DACSEIS
IST-2000-26057

Workpackage 7

A simulation study of repeated weighting estimation

Deliverable 7.3

2004

List of contributors:

Harm Jan Boonstra; Statistics Netherlands

Main responsibility:

Harm Jan Boonstra; Statistics Netherlands

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/`

Preface

A simulation study is carried out to investigate the performance of repeated weighting estimators and corresponding variance estimators. The study concerns two three-way frequency tables for a population of persons to be estimated from a sample matched to a register. Repeated weighting estimates are general regression estimates adjusted to certain marginal totals in order to achieve numerical consistency, in this case with register counts. For the particular frequency tables studied, repeated weighting and corresponding variance estimation perform well. Even for moderately small samples, mean squared errors are comparable to those for the general regression estimator that serves as starting point for the repeated weighting estimator. For the second frequency table simulated, a simplified variance estimator for the repeated weighting estimator is computed, which performs nearly as well as the original repeated weighting variance estimator.

Harm Jan Boonstra Statistics Netherlands

Contents

List of Tables

Chapter 1

Introduction

Weighting is a technique often used by statistical offices to reduce non-response bias and sampling error in sample surveys. Weighting usually employs population information available from a register to accomplish this. However, it is a well-known problem that the amount of such information that can be employed in a single weighting scheme is limited to an extent dependent on the sample size. Using too much auxiliary information results in unstable estimates due to the large number of regression coefficients that have to be estimated from the sample data at hand.

The problem just sketched is encountered when one attempts to estimate a large set of tables from several data sources in such a way that all estimates are mutually consistent, i.e. such that common margins of any two estimated tables are perfectly equal. In order to obtain consistent estimates, a procedure of repeated weighting was introduced in Renssen *et al.* (2001). There it is proposed to first weight the sample using an overall weighting scheme and then verify for each particular target table to be estimated whether it is consistent with all register counts and previously estimated tables. If it is not, a second weighting or adjustment step is performed which corrects this. The overall weighting scheme serves the purpose of non-response bias and sampling variance reduction, and using the resulting weights some subset of tables can usually already be estimated consistently. Re-weighting where necessary, one obtains consistency of the complete set of estimated tables.

Repeated weighting procedures and variance estimation for the corresponding estimators have been further discussed in Houbiers *et al.* (2003) and Boonstra *et al.* (2003). Here we continue the discussion of the latter paper by carrying out a simulation study to investigate the performance of the repeated weighting estimators and their variance estimators. We focus attention on the case where consistency must be obtained for estimates based on a single survey matched to a perfect register.

Estimates and variance estimates for two three-way frequency tables are simulated under simple random sampling for sample sizes $n = 500$ and $n = 5,000$. We use two sample sizes in our simulations since repeated weighting estimators, as well as corresponding variance estimators, are likely to be more vulnerable to smaller samples. This expectation is based on the fact that repeated weighting estimates usually involve many estimated regression coefficients and possibly also estimated population totals. For reference, we also simulate direct estimates and general regression estimates based on the overall weighting scheme.

The examples studied are somewhat contrived in the sense that, at least for $n = 5,000$, both overall weighting scheme and repeated weighting scheme could have been applied simultaneously using a single weighting scheme, presumably giving better results. However, we should keep in mind that repeated weighting is eligible for practical situations where there are many tables to be estimated, possibly not all known in advance, and in which consistency between registers and surveys cannot ordinarily be imposed by a single weighting scheme per survey.

The next chapter describes the simulation setup. Chapters 3 and 4 describe the simulations carried out on two different target tables. Chapter 5 contains conclusions. Tables of simulation results are collected in Appendix A. In Appendix B some aspects of the method of computation are explained, and Appendix C lists some S-Plus code used for the computation of the repeated weighting estimates and their variance estimates.

Chapter 2

Simulation setup

The population used is based on data of the province Noord-Brabant taken from the Dutch pseudo-universe created for the DACSEIS project, see MÜNNICH and SCHÜRLE (2003). It consists of $N = 188,216$ persons. The following characteristics of these persons are available:

1. MUN: municipality in 70 categories
2. $MUN10$: municipality in 10 categories
3. SEX: sex in 2 categories
4. AGE: age in 6 categories
5. MST: marital status in 3 categories
6. $ETHN$: ethnicity in 3 categories
7. $EMPL$: employment in 3 categories

For the simulations, $ETHN$ and $EMPL$ are considered to be the variables observed only in the sample, the remaining variables are register variables. Register variables appear as classification variables in the target tables. Some of them are also used in an initial overall weighting: the overall weighting scheme adopted is $SEX \times MUN10 + AGE$.

Simple random samples of sizes $n = 500$ and $n = 5,000$ are taken from this population. Averages for point estimates are computed for 15,000 simulation runs. For 3,000 of the 15,000 simulation runs several variance estimates are computed along with the point estimates. The 15,000 simulation runs are also used to obtain an approximation to the sampling variance for these point estimates. The variance approximations thus obtained are then used as if they were the true variances in the computation of bias and mean squared error for the variance estimators.

Performance criteria that we computed are relative bias (RB) and relative root mean squared error (RRMSE). For the point estimates these are defined as

$$\mathrm{RB}(\widehat{\theta}) = \frac{\overline{\widehat{\theta}} - \theta}{\theta}, \qquad \mathrm{RRMSE}(\widehat{\theta}) = \frac{\sqrt{\overline{\left(\widehat{\theta} - \theta\right)^2}}}{\theta}, \tag{2.1}$$

where a bar denotes the average over the simulation runs. For the variance estimates RB and RRMSE are defined similarly as

$$\mathrm{RB}(v(\widehat{\theta})) = \frac{\overline{v(\widehat{\theta})} - \nu(\widehat{\theta})}{\nu(\widehat{\theta})}, \qquad \mathrm{RRMSE}(v(\widehat{\theta})) = \frac{\sqrt{\overline{\left(v(\widehat{\theta}) - \nu(\widehat{\theta})\right)^2}}}{\nu(\widehat{\theta})}, \tag{2.2}$$

where $v(\widehat{\theta})$ is a variance estimator for $\widehat{\theta}$, and $\nu(\widehat{\theta})$ is its true sampling variance. We also give the coverage $C(\widehat{\theta}, v(\widehat{\theta}))$ for several combinations of point and variance estimators, defined as the percentages of simulated intervals $\left[\widehat{\theta} - z_{1-\alpha/2}\sqrt{v(\widehat{\theta})}\,,\ \widehat{\theta} + z_{1-\alpha/2}\sqrt{v(\widehat{\theta})}\right]$ containing the true parameter θ, where $z_{1-\alpha/2}$ is a standard normal quantile, which we take at $\alpha = 0.05$.

All variables used in the simulation study are categorical. This means that we only simulate tables of counts. The results are reported in what in S-Plus is called treatment contrast parametrization: from each categorical variable the first category is discarded and replaced by the intercept, i.e. the variable that takes the value 1 for each population unit. So, for example, the table $SEX \times EMPL$ consisting of the cells $(SEX0{+}SEX1)(EMPL0{+}EMPL1{+}EMPL2)$ is reparametrized into the cells obtained by expanding the product $(1 + SEX1)(1 + EMPL1 + EMPL2)$. This parametrization serves two purposes. First, it avoids redundancy in the re-weighting schemes consisting of two or more weighting terms. Second, it allows easy inspection of the consistency of repeatedly weighted tables.

Chapter 3

Table $SEX \times AGE \times ETHN$

The first simulation is carried out on the target table of counts $SEX \times AGE \times ETHN$ (abbreviated as SAE), in which $ETHN$ is considered as the variable observed only in the sample, and the other variables are registered for the whole population. The overall weighting scheme used is $SEX \times MUN10 + AGE$.

The direct or Horvitz-Thompson (HT) estimator of the vector of population counts t_{SAE} is given by

$$\widehat{t}^{d}_{SAE} = \sum_{i \in S} d_i y_i \ = N \overline{y}_S \,, \tag{3.1}$$

where y is a vector of dummy variables representing the target table's cells, d_i are design weights which equal N/n for simple random sampling, and $\overline{y}_S$ is the sample mean of y.

The HT estimator serves as starting point for the general regression estimator which uses an overall weighting scheme to improve on it. The general regression estimator can be written as a weighted sum

$$\widehat{t}^{w}_{SAE} = \sum_{i \in S} w_i y_i \tag{3.2}$$

with regression weights w_i calibrated to the known population totals of the overall weighting scheme.

