Bootstrapping Combined Estimator based on Register and Sample Survey Data

Léander Kuijvenhoven and Sander Scholtus

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201123)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

© Statistics Netherlands,
The Hague/Heerlen, 2011.
Reproduction is permitted.
'Statistics Netherlands' must be quoted as source.

60083201123 X-10

Bootstrapping Combined Estimators based on Register and Sample Survey Data

Summary: This paper describes how the bootstrap resampling method may be used to assess the accuracy of estimates based on a combination of data from registers and sample surveys. We consider three different estimators that may be applied in this context. The validity of the proposed bootstrap method is tested in a simulation study with realistic data from the Dutch Educational Attainment File.

Keywords: register, sample, combined data, bootstrap

1 Introduction

In this paper, we consider the situation where estimates are based on data from different sources. In particular, producers of official statistics are increasingly making use of existing registers. There are several reasons for doing this, like reducing costs and reducing the burden on respondents. Also, businesses and individuals are becoming less tolerant of surveys, as is reflected by lower response rates. National Statistical Institutes (NSIs) are therefore seeking ways to put different sources of information together, to increase the ability to produce information with good quality attributes in an efficient way.

A problem, however, is that registers are often primarily used for non-statistical purposes, and therefore not always ideal from a statistical perspective. In some cases, an additional sample survey is needed to obtain reliable statistical results. The problem of assessing the accuracy of estimates based on a combination of administrative sources and sample surveys has, therefore, become very relevant to NSIs. In this paper we examine a relatively simple way to evaluate the accuracy of an estimate based on combined data, namely by using a form of bootstrap resampling.

The primary objective of this paper is to develop a methodology for assessing the accuracy of particular estimates from combined data. We do not discuss the problem of micro integration itself, e.g. how to construct a statistical database or how to handle inconsistencies between data from different sources. Instead, we assume that a statistical database has already been constructed.

The outline of this paper is as follows. In Section 2 the setting at hand is further clarified, and three types of estimators that may be used in this context are introduced. One of these estimators is a classical regression estimator which does not use register data and serves as a benchmark for comparing the other estimators. Section 3 introduces the proposed bootstrap method for combined data. Section 4 describes a simulation study in which the proposed bootstrap method is applied to realistic data from the Dutch Educational Attainment File. Finally, Section 5 closes the paper with a short discussion and some ideas for further research.

2 Combining Register and Survey Data

2.1 Description of the Situation

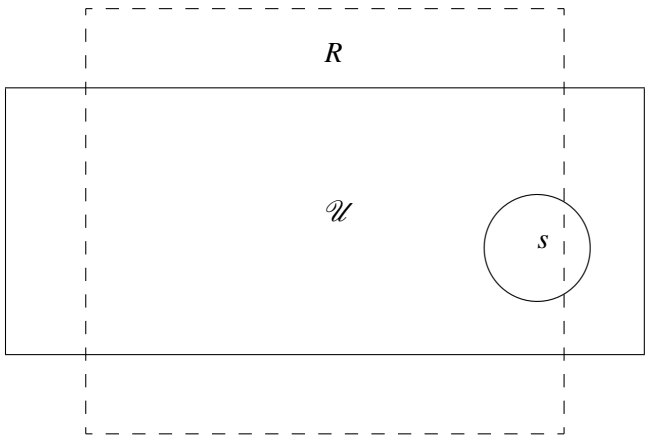
For convenience, we describe the case that a target variable is observed in one register and one sample. We denote the register by R and the sample by s. Let $\mathscr{U}$ denote the target population. Figure 1 shows the relationship between $\mathscr{U}$, R and s graphically.

Let y_k denote the value of a target variable y for an element $k \in \mathscr{U}$. The objective of the survey is to estimate the total value of y for the target population:

$$\theta_y = \sum_{k \in \mathscr{U}} y_k. \tag{1}$$

In this paper, we are interested in the case where the register only covers a selective part of the target population, so that a simple register total would not be a valid estimate

Figure 1: The target population (rectangle), the register (dashed rectangle), and the sample (circle).



for (1). This happens for instance with the Dutch registers on education, which have come into existence only recently, and hence mainly contain information on younger persons. In order to obtain valid estimates of educational attainment, the information from the register has to be combined with information from a sample survey. There are many ways to form such a combined estimator. We consider three types of estimators in the next subsection.

In general, the sample s may be drawn originally from the target population $\mathscr{U}$. Therefore, it can be partitioned into a subsample that overlaps with the register (s_R) and another subsample that does not (s_{NR}). The information from s_{NR} is our only source for inference about $\mathscr{U}_{NR}$, the part of the population that is not covered by the register. Of course, it seems more effective to draw the sample only from $\mathscr{U}_{NR}$. However, this means that the sample cannot be drawn before the register data is obtained, which in turn might conflict with timeliness constraints for the survey. Another reason why the sample may partly overlap with the register, is that an NSI may decide to use existing sample data in the data integration process, instead of introducing a new sample survey. This is, e.g., the case for the Dutch Educational Attainment File, which re-uses data from several cycles of the Dutch Labour Force Survey.

Throughout this paper, we make the following simplifying assumptions:

- The target population can be partitioned into two disjoint strata: $\mathscr{U}_R = \mathscr{U} \cap R$ and $\mathscr{U}_{NR} = \mathscr{U} \setminus \mathscr{U}_R$, and this stratification is fixed in the sense that it does not depend on an actual realisation of the register R. Note that there is supposed to be no overcoverage in the register, i.e. we assume that any register records pertaining to elements outside $\mathscr{U}$ can be identified and removed from R.

- The register contains values of a random variable $z_k = y_k + \xi_k e_k$, where ξ_k is a dichotomous variable with $\mathrm{P}(\xi_k = 1) = \lambda_k$ and $\mathrm{P}(\xi_k = 0) = 1 - \lambda_k$, indicating whether an error occurs in the recorded value for element k, and the error e_k is drawn from a distribution with mean μ_k and variance σ_k^2. Moreover, the z_k are drawn independently. Note that λ_k represents the probability that the register value for element k contains an error.

- By contrast, we assume that the target variable is recorded without error in the sample survey. In practice, of course, some measurement errors are bound to occur, but we assume that the effect of these errors is negligible compared to the sampling variance and the effect of errors in the register. This assumption reflects the fact that a statistical institute has direct control over the quality of the sample data, whereas the register data is usually collected by an external party for non-statistical purposes.

