Ängsved, Marianne

Working Paper

Estimating the finite population total under frame imperfections and nonresponse

Working Paper, No. 4/2006

Provided in Cooperation with:
Örebro University School of Business

Suggested Citation: Ängsved, Marianne (2006) : Estimating the finite population total under frame imperfections and nonresponse, Working Paper, No. 4/2006, Örebro University School of Business, Örebro

This Version is available at:
https://hdl.handle.net/10419/244423

WORKING PAPER SERIES

WORKING PAPER NO 4, 2006

ESI

Estimating the finite population total under frame imperfections and nonresponse

by

Marianne Ängsved

http://www.oru.se/esi/wps

SE-701 82 Örebro
SWEDEN

ISSN 1403-0586

Estimating the finite population total under frame imperfections and nonresponse

Marianne Ängsved[a]

Abstract
When sampling from a finite population the access to a good sampling frame is of vital importance. However, the statistician often has to face the problem with non-negligible frame imperfections, e.g. overcoverage and undercoverage. More so, error from nonresponse is an increasing problem in many surveys today. In this paper we discuss different approaches to deal with these problems simultaneously. In particular, we address the situation when there exists a new up-to-date current register and the improvement this brings along.

Keywords: Finite population sampling, target population, sampling frame, overcoverage, undercoverage, nonresponse, GREG estimator, calibration.

[a] Statistics Sweden, Methods Unit at Business and Labour Market Department, SE-701 89 Örebro, Sweden.

Contents

1 The problem

1.1 Introduction

When planning for a survey there are several decisions the statistician has to make, one being on what sampling frame to use. The access to a good sampling frame is of vital importance. Ideally the sampling frame should be a perfect match to the target population, i.e. to the population the statistician wishes to study. This property is essential for the sampling frame since it allows every element in the population to have a nonzero probability of selection, a requirement for unbiased estimation. Furthermore, a desirable feature of the sampling frame would be the existence of good auxiliary information, since this will be an aid in the statistician's work to find the best possible strategy, i.e. the best possible combination of sampling design and estimator. However, it is far from always possible in practice to come up with a perfect sampling frame. The statistician most often has to accept the fact that the sampling frame has imperfections with respect to matching the target population. Also, the auxiliary information may be more or less useful. Hence, error from frame imperfection may be substantial.

More so, error from nonresponse is an increasing problem in many surveys today. The problem with nonresponse is extensively handled in the literature. Two main approaches are distinguished for dealing with nonresponse, *reweighting* and *imputation.*

Now, the practising statistician will most likely have to deal with these two errors simultaneously. One complication when both nonresponse and frame imperfections occur is that we for elements in the nonresponse set may be unable to determine whether they belong to the target population or to the overcoverage set.

In this paper two main techniques for the estimation of a finite population total under frame imperfections and nonresponse are considered. We will not address any other type of nonsampling error.

1.2 Target population, frame population and related population sets

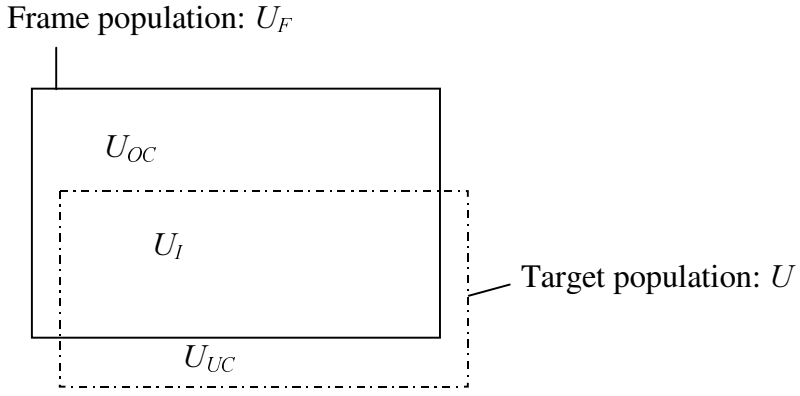
When discussing sampling from a finite population and subsequent estimation of finite population parameters several different population types can be defined, see e.g. Kish (1979) and Murthy (1983). For our purpose we define

three populations. The *target population*, denoted U, is the set of elements the statistician wishes to study, i.e. for which estimates are required. The *frame population*, denoted U_F, is the set of all elements that can be reached via the (sampling) frame. The *current register population*, denoted U_R, is the set of all elements that can be reached via an up-to-date current register. The term *current register population* was introduced in Ängsved (2004). The current register is not at hand at the sampling stage of a survey, but it may be at hand at the estimation stage. The current register could be an updated version of the frame. Or it could be a register which is different from the original frame, e.g. a newly developed register. The register population accessible from the current register matches the target population better than the frame population.