The general regression estimator on its turn is the starting point for the repeated weighting (RW) estimator. According to the so-called splitting-up procedure described in Houbiers *et al.* (2003) and Boonstra *et al.* (2003), the weighting scheme used for the RW estimator consists of all m complete $(m-1)$-way margins $\Gamma^-_1, \Gamma^-_2 \ldots \Gamma^-_m$ of the m-way table Γ to be estimated. In general then, the RW estimator is given by

$$\widehat{t}^{RW}_{\Gamma} = \widehat{t}^{w}_{\Gamma} + (\widehat{B}^{w}_{\Gamma;\Gamma^-_1+\cdots+\Gamma^-_m})^t \begin{pmatrix} \widehat{t}^{RW}_{\Gamma^-_1} - \widehat{t}^{w}_{\Gamma^-_1} \\ \vdots \\ \widehat{t}^{RW}_{\Gamma^-_m} - \widehat{t}^{w}_{\Gamma^-_m} \end{pmatrix} , \tag{3.3}$$

where $\widehat{t}^{w}_{\Gamma}$ is the general regression estimator of Γ, $\widehat{B}^{w}_{\Gamma;\Gamma^-_1+\cdots+\Gamma^-_m}$ is the matrix of weighted least squares regression coefficients for Γ in a regression on $\Gamma^-_1, \ldots \Gamma^-_m$, and $\widehat{t}^{w}_{\Gamma^-_j}$ are general

regression estimators for the margins. From the appearance of RW estimators for the marginal totals on the right hand side of (3.3) it is clear that this is a recursive definition of repeated weighting. However, the repeated weighting estimator for say Γ_j^- reduces to a table $t_{\Gamma_j^-}$ of register totals if Γ_j^- is available from the register and it equals $\widehat{t}^w_{\Gamma_j^-}$ if the latter is already consistent with the register. Additional details and a general description of repeated weighting estimation for multiple surveys can be found in BOONSTRA *et al.* (2003).

Returning to the target table $SEX \times AGE \times ETHN$, we can write its RW estimator as

$$\widehat{t}^{RW}_{SAE} = \widehat{t}^w_{SAE} + (\widehat{B}^w_{SAE;SA+SE+AE})^t \begin{pmatrix} t_{SA} - \widehat{t}^w_{SA} \\ 0 \\ 0 \end{pmatrix}, \tag{3.4}$$

where $\widehat{t}^w_{SA}$ is the $SEX \times AGE$ margin of $\widehat{t}^w_{SAE}$, and t_{SA} represents the register totals of $SEX \times AGE$. The zeroes between the brackets in (3.4) are due to the fact that the corresponding margins for $SEX \times ETHN$ and $AGE \times ETHN$ estimated using the overall weighting scheme weights are already consistent with the register.

An alternative repeated weighting estimator using only a minimal weighting scheme, i.e. the minimal re-weighting scheme such that consistency with the register is obtained, is

$$\widehat{t}^{MW}_{SAE} = \widehat{t}^w_{SAE} + (\widehat{B}^w_{SAE;SA})^t (t_{SA} - \widehat{t}^w_{SA}) \,. \tag{3.5}$$

This estimator only calibrates $\widehat{t}^w_{SAE}$ to $SEX \times AGE$ register totals and unlike (3.4) does not preserve the $SEX \times ETHN$ and $AGE \times ETHN$ margins of $\widehat{t}^w_{SAE}$.

We carried out two simulations, with 15,000 simulation runs each, the first with sample size $n = 500$ and the second with sample size $n = 5,000$. In Table A.1 in Appendix A simulation averages based on 15,000 simulation runs with $n = 500$ are displayed for the estimators (3.1) to (3.5). The first column holds the names of a complete set of the table's cells. In the second column the true population counts are given. The remaining columns list both simulation mean and simulation relative root mean squared error (RRMSE). Table A.2 contains similar results for the case of $n = 5,000$.

The following observations can be made from Tables A.1 and A.2.

- The mean squared error of the regression estimator vanishes for the SEX and AGE margins due to their presence in the overall weighting scheme. By comparing the RRMSE of the HT estimator $\widehat{t}^d_{SAE}$ and the regression estimator $\widehat{t}^w_{SAE}$, we observe that the auxiliary information used in the overall weighting scheme $SEX \times MUN10 + AGE$ improves the $SEX \times AGE$ margin as expected. However, it does not improve other cells shown in the table.

- The repeated weighting estimators $\widehat{t}^{RW}_{SAE}$ and $\widehat{t}^{MW}_{SAE}$ are calibrated for consistency with the register to $SEX \times AGE$, which explains why the corresponding marginal cells have zero RRMSE.

- The repeated weighting estimator $\widehat{t}^{RW}_{SAE}$ is the same as the regression estimator for all one and two-dimensional margins containing $ETHN$. This is due to the presence of $SEX \times ETHN$ and $AGE \times ETHN$ in the re-weighting scheme.

- Except for margins that are calibrated for some estimators, differences between the estimators are very small. In particular, the difference in performance of the two repeated weighting estimators $\widehat{t}_{SAE}^{RW}$ and $\widehat{t}_{SAE}^{MW}$ seems negligible.

Differences between Table A.1 ($n = 500$) and Table A.2 ($n = 5,000$) other than a reduction of the RRMSE by about a factor of $\sqrt{10}$ are small. Whereas for $n = 500$ the HT estimator seems to have slightly smaller MSE than the other estimators, for $n = 5,000$ it is the other way around. This is not surprising as for small sample sizes the use of auxiliary information can affect estimates adversely. The differences here are very small, however.

So already for a sample size of $n = 500$ the repeated weighting estimators for the Table $SEX \times AGE \times ETHN$ do not perform significantly worse according to the mean squared error criterion than the traditional estimators $\widehat{t}_{SAE}^{d}$ and $\widehat{t}_{SAE}^{w}$, while having the advantage of being consistent with register counts. This observation is even more encouraging when we note that some of the cells of the target table have expected cell sizes less than one under simple random sampling with $n = 500$. Moreover, several cells in the re-weighting scheme have very few observations or none at all in some simulation runs. To obtain consistency in the presence of empty cells in the re-weighting scheme, we have used slightly adjusted regression coefficients. This is explained in Appendix B. The adjusted regression coefficients are more stable against fluctuations in small samples and this appears to work quite well here.

Approximate sampling variances can be computed from the 15,000 point estimates. They are used to evaluate the performance of the following variance estimators. The variance estimates for the HT estimates $\widehat{t}_{SAE}^{d}$ are given by the diagonal elements of the estimated covariance matrix

$$v^{(d)} = \frac{N(N-n)}{n}\, s_y^2\,, \tag{3.6}$$

where $s_y^2 \equiv \frac{1}{n-1}\sum_{i\in S}(y_i - \overline{y})(y_i - \overline{y})^t$ denotes the sample covariance matrix of y.

For the regression estimator, the usual linearization variance estimates with and without g-weights are computed. They are given by the diagonal elements of

$$v^{(w)} = \frac{N(N-n)}{n}\, s_e^2\,, \tag{3.7}$$

and

$$v_g^{(w)} = \frac{N(N-n)}{n}\, s_{ge}^2\,, \tag{3.8}$$

where s_e^2 is the sample covariance matrix of the residuals e_k of the target vector-variable y with respect to the overall weighting scheme and s_{ge}^2 is the sample covariance matrix of the products $g_k e_k$, where g_k are the g-weights corresponding to the overall weighting scheme.

For the repeated weighting estimator $\widehat{t}_{SAE}^{RW}$, the linearization variance estimates are the diagonal elements of

$$v^{RW} = \frac{N(N-n)}{n}\, s_{e^{RW}}^2\,, \tag{3.9}$$

where

$$e_k^{RW} = e_k + (\widehat{B}^w_{SAE;SA+SE+AE})^t \begin{pmatrix} -e_{SA;k} \\ 0 \\ 0 \end{pmatrix}, \tag{3.10}$$

with $e_{SA;k}$ the $SEX \times AGE$ margins of e_k. For $\widehat{t}^{MW}_{SAE}$ no specific variance estimates are computed. The variance estimates are computed in 3,000 of the 15,000 simulation runs. In Table A.3 relative bias and relative root mean squared error are listed for the variance estimates for the simulations with $n = 500$. The results for $n = 5,000$ can be found in Table A.4.

The following observations are made from Tables A.3 and A.4.

- For one and two-dimensional margins involving $ETHN$ $v^{(w)}$ and v^{RW} are equal as are the point estimates $\widehat{t}^w_{SAE}$ and $\widehat{t}^{RW}_{SAE}$ whose variance they estimate. This consistency property does not hold when we take $v_g^{(w)}$ as a variance estimator for $\widehat{t}^w_{SAE}$. Perhaps it is possible to incorporate g-weights in the variance estimator for $\widehat{t}^{RW}_{SAE}$ such that it is consistent with $v_g^{(w)}$.

- All variance estimators considered are clearly downward biased, except of course $v^{(d)}$ which is known to be unbiased. The least biased is $v_g^{(w)}$, the variance estimator including the g-factors. The variance estimators $v^{(w)}$ and v^{RW} have nearly equal downward bias with v^{RW} only slightly more biased. The biases of $v^{(w)}$, $v_g^{(w)}$ and v^{RW} are significantly smaller for $n = 5,000$ than they are for $n = 500$. Indeed, these variance estimators are asymptotically unbiased.

- Although $v_g^{(w)}$ is less biased than $v^{(w)}$, its MSE is somewhat higher, with the notable exception of margin $SEX \times AGE$. For this margin we have seen from Tables A.1 and A.2 that the use of auxiliary information improves the point estimates and apparently the use of the same auxiliary information in the variance estimates by means of incorporating g-weights improves the corresponding variance estimates as well.