A straightforward calculation shows that, under the assumed error model,

$$\mathrm{E}(z_k) = \mathrm{E}_{\xi_k}[\mathrm{E}(z_k \mid \xi_k)] = \mathrm{E}_{\xi_k}(y_k + \xi_k \mu_k) = y_k + \lambda_k \mu_k, \tag{2}$$

and

$$\begin{aligned} \mathrm{V}(z_k) &= \mathrm{E}_{\xi_k}[\mathrm{V}(z_k \mid \xi_k)] + \mathrm{V}_{\xi_k}[\mathrm{E}(z_k \mid \xi_k)] \\ &= (1-\lambda_k)\mathrm{V}(z_k \mid \xi_k = 0) + \lambda_k \mathrm{V}(z_k \mid \xi_k = 1) + \mathrm{V}_{\xi_k}(y_k + \xi_k \mu_k) \\ &= 0 + \lambda_k \sigma_k^2 + \mu_k^2 \lambda_k (1-\lambda_k) \\ &= \lambda_k [\sigma_k^2 + \mu_k^2 (1-\lambda_k)]. \end{aligned} \tag{3}$$

2.2 Three Types of Estimators

2.2.1 The Ordinary Regression Estimator

The first estimator that we consider does not use any information from the register, but is solely based on the sample survey. In principle, one could use the direct (or Horvitz-Thompson) estimator $\sum_{k \in s} y_k/\pi_k$, where π_k denotes the inclusion probability of element k in the sample.[1] This is in fact an unbiased estimator of θ_y. It is common practice, however, to apply a linear regression model to increase the precision of the estimator, and to correct for nonresponse in the original sample. This leads to the well-known regression estimator:

$$\hat{\theta}_{1y} = \sum_{k \in s} w_{1k} y_k, \tag{4}$$

with

$$w_{1k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{1k}\big(\sum_{l \in s} \frac{\mathbf{x}_{1l}\mathbf{x}'_{1l}}{\pi_l}\big)^{-1}\big(\sum_{l \in \mathscr{U}} \mathbf{x}_{1l} - \sum_{l \in s} \frac{\mathbf{x}_{1l}}{\pi_l}\big)\Big]$$

the final weight of element $k \in s$. In this expression, $\mathbf{x}_{1k}$ denotes a vector of auxiliary variables that are observed for all elements of $\mathscr{U}$, corresponding to the chosen linear model. By construction, the final weights satisfy the so-called calibration equations:

$$\sum_{k \in s} w_{1k} \mathbf{x}_{1k} = \sum_{k \in \mathscr{U}} \mathbf{x}_{1k}.$$

The properties of the regression estimator are well-established (Särndal et al., 1992; Knottnerus, 2003). In particular, it is an asymptotically unbiased estimator. For future reference, we note the trivial fact that the variance of $\hat{\theta}_{1y}$ does not depend on the register, i.e. $\mathrm{V}(\hat{\theta}_{1y}) = \mathrm{V}_s(\hat{\theta}_{1y} \mid R)$.

[1] Note that π_k denotes the probability of inclusion in the sample, not the register, so it does not automatically follow that $\pi_k = 1$ for all $k \in \mathscr{U}_R$.

2.2.2 An Additive Combined Estimator

Next, we consider the following estimator for θ_y:

$$\hat{\theta}_{2y} = \sum_{k \in \mathscr{U}_R} z_k + \sum_{k \in s_R} (y_k - z_k) + \sum_{k \in s_{NR}} w_{2k} y_k, \tag{5}$$

with

$$w_{2k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{2k}\big(\sum_{l \in s_{NR}} \frac{\mathbf{x}_{2l}\mathbf{x}'_{2l}}{\pi_l}\big)^{-1}\big(\sum_{l \in \mathscr{U}_{NR}} \mathbf{x}_{2l} - \sum_{l \in s_{NR}} \frac{\mathbf{x}_{2l}}{\pi_l}\big)\Big]$$

the final weight of element $k \in s_{NR}$. In this case the regression estimator is used to calibrate s_{NR} on known or previously estimated marginal counts of $\mathscr{U}_{NR}$. This leads to an estimate for the total of y in $\mathscr{U}_{NR}$, which is added to the estimate for $\mathscr{U}_R$. The latter estimate is obtained as the observed total of z_k in $\mathscr{U}_R$, except for the elements of s_R, for which we use y_k because this value is taken to be more accurate.

In appendix A the following expressions for the bias and variance of $\hat{\theta}_{2y}$ are derived, under the assumptions made in Section 2.1:

$$\text{bias}(\hat{\theta}_{2y}) = \text{E}(\hat{\theta}_{2y}) - \theta_y \doteq \sum_{k \in \mathscr{U}_R} (1 - \pi_k)\lambda_k \mu_k \tag{6}$$

and

$$\text{V}(\hat{\theta}_{2y}) = \sum_{k \in \mathscr{U}_R} (1 - \pi_k)\lambda_k[\sigma_k^2 + \mu_k^2(1 - \lambda_k)] + \text{V}_s\Big(-\sum_{k \in s_R} \lambda_k \mu_k + \sum_{k \in s_{NR}} w_{2k} y_k\Big). \tag{7}$$

It is interesting to examine the properties of $\hat{\theta}_{2y}$ for some special cases of the error model from Section 2.1.

1. If it is assumed that $\mu_k = 0$ for all $k \in \mathscr{U}_R$, then it follows from (6) that $\hat{\theta}_{2y}$ is an asymptotically unbiased estimator for θ_y. In this case, it is expected that the errors in the register cancel out in aggregates. The variance of $\hat{\theta}_{2y}$ reduces to

   $$\text{V}(\hat{\theta}_{2y}) = \sum_{k \in \mathscr{U}_R} (1 - \pi_k)\lambda_k \sigma_k^2 + \text{V}_s\Big(\sum_{k \in s_{NR}} w_{2k} y_k\Big).$$

   Note that the first term is the expected error variance in the register (after correction with information from s_R) and the second term is the sampling variance in s_{NR}. However, the assumption that $\mu_k = 0$ may be too optimistic in practice.