In the ideal situation the target population and the frame population coincide. However, this is far from always the case. Typically there are several types of frame imperfections, see e.g. Lessler and Kalsbeek (1992) who define six sources of error that spring from frame imperfections.

In the following we will suppose that only two types of frame imperfections are present, viz. *overcoverage* and *undercoverage*. Let the set of elements in U which can be reached via the frame be denoted U_I, i.e. $U_I = U_F \cap U$, the intersection of U_F and U. The frame suffers from overcoverage if the set $U_{OC} = U_F$ - U_I is non-empty, where it is assumed that this set can not be identified from available frame information. The frame has undercoverage if the set $U_{UC} = U$ - U_I is non-empty. The two sets will be called the *overcoverage set* and the *undercoverage set*, respectively. Figure 1.1 shows how these sets are related.

Figure 1.1. *Target population and imperfect frame population*



The target population is represented by a dotted line to stress the fact that the target population membership is unknown.

Let $N = \#U$ (the number of elements in the target population), $N_F = \#U_F$, $N_I = \#U_I$, $N_{OC} = \#U_{OC}$ and $N_{UC} = \#U_{UC}$.

1.3 Further notations and definitions

Let s_F denote a sample of size n_{s_F} drawn from the frame population, U_F, with first- and second-order inclusion probabilities $\pi_{Fk} = P(k \in s_F) > 0$ and $\pi_{Fkl} = P(k, l \in s_F) > 0$ for all $k\&l \in U_F$, respectively. Let r_F denote the response set, where $r_F \subset s_F$. The response set is generated by the (unknown) response mechanism $q(r_F | s_F)$. Let o_F denote the nonresponse set. Thus we have $s_F = r_F \cup o_F$.

Let

s_I - sample elements that belong to the target population

s_{OC} - sample elements that belong to the overcoverage population

Borrowing ideas from two-phase sampling we make the following assumptions about the available auxiliary information. The vector $\mathbf{x}_{1k}$, of J_1 values, is the vector with auxiliary information defined according to (1) below. The vector $\mathbf{x}_{2k}$ is a vector with $J_2 = J - J_1$ auxiliary values available for all elements in the sample, i.e. for all $k \in s_F$. Let

$$\mathbf{x}_{I,1k} = \begin{cases} \mathbf{x}_{1k} & k \in U \\ \mathbf{0} & k \in U_{OC} \end{cases} \tag{1}$$

and, for $k \in s_F$, let

$$\mathbf{x}_{I,2k} = \begin{cases} \mathbf{x}_{2k} & k \in s_I \\ \mathbf{0} & k \in s_{OC} \end{cases} \tag{2}$$

Using these definitions we have the combined auxiliary vector

$$\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})' \tag{3}$$

available for $k \in s_I$.

Associated with each element $k \in U$ is a fixed but unknown value y_k, for the study variable y. The parameter of interest is the population total of y,

$$t_{yU} = \sum_{k \in U} y_k = \sum\nolimits_U y_k$$

Let

$$y_k = \begin{cases} y_k & k \in U \\ 0 & k \in U_{OC} \end{cases} \tag{4}$$

To simplify, we assume that all y_k ($k \in U$) are positive. For a single element, let the symbol $\check{}$ symbolize division by π_{Fk}, i.e. $\check{y}_k = y_k/\pi_{Fk}$.

Since U_I and U_{UC} are exhaustive and mutually exclusive on the set U we may write the parameter as

$$t_{yU} = t_{yU_I} + t_{yU_{UC}}$$

If it is possible to identify the set s_I, as the case may be, we can use the theory of domain estimation in order to estimate t_{yU_I}. When there is reason to assume that the undercoverage is negligible it should suffice to use $\hat{t}_{yU_I}$ as estimator for t_{yU}. However, if this is not the case we must find a way to compensate for the negative bias. One way to do this is to find a guesstimate for $t_{yU_{UC}}$, denoted by $t^{\widetilde{app}}_{yU_{UC}}$. The term guesstimate is used here to emphasize the fact that since no values on the study variable exist for the undercoverage set it is not possible to use common design based estimators to estimate the total in this set. Using $\hat{t}_{yU_I}$ together with a guesstimate $t^{\widetilde{app}}_{yU_{Uc}}$ we obtain an estimator for t_{yU}. Another way would be to use direct estimation of t_{yU}.