- Despite their negative bias, variance estimators $v^{(w)}$ and v^{RW} have smaller RRMSE than $v^{(d)}$ for the HT estimator, for almost all cells of the target table. The RRMSE of v^{RW} seems slightly smaller than the RRMSE of $v^{(w)}$. So there are no signs that variance estimation is less stable for repeated weighting estimation, even not for the rather small sample size of $n = 500$.

- The orders of magnitude of the RRMSE of the variance estimators are the same as those of the corresponding point estimators (for the same sample size), especially for the HT estimator.

The main difference between $n = 500$ and $n = 5,000$ is that the MSEs for the latter are about a factor $\sqrt{10}$ smaller. What might be surprising is that the simulation bias of the HT variance estimator is not noticeably smaller for $n = 5,000$ than for $n = 500$. However, this is because $v^{(d)}$ is unbiased and the displayed bias is due only to the simulation error resulting from the finite number (15,000) of simulation runs used to approximate the true

sampling variances and the finite number $(3,000)$ of simulation runs used for variance estimation.

In Table A.5 coverage rates are displayed for the approximate 95% confidence intervals formed by the point estimators and corresponding variance estimators. Both $n = 500$ and $n = 5,000$ rates are displayed and they are based on the $3,000$ simulation runs in which both point and variance estimators were computed.

As expected, the coverage is better for the larger sample size simulation. By comparing with the cell counts displayed in the second column of Table A.1 we see that the coverage depends mainly on the expected sample size of a cell. There is hardly any difference between the coverage of the different estimator variance estimator combinations. For $n = 500$ the coverage of the HT estimator and variance estimator is slightly better than for the other combinations, presumably due to the negative bias of the variance estimators that rely on linearization.

Chapter 4

Table $SEX \times MST \times EMPL$

The second simulation is carried out on the target table $SEX \times MST \times EMPL$ (abbreviated by SME), where $EMPL$ is considered as the variable observed only in the sample, and the other variables are register variables. The overall weighting scheme is again $SEX \times MUN10 + AGE$.

The repeated weighting estimator of the totals t_{SME} is

$$\widehat{t}^{RW}_{SME} = \widehat{t}^{w}_{SME} + (\widehat{B}^{w}_{SME;SM+SE+ME})^t \begin{pmatrix} t_{SM} - \widehat{t}^{w}_{SM} \\ 0 \\ \widehat{t}^{RW}_{ME} - \widehat{t}^{w}_{ME} \end{pmatrix}, \tag{4.1}$$

where $\widehat{t}^{w}_{SME}$ is the regression estimator of t_{SME} based on the overall weighting scheme, t_{SM} are the register totals of $SEX \times MST$, and

$$\widehat{t}^{RW}_{ME} = \widehat{t}^{w}_{ME} + (\widehat{B}^{w}_{ME;M+E})^t \begin{pmatrix} t_{M} - \widehat{t}^{w}_{M} \\ 0 \end{pmatrix}. \tag{4.2}$$

As can be seen from (4.1) and (4.2), this example involves a recursive use of the repeated weighting estimator, i.e. one of the margins is calibrated on totals estimated themselves by a repeated weighting estimator. In more complicated cases this recursion can be deeper than just one level.

In Section 6 of Boonstra *et al.* (2003) a non-recursive approximation to the RW estimator was proposed. In general, this approximation does not necessarily satisfy all consistency requirements. However, its main function is to provide a simpler variance estimator for the RW estimator, as we shall see below. Here the approximation of the RW estimator takes the form

$$\widehat{t}^{SRW}_{SME} = \widehat{t}^{w}_{SME} + (\widehat{B}^{w}_{SME;SM+SE})^t \begin{pmatrix} t_{SM} - \widehat{t}^{w}_{SM} \\ 0 \end{pmatrix}, \tag{4.3}$$

the re-weighting being only with respect to the margins $SEX \times MST$ and $SEX \times EMPL$, i.e. those margins that do not need re-weighting themselves. We shall refer to this estimator as the simplified RW estimator.

Finally, the minimal weighting scheme to adjust $\widehat{t}^{w}_{SME}$ to the register is $SEX \times MST$. The corresponding estimates are given by

$$\widehat{t}^{MW}_{SME} = \widehat{t}^{w}_{SME} + (\widehat{B}^{w}_{SME;SM})^t (t_{SM} - \widehat{t}^{w}_{SM}) \,. \tag{4.4}$$

The three repeated weighting estimators described above as well as the HT estimator and regression estimator based on the overall weighting scheme are again computed for simple random samples of sizes $n = 500$ and $n = 5,000$ for each of 15,000 simulation runs. The results are displayed in Tables A.6 and A.7 for $n = 500$ and $n = 5,000$, respectively. Relative biases are found to be small with no great differences between the different estimators and therefore are not displayed. Only RRMSEs are listed besides a column with the sizes of the target table's cells.

From Tables A.6 and A.7 it is clear that the mean squared errors of all three repeated weighting estimators $\widehat{t}^{RW}_{SME}$, $\widehat{t}^{SRW}_{SME}$ and $\widehat{t}^{MW}_{SME}$ are nearly the same. An obvious difference with the regression estimator $\widehat{t}^{w}_{SME}$ is that for (SEX)MST margins all three repeated weighting estimators have zero MSE since they are calibrated with respect to the corresponding marginal register counts. Also, cells involving $MST1EMPL2$, i.e. $MST1EMPL2$ and $SEXMST1EMPL2$, are clearly improved by repeated weighting. Cells involving $MST2EMPL2$ are slightly improved. Smaller cells, involving $EMPL1$, do not benefit significantly from the repeated weighting. The use of auxiliary information is more effective for the larger cells than it is for the smaller cells, which is not really surprising.

Five different variance estimators have been computed for a subset of 3,000 simulation runs, one for the HT estimator, two for the regression estimator $\widehat{t}^{w}_{SME}$ and two for the repeated weighting estimator $\widehat{t}^{RW}_{SME}$. For the HT and regression estimators the variance estimates computed are of the same form as given in (3.6), (3.7) and (3.8).

For the repeated weighting estimator $\widehat{t}^{RW}_{SME}$, the linearization variance estimates are the diagonal elements of

$$v^{RW} = \frac{N(N-n)}{n} \, s^2_{e^{RW}} \,, \tag{4.5}$$

where

$$e^{RW}_k = e_k + (\widehat{B}^{w}_{SME;SM+SE+ME})^t \begin{pmatrix} -e_{SM;k} \\ 0 \\ e^{RW}_{ME;k} - e_{ME;k} \end{pmatrix} \,, \tag{4.6}$$

$e_{SM;k}$ are the $SEX \times MST$ margins of the overall weighting scheme regression residuals e_k, and

$$e^{RW}_{ME;k} = e_{ME;k} + (\widehat{B}^{w}_{ME;M+E})^t \begin{pmatrix} -e_{M;k} \\ 0 \end{pmatrix} \,. \tag{4.7}$$

A simplified variance estimator can be based on the estimator $\widehat{t}^{SRW}_{SME}$. The linearization variance estimates for the latter are the diagonal elements of

$$v^{SRW} = \frac{N(N-n)}{n} \, s^2_{e^{SRW}} \,, \tag{4.8}$$

where

$$e_k^{SRW} = e_k + (\widehat{B}^w_{SME;SM+SE})^t \begin{pmatrix} -e_{SM;k} \\ 0 \end{pmatrix} . \tag{4.9}$$

We consider v^{SRW} as a variance estimator for the original repeated weighting estimator $\widehat{t}^{RW}_{SME}$. Variance estimation for $\widehat{t}^{SRW}_{SME}$ and $\widehat{t}^{MW}_{SME}$ has not been simulated.

Tables A.8 and A.9 contain simulation results of the variance estimators, where sampling variances have been approximated as the variances of the 15,000 sets of point estimates. From these tables it can be seen that variance estimators for repeated weighting estimators have clearly greater negative bias than does the regression estimator. Globally speaking, the hierarchy of relative bias appears to be $|RB(v^{SRW})| > |RB(v^{RW})| > |RB(v^{(w)})| > |RB(v_g^{(w)})|$.

Note that the repeated weighting variance estimators perform (slightly) better, in MSE sense, than HT and regression variance estimators for the smallest cells $(SEX)MST1EMPL1$ and $(SEX)MST2EMPL1$. It might then come as a surprise that for the much larger cell $(SEX)MST1EMPL2$, repeated weighting variance estimators have about twice as high a *relative* RMSE than HT and regression variance estimators. However, this surprise will not be so great after noting from Tables A.6 and A.7 that the repeated weighting point estimators for this cell perform clearly better than HT and regression estimators, the latter having about twice as large root mean squared errors. Therefore, the decrease in accuracy is only relative; the absolute mean squared errors of regression, HT and repeated weighting variance estimators are apparently not very different.

The differences between the full linearization variance estimator v^{RW} and its simplified version v^{SRW} are small, except that v^{SRW} seems to perform somewhat worse for the cell $(SEX)MST1EMPL2$. So we can cautiously conclude that v^{SRW} is a reasonable simple alternative for v^{RW}.