2. An important special case occurs when y and z are binary variables, such that $y_k = 1$ if element k belongs to a domain of $\mathscr{U}$, and $y_k = 0$ otherwise. The population total θ_y then measures the size of the domain. Errors in the register correspond to misclassifications of elements of $\mathscr{U}_R$ with respect to the domain. In this case it is natural to assume the following error model for z_k:

   $$z_k = (1 - \xi_k)y_k + \xi_k(1 - y_k),$$

   making $\text{P}(z_k = y_k) = 1 - \lambda_k$ and $\text{P}(z_k = 1 - y_k) = \lambda_k$. In the model of Section 2.1 this leads to $\mu_k = 1 - 2y_k$ and $\sigma_k^2 = 0$. From this and (6) and (7), it follows that

   $$\text{bias}(\hat{\theta}_{2y}) \doteq \sum_{k \in \mathscr{U}_R} (1 - \pi_k)\lambda_k(1 - 2y_k)$$

and

$$\begin{aligned} \mathrm{V}(\hat{\theta}_{2y}) &= \sum_{k\in\mathscr{U}_R}(1-\pi_k)\lambda_k(1-\lambda_k)(1-2y_k)^2 \\ &\quad +\mathrm{V}_s\Big[-\sum_{k\in s_R}\lambda_k(1-2y_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big]. \end{aligned}$$

3. Kuijvenhoven and Scholtus (2010) consider the errors in the register to be deterministic: $\sigma_k^2 = 0$ and either $\lambda_k = 0$ or $\lambda_k = 1$, for all $k \in \mathscr{U}_R$. Under this model, the observed register value $z_k = y_k + \lambda_k\mu_k$ with probability one, and λ_k reduces to a simple indicator of error occurrence in z_k. In this case the bias of $\hat{\theta}_{2y}$ can be written as

$$\mathrm{bias}(\hat{\theta}_{2y}) \doteq \sum_{k\in\mathscr{U}_R}(1-\pi_k)(z_k-y_k), \tag{8}$$

and the variance of $\hat{\theta}_{2y}$ simplifies to

$$\begin{aligned} \mathrm{V}(\hat{\theta}_{2y}) &= \mathrm{V}_s\Big[\sum_{k\in s_R}(y_k-z_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big] \\ &= \mathrm{V}_s\Big[\sum_{k\in\mathscr{U}_R}z_k+\sum_{k\in s_R}(y_k-z_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big] \\ &= \mathrm{V}_s\big(\hat{\theta}_{2y}\,\big|\,R\big), \end{aligned}$$

Hence, under this assumption the variance of $\hat{\theta}_{2y}$ can be evaluated by focusing purely on the sampling variance. In the remainder of this paper, we will in fact assume that the errors in the register satisfy this deterministic model.

2.2.3 A Regression-Based Combined Estimator

The last estimator of θ_y that we consider is based on two separate regression models for $\mathscr{U}_R$ and $\mathscr{U}_{NR}$. Specifically:

$$\hat{\theta}_{3y} = \sum_{k\in s_R} w_{3Rk}y_k + \sum_{k\in s_{NR}} w_{3NRk}y_k, \tag{9}$$

with

$$w_{3Rk} = \frac{1}{\pi_k}\Big[1+\Big(\sum_{l\in\mathscr{U}_R}z_l-\sum_{l\in s_R}\frac{z_l}{\pi_l}\Big)\Big(\sum_{l\in s_R}\frac{z_l^2}{\pi_l}\Big)^{-1}z_k\Big]$$

the final weight of element $k \in s_R$ and

$$w_{3NRk} = \frac{1}{\pi_k}\Big[1+\mathbf{x}'_{3NRk}\Big(\sum_{l\in s_{NR}}\frac{\mathbf{x}_{3NRl}\mathbf{x}'_{3NRl}}{\pi_l}\Big)^{-1}\Big(\sum_{l\in\mathscr{U}_{NR}}\mathbf{x}_{3NRl}-\sum_{l\in s_{NR}}\frac{\mathbf{x}_{3NRl}}{\pi_l}\Big)\Big]$$

the final weight of element $k \in s_{NR}$.

For the non-registered part of the population, this estimator uses a similar approach to $\hat{\theta}_{2y}$. For $\mathscr{U}_R$, the estimator uses a regression model with the register variable z as predictor variable, since z is likely to be highly correlated with the target variable y. Since the regression estimator is asymptotically unbiased, it holds asymptotically that $\mathrm{E}(\hat{\theta}_{3y}) = \theta_y$. Hence, the advantage of this approach is that it incorporates the information from the register into the estimation process without the risk of introducing a substantial bias. However, this approach is mainly suited for surveys with only one

target variable, since otherwise it leads to a different set of regression weights for each target variable. In particular, this type of estimator is not useful if the objective of the survey is to create a general purpose data file for researchers.

It is not difficult to see that, in the extreme case that the register contains no measurement errors at all, i.e. $z_k = y_k$ for all $k \in \mathscr{U}_R$, the two estimators $\hat{\theta}_{2y}$ and $\hat{\theta}_{3y}$ become identical if they use the same weighting model for s_{NR}.

We observe that $\hat{\theta}_{3y}$ can in fact be written as an ordinary regression estimator, which – unlike $\hat{\theta}_{1y}$ – uses auxiliary information from the register. To see this, define a new auxiliary vector $\mathbf{x}_{3k}$ for each $k \in \mathscr{U}$ by:

$$\mathbf{x}_{3k} = \begin{cases} (z_k, \mathbf{0}')' & \text{if } k \in \mathscr{U}_R \\ (0, \mathbf{x}'_{3NRk})' & \text{if } k \in \mathscr{U}_{NR} \end{cases}$$

and define new regression weights

$$w_{3k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{3k}\big(\sum_{l\in s} \frac{\mathbf{x}_{3l}\mathbf{x}'_{3l}}{\pi_l}\big)^{-1}\big(\sum_{l\in\mathscr{U}} \mathbf{x}_{3l} - \sum_{l\in s} \frac{\mathbf{x}_{3l}}{\pi_l}\big)\Big].$$

Then it is easily derived that $w_{3k} = w_{3Rk}$ for all $k \in s_R$ and $w_{3k} = w_{3NRk}$ for all $k \in s_{NR}$. Therefore it holds that $\hat{\theta}_{3y} = \sum_{k\in s} w_{3k} y_k$.

Finally, we remark that under the deterministic error model from Section 2.2.2, it clearly holds that $\mathrm{V}(\hat{\theta}_{3y}) = \mathrm{V}_s(\hat{\theta}_{3y} \,|\, R)$.

3 A Bootstrap Method for Combined Data

3.1 Introduction to the Bootstrap

Loosely speaking, the bootstrap idea is to mimic the process that generated the originally observed data, by estimating the underlying distribution from the sample and then resampling from this estimated distribution. In some special cases the bootstrap can be performed analytically, but usually one resorts to Monte Carlo approximation, by generating a large number of bootstrap replicates of the target estimate. These replicates are obtained by taking the algorithm that produces the original estimate when applied to the original sample, and applying it to resamples taken from the estimated distribution. We refer to Efron and Tibshirani (1993) for an introduction to the classical bootstrap.

An important problem with the classical bootstrap arises when it is applied to finite population sampling, namely how to mimic the effect of sampling without replacement. In order to obtain a valid measure of the variance of an estimate, it is crucial to capture the effect of the sampling design. In particular, sampling without replacement leads to a smaller variance than sampling with replacement.