In section 2 we give a short introduction to estimation under frame imperfections. The estimation setup under both frame imperfections and nonresponse is introduced in section 3. In section 4 two estimators for the total of y in U_I, t_{yU_I}, are considered. Section 5.1 suggests a method to find a guesstimate of the total of y in the undercoverage set, to be used together with t_{yU_I}, while section 5.2 presents a method for direct estimation of the total of y in the target population. In section 6 we present results from a small simulation study. Finally, some concluding remarks are given in section 7.

2 The estimation situation under frame imperfections

In the estimation setup where the only nonsampling error is coverage error the main concern in the estimation process would be the existence of undercoverage. The elements in the undercoverage set has zero probability of being selected for any sample and this may cause bias in the estimate(s).

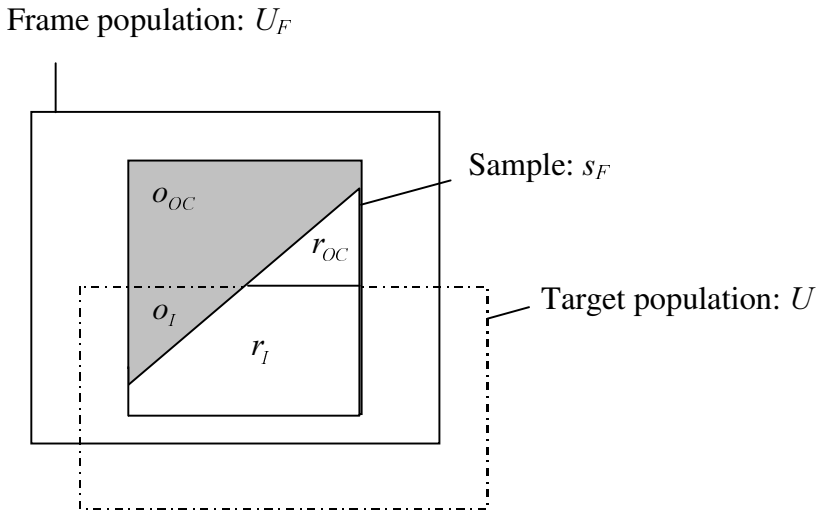
Bias resulting from the overcoverage can be avoided since we assume that we correctly can identify the sample elements that belong to the overcoverage. In this estimation setup, values on the study variable will exist for every element in the sample intersection set, s_I, i.e. the set of sample elements that belong to the target population. Since the intersection set, U_I, is a domain of the frame population we could use an appropriate and (approximately) unbiased domain estimator to estimate the total of U_I. However, using only the estimate of the total of y in the intersection set as an estimate for the total of y in the population will lead to a non-negligible negative bias (since we assume $y_k > 0$ for all $k \in U$) unless the undercoverage is negligible. Ways to adjust for this negative bias will rely on more or less speculative reasoning which may have to be applied to many study variables separately.

However, when we are in the favourable situation where a perfect current register exists at the estimation stage of the survey, the estimation setup improves considerably compared to *the standard estimation setup*, where the only information available comes from the (imperfect) frame. From the current register we can identify, for every $k \in U_F$, whether $k \in U_I$ or $k \in U_{OC}$. Furthermore, we have access to an auxiliary vector for every element in the target population, and hence for every element in the undercoverage set. This means that better estimators can be used for the intersection set. Additionally there are better prospects for guesstimating the total of the undercoverage set (Ängsved, 2004).

3 The estimation situation under frame imperfections and nonresponse - an introduction

One specific complication in the standard estimation setup when both nonresponse and frame imperfections occur is that we can not determine whether a nonresponding element belongs to the target population or to the overcoverage set. Figure 3.1 illustrates the survey situation in the estimation setup with both frame imperfections and nonresponse.

