \boldsymbol{X}_{\boldsymbol{r}} \qquad (23)$$

$$\widehat{\boldsymbol{E}_{\boldsymbol{r}}} \sim \text{from regressing } \boldsymbol{Y}_{\boldsymbol{r}} \text{ on } \widetilde{\boldsymbol{X}}_{\boldsymbol{r}} \qquad (24)$$

Here $\widetilde{\boldsymbol{X}}_{\boldsymbol{r}}$ is the net sample variant of $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}$. It is not straightforward to generalise the model based robust covariance matrix above (12) to estimates based on (22). If we use $\widehat{\boldsymbol{E}^*_{\boldsymbol{r}}}$ we do not take into account that the estimated population totals are uncertain. If we use $\widehat{\boldsymbol{E}_{\boldsymbol{r}}}$ we do not take into account calibration towards the estimated totals.

In the case of a single y-variable, Särndal and Lundström (2005) have described a design based variance estimate which uses both types of residuals. Although, their residuals are calculated from weighted regressions based on the finally calibrated weights instead of the design weights (then $\overline{\boldsymbol{WE}}$ in (16) vanish). When all population totals are available, $\widehat{\boldsymbol{E}^*_{\boldsymbol{r}}} = \widehat{\boldsymbol{E}_{\boldsymbol{r}}}$ and the univariate variance estimate of Särndal and Lundström is very similar to estimates obtain by (12) and (16).

Their formula consists of two parts. The sampling variance component (using $\widehat{\boldsymbol{E}_{\boldsymbol{r}}}$) and the nonresponse variance component (using $\widehat{\boldsymbol{E}^*_{\boldsymbol{r}}}$). The latter part can be expressed as this sum over the response set:

$$\hat{\mathrm{V}}_{\mathrm{NR}} = \sum v_i(v_i - 1)(d_i e_i^*)^2 \tag{25}$$

were d_i are design weights and $v_i = w_i/d_i$. Here we cannot omit design weights by setting them to one. However, above we used $\widetilde{\boldsymbol{w}}$ instead of design weights and we will also do this here. Then it is ensured that $\hat{\mathrm{V}}_{\mathrm{NR}} = 0$ in the case of no non-response.

We can obtain a multivariate generalization by

$$\widehat{\mathrm{Cov}}_{\mathrm{NR}} = F\left(\boldsymbol{W} \oslash \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}} \odot \widehat{\boldsymbol{E}^*_{\boldsymbol{r}}}\right) \tag{26}$$

with F as defined in (12) and where $\widetilde{\boldsymbol{W}}$ contain weights according to $\widetilde{\boldsymbol{w}}$. With no non-response $\widetilde{\boldsymbol{W}} = \boldsymbol{W}$ and $\widehat{\mathrm{Cov}}_{\mathrm{NR}} = 0$. In the special case where the gross sample is the population all relevant elements of $\widetilde{\boldsymbol{W}}$ are one and $\widehat{\mathrm{Cov}}_{\mathrm{NR}}$ becomes equivalent to the covariance estimate in (12). In cases where the columns of $\boldsymbol{W}$ are equal, (26) is a natural generalization of (25) with $d_i = \widetilde{w}_i$ and $v_i = w_i/\widetilde{w}_i$. The tricky part is covariances between variables with different weights. Similar to the text in Section 4, $(v_{i1} - 1)(v_{i2} - 1) + \sqrt{(v_{i1} - 1)(v_{i2} - 1)}$ is used as a substitute for $v_i(v_i - 1)$. Now, the difference between v_i and $(v_i - 1)$ is substantial.

The present paper does not intend to generalize the sampling variance component in Särndal and Lundström (2005). Instead we will use the non-response component (26) to adjust the estimates in (12) and (16) which is calculated from $\widehat{\boldsymbol{E}_{\boldsymbol{r}}}$. Especially the adjusted variant of (12) can be expressed as

$$\widehat{\mathrm{Cov}} = F\left(\boldsymbol{W}, \widehat{\boldsymbol{E}_{\boldsymbol{r}}}\right) - F\left(\boldsymbol{W} \oslash \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}} \odot \widehat{\boldsymbol{E}_{\boldsymbol{r}}}\right) + F\left(\boldsymbol{W} \oslash \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{W}} \odot \widehat{\boldsymbol{E}^*_{\boldsymbol{r}}}\right) \tag{27}$$

The underlying idea is that the unadjusted estimate is similar to the estimate of Särndal and Lundström in the case were $\widehat{\boldsymbol{E}_{\boldsymbol{r}}^{*}} = \widehat{\boldsymbol{E}_{\boldsymbol{r}}}$. This means that the original estimate contains a non-response component based on $\widehat{\boldsymbol{E}_{\boldsymbol{r}}}$, but this component should have been based on $\widehat{\boldsymbol{E}_{\boldsymbol{r}}^{*}}$ instead. The estimate in (16) can be adjusted the same way and we can then say that the final estimate is design based.

THE R PACKAGE – THREE MAIN FUNCTIONS

The methodology described above is implemented in the R package CalibrateSSB. The description below contains the most important input and output for three main functions in this package. The calculations can also be performed by one single function. See the package documentation for more details.

CalibrateSSB

Input:

Gross sample data with specified y-variables, the calibration model, population data or population totals, variables defining domains for calibration, specification of external package, possible sampling weights.

Output:

Calibrated weights, residuals, leverages.