There are various methods suggested in the literature to adapt the classical bootstrap to finite population sampling, including the with-replacement bootstrap (McCarthy and Snowden, 1985), the rescaled bootstrap (Rao and Wu, 1988), the mirror-match bootstrap (Sitter, 1992*b*) and the without-replacement bootstrap (Gross, 1980; Bickel and Freedman, 1984; Chao and Lo, 1985; Sitter, 1992*a*). A summary of these methods

can be found in Shao and Tu (1995). However, these methods tend to be difficult to apply in practice. Antal and Tillé (2011) describe yet another bootstrap method for finite population sampling.

A newer form of the without-replacement bootstrap has been suggested by Booth et al. (1994), Canty and Davison (1999) and Chauvet (2007). In the next section we describe a variant of the latter method and apply it to the case of combined register and survey data. In line with the deterministic error model from Section 2.2.2, we treat the register data as fixed in this bootstrap method.

3.2 The Proposed Bootstrap Method

The approach taken by Booth et al. (1994) entails generating pseudo-populations. A pseudo-population is an estimated version of the target population, obtained by taking d_k copies of each element from the original sample, where $d_k = 1/\pi_k$ is the inclusion weight. Bootstrap resamples are drawn by applying the original sampling design to the pseudo-population, and a replicate of the original estimator is calculated from each bootstrap resample. Finally, estimates of the accuracy of the original estimator, such as its variance or confidence intervals, are obtained from the distribution of these replicates, analogous to the classical bootstrap method.

In general d_k need not be an integer, which makes it necessary to round the inclusion weights. Writing $d_k = \lfloor d_k \rfloor + \varphi_k$ (with $\varphi_k \in [0,1)$), a stochastic form of rounding is used that rounds d_k down to $\lfloor d_k \rfloor$ with probability $1 - \varphi_k$, and up to $\lfloor d_k \rfloor + 1$ with probability φ_k.[2] In order to eliminate the effect of the stochastic rounding on the outcome of the bootstrap method, multiple pseudo-populations can be formed, each based on a different rounding of the inclusion weights.

The diagram in Figure 2 summarises the bootstrap method. In this description, B denotes the number of constructed pseudo-populations and C the number of replicates computed from each pseudo-population. The total number of bootstrap replicates equals $B \times C$. Suitable values of B and C are discussed in Section 4.

Following results of Chauvet (2007), a single pseudo-population could also be used as an approximation of the above-mentioned approach. Using a single pseudo-population is, as one would expect, less computer-intensive and faster than using multiple pseudo-populations. The bootstrap method with a single pseudo-population is obtained as a special case of the algorithm in Figure 2 with $B = 1$, so that Steps 1 to 3 are only run once. Note that compared to the multiple pseudo-population approach, a higher value of C is now needed to achieve convergence of the Monte Carlo approximation. In the simulation study in Section 4, both the multiple and single pseudo-population approach are investigated.

In the above algorithm we have not defined which estimator is used specifically. In fact, a different choice of $t(.)$ is used in Step 2 of the algorithm, depending on the

[2] This stochastic rounding can be executed in different ways. Kuijvenhoven and Scholtus (2010) apply Fellegi's method for consistent rounding directly to the inclusion weights. Booth et al. (1994) and Chauvet (2007) round the weights implicitly, by taking $\lfloor d_k \rfloor$ copies of each element from the original sample and then drawing an additional subsample from the original sample using the drawing probabilities φ_k.

Figure 2: A bootstrap algorithm for finite population sampling.

Step 1 Writing $d_k = \lfloor d_k \rfloor + \varphi_k$, define a random inflation weight $\delta_k = \lfloor d_k \rfloor$ with probability $1 - \varphi_k$ and $\delta_k = \lfloor d_k \rfloor + 1$ with probability φ_k. Generate a pseudo-population $\hat{\mathscr{U}}$ by taking δ_k copies of each element k from the original sample s.

Step 2 Draw a sample s^* from $\hat{\mathscr{U}}$ with the original sample design. That is, for $j \in \hat{\mathscr{U}}$ there is an inclusion probability $\pi_j^* \propto \pi_k$, if j is a copy of $k \in s$, where the π_j^* are scaled so that $\sum_{j \in \hat{\mathscr{U}}} \pi_j^*$ equals the original sample size. For each bootstrap resample, compute the replicate $\hat{\theta}^* = t(s^*, R)$, where $t(.)$ denotes the algorithm such that $\hat{\theta} = t(s, R)$.

Step 3 Step 2 is repeated C times to obtain replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_C^*$. From these replicates, compute:

$$\begin{aligned} v_{boot} &= \frac{1}{C-1} \sum_{c=1}^{C} \left(\hat{\theta}_c^* - \overline{\hat{\theta}^*} \right)^2 \\ \overline{\hat{\theta}^*} &= \frac{1}{C} \sum_{c=1}^{C} \hat{\theta}_c^* \end{aligned}$$

Step 4 Steps 1 to 3 are repeated B times to obtain $v_{boot}^1, \cdots, v_{boot}^B$. The estimated variance of the original estimator is

$$v_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} v_{boot}^b.$$

estimator. For the estimators from Section 2.2, we define the following expressions for the bootstrap replicate $\hat{\theta}^* = t(s^*, R)$:

$$\begin{aligned} t_1(s^*, R) &= \sum_{j \in s^*} w_{1j}^* y_j^* \\ t_2(s^*, R) &= \sum_{k \in \mathscr{U}_R} z_k + \sum_{j \in s_R^*} (y_j^* - z_j^*) + \sum_{j \in s_{NR}^*} w_{2j}^* y_j^* \\ t_3(s^*, R) &= \sum_{j \in s_R^*} w_{3Rj}^* y_j^* + \sum_{j \in s_{NR}^*} w_{3NRj}^* y_j^* \end{aligned}$$

In these expressions the following notation is used: s_R^* and s_{NR}^* denote the parts of the resample that consist of copies of elements from s_R and s_{NR}, respectively; y_j^* and z_j^* are by definition equal to y_k and z_k if j is a copy of k; $w_{.j}^*$ denotes a regression weight obtained from the bootstrap resample by applying the same regression model that led to $w_{.k}$ in the original sample. Thus for each bootstrap resample a new set of regression weights is obtained. In this manner the effect on the variance of the estimator due to weighting is taken into account.

Due to nonresponse only a part of the original sample is usually observed in practice. Note that with nonresponse present also nonrespondents are duplicated in the pseudo-population. Therefore, nonresponse will also occur in the bootstrap resamples, namely

when copies of original nonrespondents are drawn. Canty and Davison (1999) argue that the bootstrap is valid, provided that the same weighting model that was used to correct for nonresponse in the original sample is also applied to each bootstrap resample, under the assumption that the weighting model indeed explains nonresponse behaviour. Through this approach, each weighted bootstrap resample will correctly represent the original population. Shao and Sitter (1996) use a similar approach, but they impute data for nonrespondents instead of weighting.