The coverage (nominal value 95%) of the various estimator/variance estimator pairs is listed in Table A.10 for $n = 500$ and $n = 5,000$. No great differences are discernible other than the improved coverage rates for small cells when $n = 5,000$.

Chapter 5

Conclusions

For the simulations carried out we have seen that differences between several repeated weightings (complete, minimal, simplified) are small. In almost all cases the repeated weighting estimates perform as well as or (slightly) better than the regression estimator based on the overall weighting scheme, even for as small a sample size as $n = 500$. The main advantage of repeated weighting estimators is their numerical consistency, in our case numerical consistency with all register counts. We found no signs for any stability problems of repeated weighting point and variance estimators, which one might have expected on the grounds that repeated weighting estimators involve more estimated regression coefficients and also estimated calibration totals. Our conclusions also broadly agree with findings of a previous simulation study of repeated weighting (Van Duin and Snijders, 2003 and Houbiers *et al*., 2003). In addition, we found that although variance estimators for repeated weighting may be somewhat more negatively biased than for the regression estimator, this is compensated by smaller variances, resulting in mean squared errors that are comparable to those of the variance estimators of the regression estimator. Finally, the simplified variance estimator computed in the second simulation study seems a reasonable and simple alternative for the full linearization variance estimator of the repeated weighting estimator.

Appendix A

Simulation results

Table A.1: Average point estimates and RRMSE (in %) over 15,000 runs for table $SEX \times AGE \times ETHN$ under simple random sampling with $n = 500$. The second column contains population totals.

Cell name	t_{SAE}	$\widehat{t}^{d}_{SAE}$	$\widehat{t}^{w}_{SAE}$	$\widehat{t}^{RW}_{SAE}$	$\widehat{t}^{MW}_{SAE}$
Intercept	188216	188216 / 0	188216 / 0	188216 / 0	188216 / 0
SEX	97493	97509 / 4.32	97493 / 0	97493 / 0	97493 / 0
AGE1	34935	34960 / 9.46	34935 / 0	34935 / 0	34935 / 0
AGE2	35151	35128 / 9.30	35151 / 0	35151 / 0	35151 / 0
AGE3	28781	28762 / 10.5	28781 / 0	28781 / 0	28781 / 0
AGE4	28884	28856 / 10.4	28884 / 0	28884 / 0	28884 / 0
AGE5	33591	33634 / 9.61	33591 / 0	33591 / 0	33591 / 0
ETHN1	4425	4426.1 / 28.8	4427.8 / 29.6	4427.8 / 29.6	4450.8 / 29.7
ETHN2	9235	9214.1 / 19.7	9211.1 / 20.1	9211.1 / 20.1	9229.7 / 20.2
SEXAGE1	17592	17632 / 14.0	17618 / 9.60	17592 / 0	17592 / 0
SEXAGE2	17765	17745 / 13.9	17751 / 9.59	17765 / 0	17765 / 0
SEXAGE3	14239	14232 / 15.5	14238 / 10.9	14239 / 0	14239 / 0
SEXAGE4	14493	14464 / 15.4	14472 / 10.8	14493 / 0	14493 / 0
SEXAGE5	20486	20512 / 12.8	20496 / 7.87	20486 / 0	20486 / 0
SEXETHN1	2719	2715.0 / 36.9	2714.0 / 37.5	2714.0 / 37.5	2724.2 / 37.7
SEXETHN2	4408	4398.3 / 28.8	4396.8 / 29.1	4396.8 / 29.1	4404.8 / 29.2
AGE1ETHN1	1069	1068.1 / 59.1	1067.2 / 59.8	1067.2 / 59.8	1070.9 / 60.0
AGE2ETHN1	909	908.3 / 63.8	908.6 / 64.5	908.6 / 64.5	911.3 / 64.7
AGE3ETHN1	567	566.9 / 81.7	568.6 / 83.4	568.6 / 83.4	572.5 / 83.7
AGE4ETHN1	629	633.8 / 78.2	634.3 / 79.7	634.3 / 79.7	638.6 / 80.0
AGE5ETHN1	847	845.2 / 65.6	844.3 / 66.6	844.3 / 66.6	847.6 / 66.7
AGE1ETHN2	2169	2153.9 / 41.6	2149.9 / 41.4	2149.9 / 41.4	2152.7 / 41.6
AGE2ETHN2	2225	2225.3 / 40.7	2227.5 / 40.7	2227.5 / 40.1	2230.2 / 40.9
AGE3ETHN2	1221	1218.5 / 55.1	1220.2 / 55.6	1220.2 / 55.6	1223.7 / 55.9
AGE4ETHN2	1178	1176.1 / 56.4	1176.9 / 56.9	1176.9 / 56.9	1179.8 / 57.0
AGE5ETHN2	975	971.6 / 62.0	970.4 / 62.6	970.4 / 62.6	974.3 / 62.9
SA1ETHN1[1]	698	699.7 / 73.6	697.9 / 74.5	697.4 / 74.2	698.7 / 74.4
SA2ETHN1	524	521.8 / 83.9	521.8 / 85.5	522.6 / 85.0	523.6 / 85.1
SA3ETHN1	321	319.2 / 109	319.9 / 111	320.2 / 111	322.0 / 111
SA4ETHN1	270	271.5 / 117	271.6 / 119	272.3 / 119	274.3 / 120
SA5ETHN1	664	661.9 / 74.3	661.4 / 75.3	660.1 / 74.9	661.9 / 75.0
SA1ETHN2	1143	1130.6 / 56.6	1128.5 / 57.0	1127.9 / 56.5	1128.3 / 56.5
SA2ETHN2	1000	1003.1 / 61.2	1003.3 / 61.7	1003.8 / 61.2	1005.2 / 61.3
SA3ETHN2	624	624.7 / 77.8	625.6 / 79.1	626.4 / 78.7	627.5 / 78.9
SA4ETHN2	560	557.8 / 82.2	557.9 / 83.6	558.4 / 83.0	559.9 / 83.0
SA5ETHN2	508	507.6 / 85.4	508.4 / 87.3	506.9 / 87.1	509.0 / 87.0

[1] SA $\widehat{=}$ SEXAGE

Table A.2: Average point estimates and RRMSE (in %) over 15,000 runs for table $SEX \times AGE \times ETHN$ under simple random sampling with $n = 5,000$. The second column contains population totals.

Cell name	t_{SAE}	$\widehat{t}^{d}_{SAE}$	$\widehat{t}^{w}_{SAE}$	$\widehat{t}^{RW}_{SAE}$	$\widehat{t}^{MW}_{SAE}$
Intercept	188216	188216 / 0	188216 / 0	188216 / 0	188216 / 0
SEX	97493	97503 / 1.36	97493 / 0	97493 / 0	97493 / 0
AGE1	34935	34936 / 2.95	34935 / 0	34935 / 0	34935 / 0
AGE2	35151	35141 / 2.91	35151 / 0	35151 / 0	35151 / 0
AGE3	28781	28775 / 3.30	28781 / 0	28781 / 0	28781 / 0
AGE4	28884	28885 / 3.29	28884 / 0	28884 / 0	28884 / 0
AGE5	33591	33601 / 3.01	33591 / 0	33591 / 0	33591 / 0
ETHN1	4425	4427.8 / 8.94	4428.1 / 8.95	4428.1 / 8.95	4428.4 / 8.95
ETHN2	9235	9230.8 / 6.12	9231.3 / 6.13	9231.3 / 6.13	9231.5 / 6.13
SEXAGE1	17592	17600 / 4.39	17598 / 2.92	17592 / 0	17592 / 0
SEXAGE2	17765	17761 / 4.34	17763 / 2.91	17765 / 0	17765 / 0
SEXAGE3	14239	14228 / 4.87	14229 / 3.34	14239 / 0	14239 / 0
SEXAGE4	14493	14495 / 4.81	14494 / 3.25	14493 / 0	14493 / 0
SEXAGE5	20486	20490 / 3.99	20483 / 2.40	20486 / 0	20486 / 0
SEXETHN1	2719	2720.1 / 11.5	2719.9 / 11.4	2719.9 / 11.4	2720.1 / 11.4
SEXETHN2	4408	4406.1 / 8.99	4405.7 / 8.90	4405.7 / 8.90	4405.6 / 8.89
AGE1ETHN1	1069	1068.8 / 18.6	1068.6 / 18.4	1068.6 / 18.4	1068.6 / 18.4
AGE2ETHN1	909	910.0 / 19.9	909.4 / 19.8	909.4 / 19.8	909.5 / 19.8
AGE3ETHN1	567	568.2 / 25.3	568.4 / 25.1	568.4 / 25.1	568.5 / 25.1
AGE4ETHN1	629	628.5 / 24.1	628.7 / 24.0	628.7 / 24.0	628.7 / 24.0
AGE5ETHN1	847	849.3 / 20.7	849.1 / 20.6	849.1 / 20.6	849.3 / 20.6
AGE1ETHN2	2169	2167.9 / 12.9	2167.8 / 12.6	2167.8 / 12.6	2167.8 / 12.6
AGE2ETHN2	2225	2226.6 / 12.8	2227.4 / 12.5	2227.4 / 12.5	2227.5 / 12.5
AGE3ETHN2	1221	1220.1 / 17.2	1220.4 / 17.0	1220.4 / 17.0	1220.4 / 17.0
AGE4ETHN2	1178	1174.9 / 17.6	1174.8 / 17.3	1174.8 / 17.3	1174.8 / 17.3
AGE5ETHN2	975	976.0 / 19.2	975.9 / 19.0	975.9 / 19.0	975.9 / 19.0
SA1ETHN1[2]	698	697.5 / 22.8	697.3 / 22.6	697.1 / 22.4	697.1 / 22.4
SA2ETHN1	524	522.6 / 26.4	522.9 / 26.3	523.0 / 26.2	523.0 / 26.2
SA3ETHN1	321	321.5 / 33.8	321.5 / 33.7	321.8 / 33.6	321.8 / 33.6
SA4ETHN1	270	270.2 / 36.8	270.2 / 36.8	270.1 / 36.6	270.2 / 36.6
SA5ETHN1	664	666.0 / 23.4	665.9 / 23.2	666.0 / 23.2	666.1 / 23.2
SA1ETHN2	1143	1141.9 / 17.8	1141.8 / 17.6	1141.7 / 17.4	1141.5 / 17.4
SA2ETHN2	1000	1000.9 / 19.2	1000.9 / 19.0	1001.1 / 18.8	1001.1 / 18.8
SA3ETHN2	624	623.8 / 24.4	623.8 / 24.2	624.3 / 24.0	624.2 / 24.0
SA4ETHN2	560	558.0 / 25.3	557.9 / 25.2	557.8 / 25.0	557.8 / 25.0
SA5ETHN2	508	509.2 / 26.7	509.1 / 26.6	509.2 / 26.6	509.2 / 26.5