Figure 3.1. *Target population, imperfect frame population and sample when nonresponse has occurred. Standard estimation setup*



The dotted line in figure 3.1 is used to stress the fact that the target population membership is unknown and the shaded area to stress the fact that no there is no information available for the nonresponse set. However, from the response set r_F we can identify two subsets:

- r_I - responding/observed sample elements that belong to the target population
- r_{OC} - responding/observed sample elements that belong to the overcoverage population

Furthermore, we have, two subsets of the nonresponse set o_F:

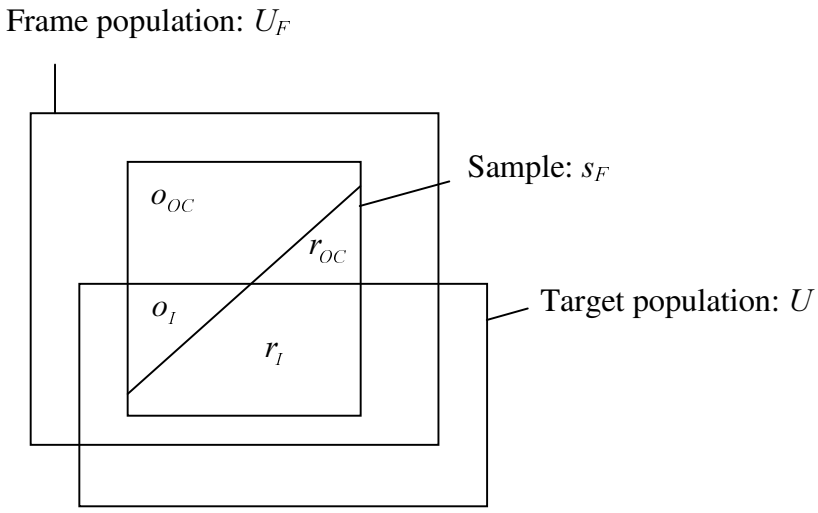
- o_I - nonresponding/non-observed sample elements that belong to the target population
- o_{OC} - nonresponding/non-observed sample elements that belong to the overcoverage population

Note that it is not possible from sample information to identify membership to either of the two subsets of o_F, i.e. to o_I or o_{OC}.

In the standard estimation setup in figure 3.1 no frame information exists for elements $k \in U_F$ on whether $k \in U_I$ or $k \in U_{OC}$. Moreover, the undercoverage is unknown. In this estimation setup the statistician has to make unverifiable assumptions concerning the undercoverage as well as of the overcoverage in the nonresponse set.

When the up-to-date current register exists we are in a much more favourable situation. The following figure illustrates this improved situation when $U_R = U$.

Figure 3.2. *Target population, imperfect frame population and sample when nonresponse has occured. Perfect current register at hand, i.e.* $U_R = U$.



We now have an improved setup for making inference to the target population. One possibility would be to draw a probability sample from the undercoverage set identified in U. This sample would enable us to calculate an objective estimate of $t_{yU_{UC}}$. However, as is often the case, the time schedule and/or the survey budget may not admit the extra selection of elements. Another possibility would be to use the additional information the current register provides. For every element $k \in U_F$, it is possible to identify whether the element belongs to the intersection set, U_I, or to the overcoverage set, U_{OC}. This implies that for elements in the nonresponding set o_F it is possible to identify whether these belong to the overcoverage o_{OC} or to the target population o_I. Thus, it is possible to identify the set s_I and we could use domain estimators under nonresponse to estimate the total t_{U_I}. Moreover, we have current auxiliary information $\mathbf{x}_k$ for all $k \in U$ which thus provides information on the elements in U_{UC}. This information could be used to find guesstimates of $t_{yU_{UC}}$.

It should be noted here that when we specify the "domain" estimator $\hat{t}_{yU_I}$ we do this realizing that only one "domain" total is to be estimated, i.e.

the total of y in U_I. In this paper we do not consider the survey case when estimation is required for other subpopulations.

Henceforth we assume that the current register is perfect in the sense that the register population equals the target population, i.e., $U_R = U$.

4 A perfect current register exists - estimation of t_{yU_I}

As noted earlier, the set U_I constitutes a domain of U_F, and when we are in the favourable situation where we have access to a perfect current register at the estimation stage of the survey it is possible to identify the set s_I in spite of the nonresponse. Thus, in order to estimate the total of U_I we could use common design based domain estimators used in the presence of nonresponse to estimate the total of U_I.