4 Simulation Study

In this section, we assess the correctness of the bootstrap method from Section 3.2 in a simulation study. For this simulation we used a small subset of the Dutch Educational Attainment File (EAF) as our target population. The EAF contains information on the highest attained education level of persons living in the Netherlands. Moreover, the EAF can be linked to other files containing background variables for these persons. The information on educational attainment is obtained from the Dutch educational registrations and from the Labour Force Survey. For persons that are present in more than one source, the scores for education levels in different sources are compared and one of the values is chosen (usually the highest one). This process is called harmonisation. We refer to Linder and Van Roon (2011) for more details on the EAF.

As our target population we select a subset of $49,647$ persons aged over 14 years[3] from the EAF. For the purpose of this simulation study, the file containing the records of these $49,647$ persons is considered to be a complete enumeration of the target population. In this target population, register information is available for $8,904$ persons, so the size of $\mathscr{U}_R$ is $8,904$. The remaining $40,743$ persons, for which no register information is available, constitute the subpopulation $\mathscr{U}_{NR}$. Note that for persons in $\mathscr{U}_R$, the education level from the register may differ from the final, harmonised education level. Using the notation from Section 2, the true value y_k corresponds to the harmonised education level and the register value z_k corresponds to the unharmonised education level from the register. For the purpose of the simulation study, differences between these two values are considered to be caused by errors in the register.

Next, we draw samples from our target population and use these samples, together with $\mathscr{U}_R$, to estimate certain parameters of the target population. Since the true values of these target parameters are also known in this study, the theoretical accuracy of the survey estimators can be measured directly. We also compute estimates of accuracy using the bootstrap method, and compare these with the theoretical accuracy. In order to comply with the assumption from Section 2.1 that measurement errors only occur in the register, for the purpose of this simulation study, we always take the harmonised education levels as observed data in our samples.

A stratified simple random sampling design is used to draw the samples, where the stratification is by *Sex* (values: Male and Female) and *Age* (values: Young, Middle, and Old). The total sample size equals $3,615$. The sampling fractions are: 30% for the

[3]The education levels are sometimes deductively imputed for persons younger than 15 years, so these cannot be considered as typical register or sample data.

two strata with $Age =$ Young, 6% for the two strata with $Age =$ Middle, and 3% for the two strata with $Age =$ Old. These sampling fractions are chosen such that the corresponding d_k have large non-integer parts: the inclusion weights are 3 $1/3$, 16 $2/3$, and 33 $1/3$ for young persons, middle-aged persons, and old persons, respectively. Thus, we expect to see a relatively large effect of the stochastic rounding on the outcome of the bootstrap method.

The education levels come from a hierarchic classification. The highest level of the classification consists of five codes, ranging from primary education (code 1) to education leading to a master's degree or higher (code 5). In this simulation study, we estimate the number of persons with educational attainment code 3 (which corresponds to the second stage of secondary education) in each stratum. These parameters can be seen as population totals of suitably defined binary variables. Table 1 displays the actual values in the target population.

We use the three estimators (4), (5), and (9) from Section 2 to estimate the target parameters. The regression weights w_{1k}, w_{2k}, and w_{3NRk} are obtained using the following linear model:

$$Region(5) \times Age(3) + Region(5) \times Sex(2) \times Marital\ Status(3),$$

where the number in brackets denotes the number of classes for each auxiliary variable.

The true statistical properties of the three estimators are approximated by drawing $20,000$ samples from the target population. Table 1 displays the approximate standard errors of the three estimators based on these $20,000$ realisations. Since estimator 2 is known to be potentially biased, Table 1 also shows the approximate relative bias of this estimator based on $20,000$ realisations. Since the target population is completely known, the bias of estimator 2 can also be calculated directly from expression (8), which leads to similar values.

It is seen that the register count overestimates the number of persons with educational attainment code 3. This bias is caused by the fact that for some persons with educational attainment code 4, not all forms of education attained by these persons have been properly registered, so that the reported educational attainment code 3 in the register is too low by mistake. Of course, this effect could be neutralised by the presence of persons with an actual educational attainment code 3 who are registered with a lower educational attainment code, but apparently the latter type of error occurs less often, so that the net result is a positive bias. The bias is much larger for the strata of young persons than for the other strata, because, as mentioned in Section 2.1, older persons are underrepresented in the register. In fact, hardly any register information is available from the strata of old persons, so that the three estimators are actually almost identical for these strata (which explains why the standard errors are more or less equal).

For the above-mentioned results it is assumed that all persons respond when sampled. It is more realistic to assume that some nonresponse occurs in the sample. To keep matters simple, we adopt the so-called fixed response model (Bethlehem et al., 2011), whereby each person in the target population either always responds or never responds when sampled. The response indicators are randomly assigned to the persons in the target population, in such a way that the weighting model in *Region*, *Age*, *Sex*, and

Table 1: Target parameters, standard errors, and relative bias based on 20,000 simulations. Abreviations: YM = Young Males, MM = Middle-Aged Males, OM = Old Males, YF = Young Females, MF = Middle-Aged Females, OF = Old Females.

	YM	MM	OM	YF	MF	OF
target parameters						
	1,178	4,459	4,423	1,164	5,386	3,880
standard errors (with full response)						
estimator 1	49	203	298	48	208	293
estimator 2	36	186	297	36	190	291
estimator 3	40	190	297	38	193	291
relative bias (with full response)						
estimator 2	+12%	+3%	+0%	+7%	+2%	+0%
standard errors (with nonresponse)						
estimator 1	58	241	349	57	248	341
estimator 2	42	223	347	43	229	339
estimator 3	47	227	347	45	231	339
relative bias (with nonresponse)						
estimator 2	+13%	+3%	−0%	+8%	+1%	−1%

Marital Status explains most of the nonresponse behaviour. The last two sections of Table 1 show the approximate standard errors of the three estimators and the relative bias of estimator 2 with nonresponse, again based on 20,000 realisations.

The approximate standard errors in Table 1 serve as a benchmark for the bootstrap results to be discussed below.

In order to apply the bootstrap method proposed in Section 3.2, suitable values for B and C have to be chosen. Chauvet (2007) reports results based on $B = 100$ and $C = 30$ for the multiple pseudo-population approach, and $C = 1,000$ for the single pseudo-population approach. In contrast to Chauvet (2007), we consider only variance estimates and not the estimation of bootstrap confidence intervals in this study. It is acknowledged in the bootstrap literature that, compared to the estimation of confidence intervals, a smaller number of replicates suffices for variance estimation. Therefore, to limit the amount of computational work, we choose $B = 50$ and $C = 30$ for the multiple pseudo-population approach in this simulation study. For the single pseudo-population approach, we choose $C = 1,000$.