[2] SA $\widehat{=}$ SEXAGE

Table A.3: Relative bias (in %) and RRMSE (in %) of variance estimates over 3,000 simulation runs under simple random sampling with $n = 500$. A '*' indicates that bias and mean squared error vanish.

Cell name	$v^{(d)}(\widehat{t}^{d}_{SAE})$	$v^{(w)}(\widehat{t}^{w}_{SAE})$	$v_g^{(w)}(\widehat{t}^{w}_{SAE})$	$v^{RW}(\widehat{t}^{RW}_{SAE})$
Intercept	*	*	*	*
SEX	-0.42 / 0.59	*	*	*
AGE1	-2.4 / 7.5	*	*	*
AGE2	0.37 / 7.2	*	*	*
AGE3	-0.11 / 8.6	*	*	*
AGE4	1.8 / 8.9	*	*	*
AGE5	-0.75 / 7.4	*	*	*
ETHN1	-1.0 / 28.4	-10 / 27.8	-6.1 / 30.5	-10 / 27.8
ETHN2	-0.08 / 18.6	-9.3 / 19.3	-5.0 / 20.5	-9.3 / 19.3
SEXAGE1	-1.2 / 12.4	-11 / 13.0	-5.0 / 8.4	*
SEXAGE2	-1.5 / 12.4	-12 / 13.8	-6.4 / 9.2	*
SEXAGE3	1.5 / 14.3	-10 / 13.0	-4.2 / 8.8	*
SEXAGE4	1.6 / 14.5	-10 / 13.2	-4.9 / 9.3	*
SEXAGE5	-0.79 / 11.3	-9.1 / 12.0	-3.1 / 8.8	*
SEXETHN1	-1.2 / 36.1	-10 / 33.9	-6.4 / 37.8	-10 / 33.9
SEXETHN2	0.21 / 28.0	-9.0 / 26.4	-4.6 / 29.6	-9.0 / 26.4
AGE1ETHN1	0.06 / 58.5	-9.3 / 52.6	-4.9 / 58.5	-9.3 / 52.6
AGE2ETHN1	-0.24 / 65.4	-9.0 / 59.1	-5.0 / 66.6	-9.0 / 59.1
AGE3ETHN1	-0.11 / 80.6	-10.5 / 71.9	-4.7 / 84.5	-10.5 / 71.9
AGE4ETHN1	-2.7 / 76.2	-12.6 / 68.3	-9.7 / 76.7	-12.6 / 68.3
AGE5ETHN1	0.50 / 66.6	-8.6 / 60.0	-4.0 / 68.7	-8.6 / 60.0
AGE1ETHN2	-2.1 / 40.0	-10.5 / 36.3	-6.2 / 40.5	-10.5 / 36.3
AGE2ETHN2	1.0 / 39.8	-8.7 / 35.2	-4.0 / 39.2	-8.7 / 35.2
AGE3ETHN2	0.73 / 55.3	-9.5 / 48.9	-4.5 / 56.1	-9.5 / 48.9
AGE4ETHN2	-0.87 / 55.8	-10.5 / 49.8	-7.4 / 55.3	-10.5 / 49.8
AGE5ETHN2	-1.7 / 61.9	-10.4 / 56.1	-5.9 / 64.0	-10.4 / 56.1
SEXAGE1ETHN1	-0.63 / 72.1	-9.4 / 65.2	-5.6 / 73.1	-10.1 / 64.2
SEXAGE2ETHN1	-1.5 / 84.6	-10.7 / 76.4	-7.8 / 85.9	-11.0 / 75.5
SEXAGE3ETHN1	-0.47 / 106	-10.6 / 95.1	-5.1 / 112	-11.3 / 93.5
SEXAGE4ETHN1	-0.90 / 117	-10.2 / 105	-7.3 / 118	-10.8 / 104
SEXAGE5ETHN1	-0.55 / 75.8	-9.2 / 68.7	-4.9 / 77.8	-9.3 / 68.2
SEXAGE1ETHN2	0.35 / 56.6	-8.6 / 50.7	-4.0 / 58.4	-9.7 / 49.3
SEXAGE2ETHN2	0.33 / 61.0	-8.8 / 54.6	-4.0 / 62.4	-9.4 / 53.3
SEXAGE3ETHN2	-0.04 / 77.5	-10.4 / 68.8	-4.8 / 79.9	-11.5 / 67.1
SEXAGE4ETHN2	-2.6 / 78.3	-11.9 / 70.5	-9.4 / 79.9	-12.2 / 69.2
SEXAGE5ETHN2	-0.46 / 85.5	-10.1 / 76.9	-4.4 / 91.1	-10.5 / 75.8

Table A.4: Relative bias (in %) and RRMSE (in %) of variance estimates over $3,000$ simulation runs under simple random sampling with $n = 5,000$. A '*' indicates that bias and mean squared error vanish.

Cell name	$v^{(d)}$	$v^{(w)}$	$v_g^{(w)}$	v^{RW}
Intercept	*	*	*	*
SEX	-1.7 / 1.7	*	*	*
AGE1	-1.6 / 2.8	*	*	*
AGE2	0.31 / 2.3	*	*	*
AGE3	-0.97 / 2.9	*	*	*
AGE4	-1.0 / 2.9	*	*	*
AGE5	-0.88 / 2.5	*	*	*
ETHN1	1.3 / 9.0	0.42 / 8.8	-0.87 / 8.9	0.42 / 8.8
ETHN2	0.59 / 5.8	-0.37 / 5.7	0.12 / 5.8	-0.37 / 5.7
SEXAGE1	-1.9 / 4.3	-2.0 / 3.0	-1.5 / 2.7	*
SEXAGE2	-0.68 / 4.0	-2.5 / 3.3	-1.8 / 2.9	*
SEXAGE3	0.20 / 4.5	-1.9 / 3.2	-1.2 / 2.9	*
SEXAGE4	0.97 / 4.6	0.26 / 2.8	0.90 / 2.9	*
SEXAGE5	0.08 / 3.5	-0.56 / 2.8	0.08 / 2.6	*
SEXETHN1	0.55 / 11.5	-0.39 / 11.2	-0.00 / 11.3	-0.39 / 11.2
SEXETHN2	0.46 / 8.8	-0.48 / 8.5	-0.05 / 8.6	-0.48 / 8.5
AGE1ETHN1	-1.2 / 18.6	-2.2 / 18.1	-2.0 / 18.1	-2.2 / 18.1
AGE2ETHN1	1.1 / 20.4	-0.33 / 19.7	0.22 / 19.9	-0.33 / 19.7
AGE3ETHN1	1.1 / 25.6	0.20 / 24.9	0.90 / 25.3	0.20 / 24.9
AGE4ETHN1	-0.25 / 24.1	-1.4 / 23.4	-0.88 / 23.8	-1.4 / 23.4
AGE5ETHN1	-0.20 / 20.9	-1.1 / 20.3	-0.71 / 20.4	-1.1 / 20.3
AGE1ETHN2	-0.87 / 12.9	-1.8 / 12.2	-1.6 / 12.4	-1.8 / 12.2
AGE2ETHN2	0.12 / 12.8	-1.1 / 12.1	-0.55 / 12.1	-1.1 / 12.1
AGE3ETHN2	0.58 / 17.2	-0.65 / 16.4	-0.04 / 16.6	-0.65 / 16.4
AGE4ETHN2	-0.19 / 17.3	-0.75 / 16.6	-0.26 / 16.9	-0.75 / 16.6
AGE5ETHN2	1.4 / 19.4	0.72 / 18.7	1.2 / 19.0	0.72 / 18.7
SEXAGE1ETHN1	0.24 / 23.4	-0.36 / 22.8	-0.19 / 22.9	-0.27 / 22.7
SEXAGE2ETHN1	0.25 / 26.9	-1.3 / 26.1	-0.79 / 26.6	-1.6 / 25.8
SEXAGE3ETHN1	-0.44 / 34.1	-1.4 / 33.4	-0.86 / 33.8	-1.9 / 33.1
SEXAGE4ETHN1	0.88 / 37.2	-0.59 / 36.4	-0.07 / 37.1	-0.50 / 36.1
SEXAGE5ETHN1	0.16 / 23.4	-0.93 / 22.8	-0.57 / 23.1	-1.2 / 22.7
SEXAGE1ETHN2	-0.48 / 17.8	-1.3 / 17.1	-1.1 / 17.4	-1.6 / 16.7
SEXAGE2ETHN2	-0.09 / 19.0	-1.3 / 18.3	-0.81 / 18.5	-1.6 / 17.8
SEXAGE3ETHN2	-1.2 / 23.8	-2.5 / 23.1	-2.0 / 23.5	-2.6 / 22.7
SEXAGE4ETHN2	0.62 / 25.4	-0.41 / 24.7	-0.03 / 25.1	-0.39 / 24.3
SEXAGE5ETHN2	0.83 / 26.8	-0.07 / 26.2	0.41 / 26.6	-0.47 / 25.8