Two main approaches are generally distinguished for dealing with nonresponse, *reweighting* and *imputation*. One frequently used approach is when imputation is used for item nonresponse only and then reweighting is applied to compensate for the unit nonresponse. In this paper we will focus on reweighting.

The *calibration appoach for nonresponse* and the *two-phase approach to reweighting,* where the *response homogeneity groups model* (RHG) is an often used model, are two methods that use reweighting for dealing with nonresponse. The RHG model states that elements respond independently and that the sample consists of nonoverlapping groups with the property that elements within each group respond with the same probability but that the response probability may differ between groups. The response distribution is explicitly modeled. Under the RHG model, regression estimators for two-phase sampling can easily be adapted to the nonresponse situation. An estimator based on the RHG model has negligible nonresponse bias if the response model agrees with the true but unknown response distribution. However, the model does not have to be a perfect image of the real world in order to reduce nonresponse bias. The assumption of constant probability within well constructed groups will reduce nonresponse bias compared to more simple models. The RHG method is discussed in detail in Särndal, Swensson, and Wretman (1992).

In calibration for nonresponse no modeling of the response probabili-

ties is done. The properties of the calibration estimator will depend on the formulation of the auxiliary vector, the strength of association between the auxiliary vector and the study variable and of the response behaviour (Särndal and Lundström, 2005). When auxiliary information exists at both the sample level and the population level Estevao and Särndal (2002) show that, although in the context of two-phase sampling, different options exist for the use of auxiliary information originating from the two different levels. Särndal and Lundström (2005) consider three different calibration estimators adjusting for nonresponse using auxiliary information from both the sample level and the population level. They distinguish two types of procedures, single-step calibration and two-step calibration. For the two-step calibration procedure they consider two different alternatives, A and B, for the use of the auxiliary information. They conclude that although the three calibration estimators are not identical, one can expect the estimators to have only minor differences in regard to their capacity to provide effective protection against nonresponse bias. They also state that the two-step B alternative uses final weights that are less controlled, which may have some effect on the bias and the variance. Henceforth, we will use the single-step calibration procedure.

In section 4.1 we will present a RHG model suitable for the estimation of t_{yU_I}. In section 4.2 we will consider estimation of t_{yU_I} using calibration for nonresponse.

4.1 Regression estimator under the response homogeneity groups model

A simple and naive assumption about the response probabilities is that $\Pr(k \in r_F \,|s_F) = \theta_k = \theta$ for all k. A more useful response modeling would be to use *response homogeneity groups* (RHG). The use of response homogeneity groups allows more realistic response models in that it will give every element within a group the same response probability but the response probabilities are allowed to vary between the groups.

The general formulation of the response model, adapted to the situation with frame imperfections, is as follows: the realized sample s_F is partitioned into groups s_{Fh}, $h = 0, 1, \ldots, h, \ldots, H_{s_F}$, such that response probabilities are constant within groups, but are allowed to vary between groups. For our purpose we define the RHG-group with $h = 0$, i.e. s_{F0}, as the group that corresponds to the set s_{OC}, the sample elements that belong to the overcoverage

population. This can be done since we have information from the current register on which elements in s_F that belong to RHG-group $s_{F0} = s_{OC}$. Furthermore, for $h = 1, \ldots, H_{s_F}$, let $s_{Ih} = s_{Fh}$. In the following we will use the notation s_{Ih} to emphasize the fact that the RHG-group s_{OC} is excluded. For s_{Ih} $(h = 1, \ldots, H_{s_F})$ we have the response set r_{Ih}. The (random) number of elements in s_{Ih} and r_{Ih} are denoted $n_{s_{Ih}}$ and $n_{r_{Ih}}$ respectively. Given s_F we have

$$\begin{cases} \Pr\left(k \in r_I \,|s_I\right) = \pi_{k|s_I} = \theta_{hs_I} > 0 & k \in s_{Ih} \\ \Pr\left(k\&l \in r_I \,|s_I\right) = \pi_{kl|s_I} = \Pr\left(k \in r_I \,|s_I\right) \Pr\left(l \in r_I \,|s_I\right) & k \neq l \in s_I \end{cases} \tag{5}$$

for $h = 1, \ldots, H_{s_F}$.