The function computes calibrated weights. One alternative is to base all computations on the package ReGenesees (Zardetto, 2015). Then, unavailable population totals are not allowed and leverages are not computed. Weights ($\boldsymbol{w}$) can alternatively be computed by the package survey (Lumley, 2014) or according to (4) without any external package. The unavailable totals are estimated by the method of calibration from the gross sample as described in Section 8. The weights, $\widetilde{\boldsymbol{w}}$, are also included in the output object together with two types of residuals and leverages according to (23) and (24).

WideFromCalibrate

Input:

Output object from CalibrateSSB including the original y-data, variables defining the panel waves, sampling unit identifier, variables that split the data into estimation domains, possible extra variables.

Output:

Reorganised data.

This function reorganises the data so that a matrix is created for each variable. The rows represent the sampling units (persons) and it is one column for each wave. Furthermore, the data may be split into sub-datasets. Possible extra variables can be included. I practice this can be used to include a clustering variable (families).

PanelEstimation
Input:
Output object from WideFromCalibrate, possible specification of numerator and denominator, matrix defining linear combinations, estimation type, leverage power, possible clustering variable.
Output:
Estimates and corresponding variances.

This function performs all calculations separately within each estimation domain as specified when running WideFromCalibrate. Initially this function calculates a covariance matrix according to the theory above. The estimation type parameter determines whether this is based on (9), (12), (12) without the last row, (16) or a cluster robust variant.

The residuals can be adjusted by leverages by using a nonzero leverage power, 1/2 when (13) and 1 when (14). When the input contains two types of residuals (caused by unavailable population totals), adjustment according to (27) is performed. When numerator and denominator are not specified, linear combinations and corresponding variances are calculated directly according to (17) with $\boldsymbol{M}$ taken from input. When numerator and denominator are specified an extra round of initial computations of ratios are performed. The covariance matrix for the ratios is calculated as described in Section 6. Additional functions to compute the linear combination matrix are supplied so that various changes and mean changes can be computed easily.

SUMMARY

CalibrateSSB is an R-package that handles repeated surveys with partially overlapping samples. Initially the samples are weighted by linear calibration using known or estimated population totals. A robust model based covariance matrix for all relevant estimated totals is calculated from the residuals according to the calibration model. Alternatively a design based covariance matrix is calculated in a very similar way. A cluster robust version is also possible. In the case of estimated populations totals the covariance matrix is adjusted by utilizing the theory of Särndal and Lundström (2005). Variances of linear combinations (changes and averages) and ratios are calculated from this covariance matrix. The linear combinations and ratios can involve variables within and/or between sample waves. In summary, various estimates based on data from several waves are calculated with standard errors.

References

1. **Berger, Y. G.** and **Priam, R.**, 2016, *"A simple variance estimator of change for rotating repeated surveys: an application to the European Union Statistics on Income and Living Conditions household surveys"*, J. R. Statist. Soc. A, Vol. 179, Part 1, pp. 251-272.
2. **Cameron, A. C.** and **Miller, D. L.**, 2015, *"A Practitioner's Guide to Cluster-Robust Inference, Journal of Human Resources"*, Vol. 50, No. 2, pp. 317-373.
3. **Estevao, V. M.** and **Särndal, C.-E.**, 2002, *"The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling"*, Journal of Official Statistics, Vol. 18, No. 2, pp. 233-255.
4. **Hagesæther, N.** and **Zhang, L.-C.**, 2007, *"Om estimeringsusikkerhet og utvalgsplan i AKU. Notater2007/22, Statistics Norway"* (Norwegian only), ***https://www.ssb.no/a/publikasjoner/pdf/notat_200722/notat_200722.pdf***
5. **Hamre, J.** og **Heldal, J.**, 2013, *"Improved calculation and dissemination of coefficients of variation in the Norwegian LFS"*, Documents 46/2013, Statistics Norway. ***http://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/_attachment/148090?_ts=142476a8ad0***
6. **Kauermann, G.** and **Carroll, R . J.**, 2001, *"A Note on the Efficiency of Sandwich Covariance Matrix Estimation"*, Journal of the American Statistical Association, Vol. 96, No. 456, pp. 1387-1396.
7. **Langsrud, Ø.**, 2016, *"CalibrateSSB: Variance estimation for repeated surveys with partially overlapping samples"*, R package intended to be published on CRAN.
8. **Lumley, T**, 2014, *"Survey: analysis of complex survey samples"*, R package version 3.30, CRAN. ***http://r-survey.r-forge.r-project.org/survey/***
9. **Osier, G., Berger, Y. G.** and **Goedemé, T.**, 2013, *"Standard Error Estimation for the EU-SILC Indicators of Poverty and Social Exclusion"*, Eurostat Methodologies and Working Papers series. ***http://ec.europa.eu/eurostat/documents/3888793/5855973/KS-RA-13-024-EN.PDF***
10. **Särndal, C.-E.** and **Lundström, S.**, 2005, *Estimation in Surveys with Nonresponse*, John Wiley and Sons, New York.
11. **Shao, J.**, 1993, *"Linear Model Selection by Cross-Validation"*, Journal of the American Statistical Association, Vol. 88, No. 422, pp. 486-494.
12. **Vaillant, R., Dorfman, A. H.,** and **Royall, R. M.** ,2000, *Finite Population Sampling and Inference, a Prediction Approach*. Wiley, New York.
13. **Zardetto, D.**, 2015, *ReGenesees: "R Evolved Generalized Software for Sampling Estimates and Errors"*, R package version 1.7, Istat. ***http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/regenesees***