Table A.5: Coverage rates for some estimator variance estimator combinations for sample sizes $n = 500$ and $n = 5,000$.

Cell name	$C(\widehat{t}^{d}_{SAE}, v^{(d)})$	$C(\widehat{t}^{w}_{SAE}, v^{(w)})$	$C(\widehat{t}^{w}_{SAE}, v_g^{(w)})$	$C(\widehat{t}^{RW}_{SAE}, v^{RW})$
Intercept	*	*	*	*
SEX	94.8 / 94.3	*	*	*
AGE1	95.4 / 94.3	*	*	*
AGE2	94.9 / 95.0	*	*	*
AGE3	94.5 / 94.7	*	*	*
AGE4	94.3 / 94.3	*	*	*
AGE5	95.5 / 94.8	*	*	*
ETHN1	93.6 / 95.1	91.2 / 95.0	91.7 / 95.1	91.2 / 95.0
ETHN2	94.5 / 95.2	93.1 / 95.3	93.6 / 95.4	93.1 / 95.3
SEXAGE1	94.9 / 94.7	92.9 / 94.8	94.3 / 94.8	*
SEXAGE2	94.5 / 94.8	92.7 / 94.4	94.1 / 94.5	*
SEXAGE3	96.0 / 95.2	93.9 / 94.8	94.7 / 94.9	*
SEXAGE4	94.1 / 94.7	93.3 / 94.6	94.0 / 94.6	*
SEXAGE5	94.2 / 95.5	94.0 / 95.0	94.7 / 95.2	*
SEXETHN1	92.5 / 94.6	91.5 / 94.6	91.0 / 94.5	91.5 / 94.6
SEXETHN2	94.1 / 94.7	92.0 / 95.0	92.6 / 95.0	92.0 / 95.0
AGE1ETHN1	94.1 / 93.9	88.0 / 93.3	86.7 / 93.4	88.0 / 93.3
AGE2ETHN1	90.0 / 93.4	89.9 / 93.4	86.2 / 93.3	89.9 / 93.4
AGE3ETHN1	78.1 / 92.6	77.5 / 93.2	77.9 / 93.4	77.5 / 93.2
AGE4ETHN1	81.0 / 93.5	80.5 / 93.6	80.8 / 93.8	80.5 / 93.6
AGE5ETHN1	89.1 / 92.1	88.6 / 93.8	86.4 / 94.1	88.6 / 93.8
AGE1ETHN2	92.4 / 94.9	90.9 / 94.4	90.3 / 94.5	90.9 / 94.4
AGE2ETHN2	93.5 / 93.0	92.1 / 94.6	91.7 / 94.7	92.1 / 94.6
AGE3ETHN2	83.8 / 94.2	84.2 / 94.0	87.4 / 94.2	84.2 / 94.0
AGE4ETHN2	81.7 / 94.1	83.6 / 94.2	86.3 / 93.9	83.6 / 94.2
AGE5ETHN2	91.5 / 94.7	90.1 / 94.2	86.0 / 94.1	90.1 / 94.2
SEXAGE1ETHN1	84.9 / 94.3	84.4 / 93.5	84.2 / 93.5	84.3 / 93.1
SEXAGE2ETHN1	73.8 / 92.0	73.7 / 91.8	73.8 / 92.1	73.7 / 92.0
SEXAGE3ETHN1	57.8 / 91.5	57.3 / 91.4	57.7 / 91.5	57.3 / 91.4
SEXAGE4ETHN1	51.0 / 92.5	50.9 / 92.5	51.0 / 92.5	50.8 / 92.5
SEXAGE5ETHN1	82.0 / 91.9	81.8 / 92.9	81.7 / 93.2	81.7 / 92.9
SEXAGE1ETHN2	80.1 / 93.5	83.7 / 93.6	86.3 / 93.6	85.2 / 93.3
SEXAGE2ETHN2	92.8 / 93.2	91.1 / 94.1	87.6 / 94.3	89.9 / 94.4
SEXAGE3ETHN2	80.8 / 93.2	80.3 / 93.6	80.4 / 93.5	80.3 / 93.8
SEXAGE4ETHN2	77.9 / 92.3	77.7 / 93.7	77.8 / 93.7	77.5 / 93.9
SEXAGE5ETHN2	73.4 / 92.0	73.0 / 93.7	73.3 / 93.7	72.9 / 93.8

Table A.6: RRMSE (in %) over 15,000 runs for table $SEX \times MST \times EMPL$ under simple random sampling with $n = 500$. The second column contains population totals.

Cell name	t_{SME}	$\widehat{t}^{d}_{SME}$	$\widehat{t}^{w}_{SME}$	$\widehat{t}^{RW}_{SME}$	$\widehat{t}^{SRW}_{SME}$	$\widehat{t}^{MW}_{SME}$
Intercept	188216	0	0	0	0	0
SEX	97493	4.3	0	0	0	0
MST1	24322	11.6	10.9	0	0	0
MST2	52810	7.1	5.1	0	0	0
EMPL1	4297	29.2	29.8	29.8	29.8	29.9
EMPL2	89278	4.7	3.7	3.7	3.7	3.7
SEXMST1	17498	14.0	12.7	0	0	0
SEXMST2	23109	11.9	9.9	0	0	0
SEXEMPL1	2550	38.5	39.0	39.0	39.0	39.1
SEXEMPL2	58596	6.6	4.7	4.7	4.7	4.7
MST1EMPL1	528	84.1	86.3	85.3	85.6	86.1
MST2EMPL1	1485	50.0	51.1	50.8	50.9	50.8
MST1EMPL2	18469	13.5	12.2	6.8	6.7	6.8
MST2EMPL2	18127	13.6	13.6	11.1	11.2	11.0
SEXMST1EMPL1	322	108	110	110	110	110
SEXMST2EMPL1	612	78.2	79.8	79.4	79.4	79.4
SEXMST1EMPL2	14422	15.5	13.9	7.0	6.8	6.8
SEXMST2EMPL2	9309	19.4	17.8	15.7	15.7	15.5

Table A.7: RRMSE (in %) over 15,000 runs for table $SEX \times MST \times EMPL$ under simple random sampling with $n = 5,000$. The second column contains population totals.

Cell name	t_{SME}	$\widehat{t}^{d}_{SME}$	$\widehat{t}^{w}_{SME}$	$\widehat{t}^{RW}_{SME}$	$\widehat{t}^{SRW}_{SME}$	$\widehat{t}^{MW}_{SME}$
Intercept	188216	0	0	0	0	0
SEX	97493	1.4	0	0	0	0
MST1	24322	3.6	3.3	0	0	0
MST2	52810	2.2	1.6	0	0	0
EMPL1	4297	9.1	9.1	9.1	9.1	9.1
EMPL2	89278	1.5	1.1	1.1	1.1	1.2
SEXMST1	17498	4.4	3.9	0	0	0
SEXMST2	23109	3.7	3.0	0	0	0
SEXEMPL1	2550	12.0	11.9	11.9	11.9	11.9
SEXEMPL2	58596	2.1	1.4	1.4	1.4	1.4
MST1EMPL1	528	26.2	26.2	26.0	26.0	26.0
MST2EMPL1	1485	15.6	15.6	15.5	15.5	15.5
MST1EMPL2	18469	4.2	3.7	2.1	2.0	2.0
MST2EMPL2	18127	4.3	3.5	3.4	3.4	3.4
SEXMST1EMPL1	322	33.6	33.7	33.4	33.4	33.4
SEXMST2EMPL1	612	24.6	24.6	24.5	24.5	24.5
SEXMST1EMPL2	14422	4.8	4.3	2.1	2.1	2.1
SEXMST2EMPL2	9309	6.1	5.5	4.8	4.8	4.7

Table A.8: Relative bias (in %) and RRMSE (in %) of variance estimates over 3,000 simulation runs under srs with $n = 500$. A '*' indicates that bias and mean squared error vanish.