Conditioning on the response count vector $\mathbf{n}_{r_I} = (n_{r_{I1}}, \ldots, n_{r_{Ih}}, \ldots, n_{r_{IH_{s_F}}})$ the conditional response probabilities (for $h = 1, \ldots, H_{s_F}$) are

$$\pi_{k\left|_{s_I, \mathbf{n}_{r_I}}\right.} = \Pr\left(k \in r_I \,|s_I, \mathbf{n}_{r_I}\right) = \pi_{kk\left|_{s_I, \mathbf{n}_{r_I}}\right.} = f_{Ih} = \frac{n_{r_{Ih}}}{n_{s_{Ih}}} \tag{6a}$$

for $k \in s_{Ih}$ and

$$\pi_{kl\left|_{s_I, \mathbf{n}_{r_I}}\right.} = \Pr\left(k\&l \in r_I \,|s_I, \mathbf{n}_{r_I}\right) = \begin{cases} \dfrac{f_{Ih}\left(n_{r_{Ih}} - 1\right)}{\left(n_{s_{Ih}} - 1\right)}; & k \neq l \in s_{Ih} \\ f_{Ih} f_{Ih'}; & k \in s_{Ih}, l \in s_{Ih'}; h \neq h' \end{cases} \tag{6b}$$

Under the RHG model given by (5) and using the auxiliary information defined by (1) and (3) respectively we obtain an estimator for t_{yU_I}, i.e. we obtain

$$\hat{t}_{yU_I cr} = \sum\nolimits_{U_I} \hat{y}_{1k} + \sum_{h=1}^{H_{s_F}} \left(\sum\nolimits_{s_{Ih}} \frac{\left(\hat{y}_k - \hat{y}_{1k}\right)}{\pi_{Fk}} + f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\left(y_k - \hat{y}_k\right)}{\pi_{Fk}} \right) \tag{7}$$

where

$$\begin{aligned} \hat{y}_{1k} &= \mathbf{x}'_{1k} \widehat{\mathbf{B}}_{1r_I}^{RHG} \\ &= \mathbf{x}'_{1k} \left(\sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{x}_{1k} \mathbf{x}'_{1k}}{\pi_{Fk} \sigma_{1k}^2}^{-1} \right)^{-1} \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{x}_{1k} y_k}{\pi_{Fk} \sigma_{1k}^2} \end{aligned} \tag{8}$$

for $k \in U_I$, and

$$\begin{aligned}\hat{y}_k &= \mathbf{x}_k' \widehat{\mathbf{B}}_{r_I}^{RHG} \\ &= \mathbf{x}_k' \left(\sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_{Fk}\sigma_k^2}^{-1} \right)^{-1} \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{x}_k y_k}{\pi_{Fk}\sigma_k^2} \end{aligned} \tag{9}$$

for $k \in s_I$. Here σ_{1k}^2 and σ_k^2 expresses the statistician's best opinion of the residual variability of y in a linear relationship with $\mathbf{x}_{1k}$ and $\mathbf{x}_k$, respectively. For details, see Särndal et al. (1992).

For large samples the approximate conditional expectation of $\hat{t}_{yU_I cr}$ given s_F, is, assuming that the RHG model in (5) holds,

$$\begin{aligned} E_{RD}\left(\hat{t}_{yU_I cr} \,|s_F\right) &= E_{\mathbf{n}_{r_F}} E_{RD}\left(\hat{t}_{yU_I cr} \,|s_F, \mathbf{n}_{r_F}\right) \\ &\approx \hat{t}_{yU_I reg} \end{aligned} \tag{10}$$

where

$$\hat{t}_{yU_I reg} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{1k} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{1k} \right)' \widehat{\mathbf{B}}_{1s_I} \tag{11}$$

with

$$\begin{aligned} \widehat{\mathbf{B}}_{1s_I} &= \widehat{\mathbf{T}}_{\mathbf{x}_1 s_I}^{-1} \widehat{\mathbf{t}}_{\mathbf{x}_1 y s_I} \\ &= \left(\sum\nolimits_{s_I} \frac{\mathbf{x}_{1k}\mathbf{x}_{1k}'}{\pi_{Fk}\sigma_{1k}^2} \right)^{-1} \sum\nolimits_{s_I} \frac{\mathbf{x}_{1k} y_k}{\pi_{Fk}\sigma_{1k}^2} \end{aligned} \tag{12}$$

The estimator $\hat{t}_{yU_I reg}$ is a domain regression estimator that could have been used in case of full response. Note that $\hat{t}_{yU_I reg}$ is approximately unbiased for t_{yU_I} and thus $\hat{t}_{yU_I cr}$ is approximatly unbiased. In section 6 results from a small Monte Carlo simulation (see table 6.9) indicate confirmation of these results.