Table 2 reports the estimated standard errors for the three estimators obtained from the bootstrap method with multiple pseudo-populations, both without and with nonresponse. To trace the sampling variability of the bootstrap estimates, these results are based on 20 realisations of the bootstrap method, and Table 2 shows both the mean and the relative standard deviation of 20 bootstrap estimates. Similar results for the

Table 2: Standard errors from the bootstrap method with multiple pseudo-populations (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	50	203	299	49	212	289
	(4%)	(2%)	(2%)	(3%)	(2%)	(3%)
estimator 2	37	188	297	38	195	289
	(7%)	(2%)	(2%)	(6%)	(2%)	(3%)
estimator 3	40	192	297	40	198	289
	(6%)	(2%)	(2%)	(5%)	(2%)	(3%)
standard errors (with nonresponse)						
estimator 1	60	244	350	59	250	339
	(9%)	(2%)	(3%)	(4%)	(2%)	(4%)
estimator 2	45	230	347	47	233	337
	(9%)	(3%)	(3%)	(9%)	(3%)	(3%)
estimator 3	49	233	347	49	237	337
	(8%)	(2%)	(3%)	(9%)	(3%)	(3%)

bootstrap method with a single pseudo-population are reported in Table 3.

These results do not exhibit large differences between the multiple and single pseudo-population approaches. The estimated standard errors are in both cases close to the approximate true values from Table 1, with a tendency to slightly overestimate the standard errors in most strata. The relative standard deviations of the bootstrap estimates are small and the two approaches perform about equally well in this respect also. The similar performance of the multiple and single pseudo-population approaches seen here is in line with results reported by Chauvet (2007) in a simulation study involving an artificial population of normally distributed data.

For the stratified simple random sampling design used in this simulation study, a practical alternative method for estimating the variance of a regression estimator is to apply Taylor linearisation (Särndal et al., 1992; Knottnerus, 2003). For estimator 1, the following variance estimator is readily found in the literature:

$$\hat{\mathrm{V}}(\hat{\theta}_{1y}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s^2_{\hat{\varepsilon}_{1h}}}{n_h},$$

where H denotes the number of strata (in this case: $H = 6$), N_h and n_h denote the population and sample size in stratum h, and $s^2_{\hat{\varepsilon}_{1h}}$ is the sample variance of the residuals of the fitted regression model. A similar expression is found for estimator 3, since we already noted that this estimator can be written in the same form as estimator 1. For estimator 2, we use the following variance estimator:

$$\hat{\mathrm{V}}(\hat{\theta}_{2y}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s^2_{u_{1h}} + s^2_{u_{2h}} - 2s_{u_{1h}u_{2h}}}{n_h},$$

Table 3: Standard errors from the bootstrap method with a single pseudo-population (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	50	205	297	50	209	290
	(5%)	(2%)	(3%)	(4%)	(2%)	(3%)
estimator 2	37	189	295	38	192	289
	(6%)	(3%)	(3%)	(6%)	(2%)	(3%)
estimator 3	41	193	295	41	195	290
	(6%)	(3%)	(3%)	(5%)	(2%)	(3%)
standard errors (with nonresponse)						
estimator 1	59	246	349	59	248	337
	(7%)	(3%)	(3%)	(4%)	(3%)	(4%)
estimator 2	45	230	346	47	232	335
	(8%)	(3%)	(3%)	(8%)	(3%)	(4%)
estimator 3	49	235	346	49	235	335
	(7%)	(3%)	(3%)	(7%)	(3%)	(4%)

with

$$u_{1hk} = \begin{cases} (y_{hk} - z_{hk}) n_h / N_h & \text{if } k \in s_R \\ 0 & \text{if } k \in s_{NR} \end{cases}$$

and

$$u_{2hk} = \begin{cases} 0 & \text{if } k \in s_R \\ \hat{\varepsilon}_{2hk} & \text{if } k \in s_{NR} \end{cases}$$

Here, $s_{u_{1h}u_{2h}}$ denotes the sample covariance of u_1 and u_2 in stratum h. This formula is obtained from the derivation given in appendix A below expression (12), for the particular case of stratified simple random sampling and the deterministic error model. In addition, the population (co)variances are estimated by their sample equivalents.

Table 4 reports the results of 20 realisations of the linearisation method. In contrast to the bootstrap method, it is seen that the linearisation method tends to underestimate the standard errors, and this effect becomes more pronounced in the situation with nonresponse. On the other hand, the linearisation method appears to be less sensitive to sampling variability, since the relative standard deviations of the 20 realisations are mostly smaller than for the bootstrap method.

To conclude, we give some information on the practical execution of the above simulation study. Most of the computational work for the bootstrap method was done in Blaise, a survey processing system developed at Statistics Netherlands. The bootstrap method was implemented as a series of so-called Manipula setups in Blaise, and the Blaise weighting tool Bascula was used to compute the regression weights. Finally, the statistical software R was used to compile and analyse the results of the simulation

Table 4: Standard errors from the linearisation method (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	48	200	294	47	206	284
	(2%)	(1%)	(1%)	(2%)	(< 1%)	(2%)
estimator 2	33	184	292	34	187	282
	(2%)	(1%)	(1%)	(2%)	(1%)	(2%)
estimator 3	37	187	292	37	190	282
	(2%)	(1%)	(1%)	(2%)	(1%)	(2%)
standard errors (with nonresponse)						
estimator 1	56	229	343	56	235	325
	(2%)	(1%)	(2%)	(2%)	(1%)	(2%)
estimator 2	39	211	340	40	213	323
	(2%)	(1%)	(2%)	(2%)	(1%)	(3%)
estimator 3	43	214	339	43	216	322
	(2%)	(1%)	(2%)	(2%)	(1%)	(2%)

study. The estimated standard errors for the linearisation method were also calculated in R.

5 Discussion

In this paper, we have described different estimators based on a combination of register data and sample data, and we have introduced a bootstrap method for assessing the variance of these estimators. Moreover, the performance of the bootstrap method was examined in a simulation study, using realistic data from the Dutch Educational Attainment File. The results of this simulation study show that the bootstrap provides valid variance estimates for estimators based on combined data, and that the quality of the bootstrap estimates compares favourably to an alternative method based on linearisation. It also appears from the study that the bootstrap method with a single pseudo-population is not outperformed by the multiple pseudo-population approach, although the latter has a more sound theoretical basis (Chauvet, 2007). The single pseudo-population approach is less complex than the multiple pseudo-population approach and, in principle, requires less computational work. However, in practice both approaches require the computation of a similar number of replicates; in our simulation study, the total number of replicates BC equals $1,500$ for the multiple pseudo-population approach and $1,000$ for the single pseudo-population approach.