Cell name	$v^{(d)}(\widehat{t}^{d}_{SME})$	$v^{(w)}(\widehat{t}^{w}_{SME})$	$v_g^{(w)}(\widehat{t}^{w}_{SME})$	$v^{RW}(\widehat{t}^{RW}_{SME})$	$v^{SRW}(\widehat{t}^{RW}_{SME})$
Intercept	*	*	*	*	*
SEX	-0.42 / 0.59	*	*	*	*
MST1	-0.77 / 9.8	-9.3 / 12.6	-5.1 / 11.1	*	*
MST2	0.45 / 4.3	-10.2 / 13.1	-5.8 / 11.0	*	*
EMPL1	0.13 / 28.5	-9.4 / 27.3	-4.8 / 30.5	-9.4 / 27.3	-9.4 / 27.3
EMPL2	1.2 / 1.3	-9.1 / 10.5	-4.7 / 7.8	-9.1 / 10.5	-9.1 / 10.5
SEXMST1	-0.02 / 12.4	-7.9 / 12.6	-3.6 / 11.7	*	*
SEXMST2	0.15 / 10.0	-10.4 / 12.9	-5.7 / 10.5	*	*
SEXEMPL1	-1.4 / 36.8	-10.4 / 34.4	-5.9 / 38.6	-10.4 / 34.4	-10.4 / 34.4
SEXEMPL2	0.07 / 3.6	-8.9 / 10.1	-4.4 / 7.1	-8.9 / 10.1	-8.9 / 10.1
MST1EMPL1	-0.03 / 84.5	-9.9 / 76.6	-5.2 / 88.0	-11.3 / 74.1	-12.2 / 73.0
MST2EMPL1	0.88 / 50.9	-9.0 / 46.4	-4.6 / 52.7	-9.1 / 46.0	-9.2 / 46.0
MST1EMPL2	0.21 / 11.9	-8.4 / 12.8	-4.0 / 11.6	-11.4 / 19.3	-16.6 / 22.7
MST2EMPL2	0.70 / 11.9	-8.3 / 14.1	-4.0 / 13.3	-8.3 / 11.6	-9.1 / 12.3
SM1EMPL1[3]	0.58 / 107	-9.0 / 97.2	-4.6 / 112	-11.0 / 93.0	-11.5 / 92.9
SM2EMPL1	-0.25 / 78.2	-9.4 / 71.1	-4.4 / 81.8	-10.2 / 69.7	-10.4 / 69.4
SM1EMPL2	0.33 / 14.1	-8.0 / 13.6	-3.6 / 12.7	-12.8 / 25.6	-16.5 / 27.9
SM2EMPL2	1.5 / 18.5	-8.7 / 16.7	-4.4 / 16.5	-9.9 / 14.9	-9.5 / 14.8

[3] SM $\widehat{=}$ SEXMST

Table A.9: Relative bias (in %) and RRMSE (in %) of variance estimates over 3,000 simulation runs under srs with $n = 5,000$. A '*' indicates that bias and mean squared error vanish.

Cell name	$v^{(d)}(\widehat{t}^{w}_{SME})$	$v^{(w)}(\widehat{t}^{w}_{SME})$	$v_g^{(w)}(\widehat{t}^{w}_{SME})$	$v^{RW}(\widehat{t}^{RW}_{SME})$	$v^{SRW}(\widehat{t}^{RW}_{SME})$
Intercept	*	*	*	*	*
SEX	-1.7 / 1.7	*	*	*	*
MST1	0.59 / 3.1	-0.41 / 2.9	0.07 / 3.0	*	*
MST2	-0.55 / 1.5	-2.3 / 3.6	-1.9 / 3.4	*	*
EMPL1	-0.11 / 8.9	-1.0 / 8.8	-0.56 / 8.9	-1.0 / 8.8	-1.0 / 8.8
EMPL2	0.45 / 0.48	0.90 / 2.0	1.4 / 2.3	0.90 / 2.0	0.90 / 2.0
SEXMST1	-0.31 / 3.9	-0.71 / 3.3	-0.25 / 3.3	*	*
SEXMST2	0.25 / 3.2	-1.3 / 2.9	-0.82 / 2.8	*	*
SEXEMPL1	-1.5 / 12.0	-2.7 / 11.8	-2.2 / 11.9	-2.7 / 11.8	-2.7 / 11.8
SEXEMPL2	0.75 / 1.4	-0.20 / 1.6	0.25 / 1.6	-0.20 / 1.6	-0.20 / 1.6
MST1EMPL1	1.1 / 26.5	-0.02 / 26.1	0.58 / 26.6	-0.70 / 25.4	-0.94 / 25.3
MST2EMPL1	1.1 / 16.0	-0.18 / 15.6	0.31 / 15.9	-0.40 / 15.4	-0.32 / 15.4
MST1EMPL2	0.23 / 3.7	-0.95 / 3.5	-0.48 / 3.4	-2.1 / 5.6	-6.5 / 8.3
MST2EMPL2	0.12 / 3.8	-0.89 / 4.0	-0.38 / 3.9	-1.1 / 3.0	-1.6 / 3.2
SM1EMPL1[4]	0.55 / 33.4	-0.46 / 33.0	0.01 / 33.6	-1.1 / 32.2	-1.3 / 32.2
SM2EMPL1	-2.1 / 24.7	-3.3 / 24.4	-2.9 / 24.8	-3.5 / 24.0	-3.5 / 24.0
SM1EMPL2	-0.23 / 4.4	-1.2 / 3.8	-0.72 / 3.8	-1.7 / 7.9	-5.2 / 9.4
SM2EMPL2	0.86 / 5.9	-0.91 / 5.0	-0.38 / 4.9	-0.32 / 3.9	0.39 / 3.9

[4]SM $\widehat{=}$ SEXMST

Table A.10: Coverage rates for some estimator variance estimator combinations for sample sizes $n = 500$ and $n = 5,000$ and target table $SEX \times MST \times EMPL$.

Cell name	$C(\widehat{t}^{d}, v^{(d)})$	$C(\widehat{t}^{w}, v^{(w)})$	$C(\widehat{t}^{w}, v_g^{(w)})$	$C(\widehat{t}^{RW}, v^{RW})$	$C(\widehat{t}^{RW}, v^{SRW})$
Intercept	*	*	*	*	*
SEX	94.8 / 94.3	*	*	*	*
MST1	94.5 / 95.0	93.1 / 94.8	93.8 / 94.8	*	*
MST2	95.2 / 94.8	93.7 / 94.6	94.4 / 94.8	*	*
EMPL1	93.3 / 95.1	91.9 / 94.8	92.0 / 94.9	91.9 / 94.8	91.9 / 94.8
EMPL2	95.4 / 95.3	94.6 / 95.6	94.8 / 95.6	94.6 / 95.6	94.6 / 95.6
SEXMST1	93.7 / 95.4	93.3 / 95.5	93.9 / 95.5	*	*
SEXMST2	95.0 / 96.0	93.8 / 95.9	94.6 / 95.9	*	*
SEXEMPL1	90.3 / 94.0	90.4 / 93.6	91.1 / 93.7	90.4 / 93.6	90.4 / 93.6
SEXEMPL2	94.4 / 95.2	94.0 / 95.5	94.6 / 95.7	94.0 / 95.5	94.0 / 95.5
MST1EMPL1	75.1 / 93.4	74.8 / 93.3	75.0 / 93.5	74.7 / 93.3	74.7 / 93.3
MST2EMPL1	90.1 / 93.5	89.5 / 94.0	88.7 / 93.9	89.6 / 94.2	89.6 / 94.2
MST1EMPL2	94.5 / 95.3	93.6 / 95.1	94.0 / 95.2	92.1 / 94.8	91.1 / 94.3
MST2EMPL2	95.8 / 95.2	94.3 / 95.0	95.0 / 94.9	94.1 / 95.2	93.8 / 95.1
SM1EMPL1[5]	57.7 / 92.4	57.5 / 92.4	57.7 / 92.4	57.4 / 92.2	57.4 / 92.2
SM2EMPL1	80.0 / 91.9	79.6 / 92.3	79.7 / 92.5	79.6 / 92.6	79.6 / 92.6
SM1EMPL2	94.1 / 95.2	93.7 / 95.1	94.2 / 95.3	91.7 / 95.1	90.9 / 94.4
SM2EMPL2	96.1 / 94.8	93.7 / 94.8	94.2 / 94.9	93.9 / 94.9	93.9 / 95.0

[5] SM $\widehat{=}$ SEXMST

Appendix B

Method of computation

All computations have been performed in S-Plus. For the general regression estimator we have used a generic function that takes as input a regression model, prepared tables of population totals, sample data, and a vector of target variables (dummies of a target table), and outputs the regression total estimator for the target variables, and optionally regression weights and residuals. This function is used to compute the overall weighting scheme regression estimates and corresponding variance estimates. It can also be used in the repeated weighting step. The repeated weighting estimators and variance estimators have been computed using S-Plus scripts specific to the target tables simulated.

In order to carry out the simulations for small sample size ($n = 500$ in our case) we had to slightly change the repeated weighting estimator so as to avoid the problem of empty cells in the re-weighting scheme. The usual (weighted least squares) regression weighting can not be used to impose all consistency constraints corresponding to the (re-)weighting scheme's cells when some of these cells contain no observations. By slightly modifying the regression coefficients as described below this problem is circumvented.

Using the fact that re-weighting schemes for the frequency tables considered consist of margins of these tables, the repeated weighting estimator can be written as

$$\widehat{t}^{RW} = \widehat{t}^{w} + \widehat{B}^{t}(r - \Delta\widehat{t}^{w}), \tag{B.1}$$

where $\widehat{t}^{w}$ is the vector of overall weighting scheme regression estimates of the table being re-weighted, r is a vector of marginal totals, exactly known or estimated, and Δ is the corresponding aggregation matrix indicating which marginal cells of the target table are in the re-weighting scheme. The weighted least squares regression coefficients take the form

$$\widehat{B}^{t} = \Lambda\Delta^{t}(\Delta\Lambda\Delta^{t})^{-1} \tag{B.2}$$

where $\Lambda = \mathrm{diag}(\widehat{t}^{w})$. By changing the matrix Λ into $\tilde{\Lambda} = \mathrm{diag}(\widehat{t}^{w} + \lambda)$ with λ a vector of positive constants we ensure that the repeated weighting estimator satisfies all consistency constraints even if some weighting cells contain no observations. For details, see Boonstra (2004). For the vector λ we have simply taken all components equal to $\frac{N}{2n}$, which is like adding half an observation to each cell. The estimators thus obtained are for all practical purposes the same as the original repeated weighting estimator and only differ when some weighting cells have zero or nearly zero sample count.

Appendix C

S-Plus code

Here we list an S-Plus function for the general regression estimator and a function that computes the repeated weighting estimator for the table $SEX \times MST \times EMPL$. The latter function is given as an example of how repeated weighting estimation and variance estimation can be done in S-Plus.

```
# Function that computes regression estimates for finite population totals

# Input:
#   - formula: formula object containing both the response, which may
#            consist of multiple columns, and the regression model
#     If the response consists of factor variables these must have been
#     dummified, e.g. using model.matrix().
#   - weights: the weights used in the regression
#   - poptables: a list of table array objects with names(dimnames())
#       set to the appropriate table names
#     The population tables must be named after the variables, separated
#     by colons for multiway tables, in the standard S-Plus naming
#     convention for interaction effects.
#     An intercept must be present with corresponding population total
#     named "(Intercept)".

# Output:
#   - tR: (vector of) regression estimate(s)
#   - coefficients: (matrix of) regression coefficients
#   - wR: regression weights, optional
#   - residuals: (matrix of) residuals, optional

RegrEst <- function(formula, weights, poptables, Rweights = F,
                    Residuals = F) {

    # use treatment contrasts (S-Plus default is Helmert)
    options.old <- options(contrasts = c(factor = "contr.treatment",
        ordered = "contr.poly"))
    on.exit(options(options.old))  # restore original contrast settings

    Reg <- lm(formula, weights=weights, x=Rweights)  # lm or mlm object
    # model matrix Reg$x needed for computation of regression weights
```

```
    # Construct vector of population totals in parametrisation of Reg
    PopVector <- vector(mode = "numeric", length = Reg$rank)
    # Projected form of regression estimator is used, so make sure that
    # an intercept is present.
    if (is.null(Reg$assign$"(Intercept)")) stop("Intercept is required")
    for (i in 1:length(Reg$assign)) {
            PopTable <- poptables[[names(Reg$assign)[[i]]]]
            if (is.null(PopTable)) stop(paste("Population table",
                  names(Reg$assign)[[i]], "is missing"))
            PopVector[Reg$assign[[i]]] <- PopTable
    }

    # Now we have the required PopVector in the appropriate
    # parametrization
    # The regression estimate is simply the inner product B^t X:
    tR <- as.vector(t(coef(Reg)) %*% PopVector) # Note the projection form

    # Compute regression weights; note again the projection form
    if (Rweights)
            wR <- weights * as.vector(Reg$x %*% backsolve(Reg$R,
                                    solve(t(Reg$R), PopVector)))
    else
            wR <- NULL

    if (Residuals)
            dimnames(Reg$residuals) <- NULL  # drop dimnames attribute
    else
            Reg$residuals <- NULL

    # output
    list(tR = tR, wR = wR, coefficients = coef(Reg),
         residuals = Reg$residuals)

}  # end function RegrEst
```

```
# Function to compute repeated weighting estimator for SEX x MST x EMPL
# as well as (super-)residuals.
# Uses function RegrEst() which computes a regression estimator including
# residuals.
# Note: this function is not generic. It is specifically written for this
# particular target table.
# The existence of the following (global) variables is assumed:
# - SEX, MST, EMPL: factor variables
# - Y: dummified table SEX x MST x EMPL
# - lm object Reg containing the information on the overall weighting
#   scheme regression estimates:
#       - Reg$tR: table of regression estimates
#       - Reg$wR: weights corresponding to overall weighting scheme
#       - Reg$residuals: regression residuals of target variables
# - poptables containing all register totals required for the reweightings
# Output:
#   - tRW: (vector of) repeated weighting estimate(s)
#   - residuals: (matrix of) superresiduals from which variance estimates
#     can be computed
```

```
SEX.MST.EMPL <- function() {

    # Reweighting according to complete splitting-up weighting scheme
    # Reweighting scheme SEX x MST + SEX x EMPL + MST x EMPL
    # SEX x MST in register, SEX x EMPL from sample,
    # MST x EMPL needs reweighting
    # Add estimated population tables SEX x EMPL to poptables
    poptables$"SEX:EMPL" <- Reg$tR[attr(Y, "assign")$"SEX:EMPL"]
    poptables$EMPL <- Reg$tR[attr(Y, "assign")$EMPL]

    # First compute MST x EMPL repeated weighting estimate
    RegRep <- RegrEst(Y[, attr(Y, "assign")$"MST:EMPL"] ~ MST + EMPL,
              Reg$wR, poptables, Rweights=FALSE, Residuals=TRUE)
    poptables$"MST:EMPL" <- RegRep$tR  # add to collection of poptables

    # Compute contribution to superresiduals from MST x EMPL margin
    ekME <- Reg$residuals[, attr(Y, "assign")$MST]
    ekME <- -(ekME %*% RegRep$coefficients[attr(RegRep$coefficients,
             "assign")$MST,])

    # Now compute the repeated weighting estimates of SEX x MST x EMPL
    RegRep <- RegrEst(Y ~ SEX * MST + SEX * EMPL + MST * EMPL, Reg$wR,
                  poptables, Rweights=FALSE, Residuals=TRUE)

    # Compute superresiduals for repeated weighting estimates RegRep$tR
    ek <- Reg$residuals
    ek <- ek - (Reg$residuals[, attr(Y, "assign")$MST] %*%
          RegRep$coefficients[attr(RegRep$coefficients, "assign")$MST,])
    ek <- ek - (Reg$residuals[, attr(Y, "assign")$"SEX:MST"] %*%
       RegRep$coefficients[attr(RegRep$coefficients, "assign")$"SEX:MST"
            ,])
    ek <- ek + (ekME %*% RegRep$coefficients[attr(RegRep$coefficients,
             "assign")$"MST:EMPL",])
    }
    else
            ek <- NULL

    list(tRW = RegRep$tR, residuals = ek)
} # end function SEX.MST.EMPL
```

References

Boonstra, H. J. H. (2004): *Calibration of tables of estimates.* Heerlen: Statistics Netherlands.

Boonstra, H. J. H., Van den Brakel, J. A., Knottnerus, P., Nieuwenbroek, N. and Renssen, R. J. (2003): *A strategy to obtain consistency among tables of survey estimates.* DACSEIS deliverable D7.2.

Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H. and Snijders, V. (2003): *Estimating consistent table sets: Position paper on repeated weighting.* Voorburg: Statistics Netherlands.

Münnich, R. and Schürle, J. (2003): *Monte-Carlo simulation study of European surveys.* DACSEIS deliverables 3.1 and 3.2.

Renssen, R. H., Kroese, A. H. and Willeboordse, A. (2001): *Aligning Estimates by Repeated Weighting.* Heerlen: Statistics Netherlands.

Van Duin, C. and Snijders, V. (2003): *Simulation studies of repeated weighting.* Voorburg: Statistics Netherlands.