Given a combination of register data and sample survey data, there are of course many different estimators that one could consider. In this paper we have only treated three such estimators. Another interesting estimator, suggested by De Heij (2011), is the

following:

$$\hat{\theta}_{4y} = \sum_{k \in \mathscr{U}_R} z_k + \sum_{k \in s_R} \frac{y_k - z_k}{\pi_k} + \sum_{k \in s_{NR}} w_{4k} y_k,$$

with w_{4k} a regression weight, defined analogously to w_{2k}. Like our additive estimator $\hat{\theta}_{2y}$, this estimator uses a regression estimator for $\mathscr{U}_{NR}$ and an adjusted register total for $\mathscr{U}_R$, where the adjustment is based on information from s_R. In $\hat{\theta}_{2y}$ the adjustment term only corrected the observed individual errors in the overlap. Consequently, as was confirmed in the simulation study, this estimator is stable but it may have a large bias. In $\hat{\theta}_{4y}$, the adjustment term is based on a Horvitz-Thompson estimate of the total error in the register. This approach has the advantage that it leads to an asymptotically unbiased estimator, unlike $\hat{\theta}_{2y}$. On the other hand, the variance of the adjustment term – and hence of the estimator as a whole – might be large. It would be interesting for a future study to compare the performance of $\hat{\theta}_{4y}$ and the other estimators in practical situations.

The bootstrap method described here only considers the variance due to sampling and treats the observed register data as fixed. In Section 2.1 we considered a general measurement error model for register values, which includes the possibility of stochastic errors in the register. From a theoretical point of view, it might be an interesting topic for future research to extend the bootstrap method so that it can also be used when the errors in the register are of a stochastic nature. However, a practical application of this theory would require accurate estimates of the model parameters λ_k, μ_k, and σ_k^2, and these might be difficult to obtain if s_R is our only source of information.

Another assumption made in this paper is that the target variable is observed without error in the sample survey, or, if errors do occur, that the effect of these errors on the estimates is negligible compared to the sampling variance. It may be of interest to relax this assumption and to also assume a model for measurement errors in the sample survey. Note that this implies that more complex estimators are needed, because we can no longer simply use the sample data to correct errors in the register data.

Acknowledgements

The research for this paper was conducted as part of Work Package 2 (Task 2.1.2) of the ESSnet project Data Integration. The authors would like to thank Li-Chun Zhang, Johan Fosen, Paul Knottnerus, Jeroen Pannekoek, Frank Linder, and Dominique van Roon for their comments on earlier drafts of this paper.

References

Antal, E. and Tillé, Y. (2011), 'A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population', *Journal of the American Statistical Association* **106**, pp. 534–543.

Bethlehem, J., Cobben, F. and Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*, John Wiley & Sons, New Jersey.

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

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

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

Chao, M.-T. and Lo, S.-H. (1985), 'A Bootstrap Method for Finite Populations', *Sankhyā Series A* **47**, pp. 399–405.

Chauvet, G. (2007), Méthodes de Bootstrap en Population Finie, PhD thesis, L'Université de Rennes.

De Heij, V. (2011), 'Samples and Registers', Internal Memo (in Dutch), Statistics Netherlands, The Hague.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, London.

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

Knottnerus, P. (2003), *Sample Survey Theory: Some Pythagorean Perspectives*, Springer-Verlag, New York.

Kuijvenhoven, L. and Scholtus, S. (2010), 'Estimating Accuracy for Statistics Based on Register and Survey Data', Discussion Paper 10007, Statistics Netherlands, The Hague.

Linder, F. and Van Roon, D. (2011), 'Combining Data from Administrative Sources and Sample Surveys; the Single-Variable Case. Case Study: Educational Attainment', Report for Work Package 4.2 of the ESSnet project Data Integration.

McCarthy, P. J. and Snowden, C. B. (1985), 'The Bootstrap and Finite Population Sampling', Technical Report, National Center for Health Statistics.

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

Särndal, C., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

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

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

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

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

Appendix A Derivation of Expressions for the Bias and Variance of $\hat{\theta}_{2y}$

We begin by evaluating the bias of $\hat{\theta}_{2y}$. From $\mathrm{E}(\hat{\theta}_{2y}) = \mathrm{E}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)]$ and expression (2), we obtain:

$$\begin{aligned}
\mathrm{E}(\hat{\theta}_{2y}) &= \mathrm{E}_s\Big[\mathrm{E}\big(\sum_{k\in\mathscr{U}_R} z_k + \sum_{k\in s_R}(y_k - z_k) + \sum_{k\in s_{NR}} w_{2k}y_k \,\big|\, s\big)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathscr{U}_R}\mathrm{E}(z_k) + \sum_{k\in s_R}\big(y_k - \mathrm{E}(z_k)\big) + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathscr{U}_R}(y_k + \lambda_k\mu_k) - \sum_{k\in s_R}\lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \sum_{k\in\mathscr{U}_R}(y_k + \lambda_k\mu_k) + \mathrm{E}_s\Big(-\sum_{k\in s_R}\frac{\pi_k\lambda_k\mu_k}{\pi_k} + \sum_{k\in s_{NR}} w_{2k}y_k\Big) \\
&\doteq \sum_{k\in\mathscr{U}_R}(y_k + \lambda_k\mu_k) - \sum_{k\in\mathscr{U}_R}\pi_k\lambda_k\mu_k + \sum_{k\in\mathscr{U}_{NR}} y_k \\
&= \theta_y + \sum_{k\in\mathscr{U}_R}(1-\pi_k)\lambda_k\mu_k,
\end{aligned}$$

In the second last line, it is used that $\sum_{k\in s_R} x_k/\pi_k$ is an unbiased estimator for $\sum_{k\in\mathscr{U}_R} x_k$, for any variable x. It is also used that $\sum_{k\in s_{NR}} w_{2k}y_k$ is an asymptotically unbiased estimator for $\sum_{k\in\mathscr{U}_{NR}} y_k$. Expression (6) now follows.

Next, we evaluate the variance of $\hat{\theta}_{2y}$ by means of the decomposition

$$\mathrm{V}(\hat{\theta}_{2y}) = \mathrm{E}_s[\mathrm{V}(\hat{\theta}_{2y} \mid s)] + \mathrm{V}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)].$$

Using the assumption that the z_k are independent, it follows from expression (3) that

$$\begin{aligned}
\mathrm{E}_s[\mathrm{V}(\hat{\theta}_{2y} \mid s)] &= \mathrm{E}_s\Big[\mathrm{V}\big(\sum_{k\in\mathscr{U}_R\setminus s_R} z_k \,\big|\, s\big)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathscr{U}_R\setminus s_R}\mathrm{V}(z_k)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathscr{U}_R\setminus s_R}\lambda_k(\sigma_k^2 + \mu_k^2(1-\lambda_k))\Big] \\
&= \sum_{k\in\mathscr{U}_R}(1-\pi_k)\lambda_k[\sigma_k^2 + \mu_k^2(1-\lambda_k)]. \qquad (10)
\end{aligned}$$

The proof of the last line is analogous to the last four lines in the evaluation of $\mathrm{E}(\hat{\theta}_{2y})$.

For the second component, we find

$$\begin{aligned}
\mathrm{V}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)] &= \mathrm{V}_s\Big[\sum_{k\in\mathscr{U}_R}(y_k + \lambda_k\mu_k) - \sum_{k\in s_R}\lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \mathrm{V}_s\Big(-\sum_{k\in s_R}\lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big). \qquad (11)
\end{aligned}$$

Combining (10) and (11) yields expression (7).

It is interesting to examine expression (11) in more detail. The weights w_{2k} have been found by fitting a regression model to the observations from $\mathscr{U}_{NR}$, say, $y_k = \beta_2'\mathbf{x}_{2k} + \varepsilon_{2k}$. Denote the vector of fitted regression coefficients by $\hat{\beta}_2$. By a standard argument, it

holds that

$$\begin{aligned}\sum_{k\in s_{NR}} w_{2k}y_k &= \sum_{k\in s_{NR}}\frac{y_k}{\pi_k}+\hat{\beta}_2'\Big(\sum_{k\in\mathscr{U}_{NR}}\mathbf{x}_{2k}-\sum_{k\in s_{NR}}\frac{\mathbf{x}_{2k}}{\pi_k}\Big)\\ &\doteq \sum_{k\in s_{NR}}\frac{y_k}{\pi_k}+\beta_2'\Big(\sum_{k\in\mathscr{U}_{NR}}\mathbf{x}_{2k}-\sum_{k\in s_{NR}}\frac{\mathbf{x}_{2k}}{\pi_k}\Big)\\ &= \beta_2'\sum_{k\in\mathscr{U}_{NR}}\mathbf{x}_{2k}+\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k},\end{aligned}$$

since the discarded term $(\hat{\beta}_2-\beta_2)'\Big(\sum_{k\in\mathscr{U}_{NR}}\mathbf{x}_{2k}-\sum_{k\in s_{NR}}\frac{\mathbf{x}_{2k}}{\pi_k}\Big)$ is asymptotically irrelevant. Hence, for sufficiently large samples, we have

$$\begin{aligned}\mathrm{V}_s\Big(-\sum_{k\in s_R}\lambda_k\mu_k+\sum_{k\in s_{NR}}w_{2k}y_k\Big) &\doteq \mathrm{V}_s\Big(-\sum_{k\in s_R}\lambda_k\mu_k+\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\ &= \mathrm{V}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k\Big)+\mathrm{V}_s\Big(\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\ &\quad -2\mathrm{Cov}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k,\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big).\end{aligned}\tag{12}$$

Note that $\lambda_k\mu_k$ is only defined for $k\in\mathscr{U}_R$, while ε_{2k} is only defined for $k\in\mathscr{U}_{NR}$. For convenience, define $\lambda_k\mu_k=0$ for $k\in\mathscr{U}_{NR}$, and define $\varepsilon_{2k}=0$ for $k\in\mathscr{U}_R$. The first variance term may now be evaluated as follows:

$$\begin{aligned}\mathrm{V}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k\Big) &= \mathrm{V}_s\Big(\sum_{k\in s}\frac{\pi_k\lambda_k\mu_k}{\pi_k}\Big)\\ &= \sum_{k\in\mathscr{U}}\sum_{l\in\mathscr{U}}(\pi_{kl}-\pi_k\pi_l)\frac{\pi_k\lambda_k\mu_k}{\pi_k}\frac{\pi_l\lambda_l\mu_l}{\pi_l}\\ &= \sum_{k\in\mathscr{U}_R}\sum_{l\in\mathscr{U}_R}(\pi_{kl}-\pi_k\pi_l)\lambda_k\mu_k\lambda_l\mu_l,\end{aligned}$$

where we have used a standard formula for the variance of a Horvitz-Thompson estimator; see e.g. Särndal et al. (1992, p. 43). Similarly, the second variance term yields

$$\mathrm{V}_s\Big(\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)=\sum_{k\in\mathscr{U}_{NR}}\sum_{l\in\mathscr{U}_{NR}}(\pi_{kl}-\pi_k\pi_l)\frac{\varepsilon_{2k}}{\pi_k}\frac{\varepsilon_{2l}}{\pi_l}.$$

Finally, the covariance term may be evaluated as follows:

$$\begin{aligned}\mathrm{Cov}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k,\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big) &= \mathrm{Cov}_s\Big(\sum_{k\in s}\frac{\pi_k\lambda_k\mu_k}{\pi_k},\sum_{k\in s}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\ &= \sum_{k\in\mathscr{U}}\sum_{l\in\mathscr{U}}(\pi_{kl}-\pi_k\pi_l)\frac{\pi_k\lambda_k\mu_k}{\pi_k}\frac{\varepsilon_{2l}}{\pi_l}\\ &= \sum_{k\in\mathscr{U}_R}\sum_{l\in\mathscr{U}_{NR}}(\pi_{kl}-\pi_k\pi_l)\lambda_k\mu_k\frac{\varepsilon_{2l}}{\pi_l}.\end{aligned}\tag{13}$$

In the second last line, use is made of a standard formula for the covariance of two Horvitz-Thompson estimators; see e.g. Särndal et al. (1992, p. 170).

In general, expression (13) may be non-zero. There exist, however, a few special cases where the covariance term always vanishes. For simple random sampling, we have $\pi_k=\pi_l=\frac{n}{N}$, $\pi_{kl}=\frac{n(n-1)}{N(N-1)}$ (for $k\neq l$), and hence

$$\sum_{k\in\mathscr{U}_R}\sum_{l\in\mathscr{U}_{NR}}(\pi_{kl}-\pi_k\pi_l)\lambda_k\mu_k\frac{\varepsilon_{2l}}{\pi_l}=\Big[\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\Big]\sum_{k\in\mathscr{U}_R}\lambda_k\mu_k\frac{N}{n}\sum_{l\in\mathscr{U}_{NR}}\varepsilon_{2l}=0,$$

since the sum of the residuals over $\mathscr{U}_{NR}$ equals zero by construction. Similarly, the covariance term also vanishes for stratified simple random sampling, provided that a separate regression model is fitted for each stratum.