Remark 1 *When the response model is a poor reflection of the true response behavior the regression estimator under the RHG model is biased, at worst severely so (Tångdahl, 2004). This scenario will not be considered here.*

For the variance and a variance estimator for $\hat{t}_{yU_I cr}$ we need the following residuals,

$$E_{1k} = y_k - \mathbf{x}_{1k}' \mathbf{B}_{1U_I} \tag{13}$$

for $k \in U_I$, with

$$\mathbf{B}_{1U_I} = \left(\sum\nolimits_{U_I} \frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\sigma^2_{1k}} \right)^{-1} \sum\nolimits_{U_I} \frac{\mathbf{x}_{1k} y_k}{\sigma^2_{1k}} \tag{14}$$

and

$$\hat{E}_k = y_k - \mathbf{x}'_k \widehat{\mathbf{B}}_{s_I} \tag{15}$$

for $k \in s_I$, with

$$\widehat{\mathbf{B}}_{s_I} = \left(\sum\nolimits_{s_I} \frac{\mathbf{x}_k \mathbf{x}'_k}{\sigma^2_k \pi_{Fk}} \right)^{-1} \sum\nolimits_{s_I} \frac{\mathbf{x}_k y_k}{\sigma^2_k \pi_{Fk}} \tag{16}$$

Moreover, let

$$e_{1k} = y_k - \hat{y}_{1k} \tag{17}$$

and

$$e_k = y_k - \hat{y}_k \tag{18}$$

Using the principles in Särndal et al. (1992) the approximate variance is

$$AV\left(\hat{t}_{yU_I cr}\right) = AV_{SAM}\left(\hat{t}_{yU_I cr}\right) + AV_{NR}\left(\hat{t}_{yU_I cr}\right)$$

with

$$AV_{SAM}\left(\hat{t}_{yU_I cr}\right) = \sum\sum\nolimits_{U_I} \left(\pi_{Fkl} - \pi_{Fk}\pi_{Fl}\right) \check{E}_{1k} \check{E}_{1l} \tag{19}$$

and

$$AV_{NR}\left(\hat{t}_{yU_I cr}\right) = E_p E_{\mathbf{n}_{r_F}} \left(\sum_{h=1}^{H_{s_F}} n^2_{s_{Ih}} \frac{1 - f_h}{n_{r_{Ih}}} S^2_{\check{\check{E}} s_{Ih}} | s_I \right) \tag{20}$$

where $S^2_{\check{\check{E}} s_{Ih}}$ is the variance in the set s_{Ih} of $\check{\check{E}}_k = \hat{E}_k / \pi_{Fk}$ with $\hat{E}_k$ given by (15) . The first component in $AV\left(\hat{t}_{yU_I cr}\right)$ reflects the sampling variance. The second component represents the increase in variance caused by nonresponse, i.e. the nonresponse variance. Note that, for this particular estimator, the AV_{NR} component is equal to zero for full response.

A variance estimator is given by

$$\hat{V}\left(\hat{t}_{yU_I cr}\right) = \hat{V}_{SAM}\left(\hat{t}_{yU_I cr}\right) + \hat{V}_{NR}\left(\hat{t}_{yU_I cr}\right)$$

with

$$\hat{V}_{SAM}\left(\hat{t}_{yU_I cr}\right) = \sum\sum\nolimits_{r_I} \frac{\pi_{Fkl} - \pi_{Fk}\pi_{Fl}}{\pi_{Fkl}\pi_{kl}\big|_{s_F,\mathbf{n}_{r_F}}} \check{e}_{1k}\check{e}_{1l} \tag{21}$$

and

$$\hat{V}_{NR}\left(\hat{t}_{yU_I cr}\right) = \sum_{h=1}^{H_{s_F}} n_{s_{Ih}}^2 \frac{1 - f_{Ih}}{n_{r_{Ih}}} S_{\check{e}r_{Ih}}^2 \tag{22}$$

where $S_{\check{e}r_{Ih}}^2$ is the variance in the set r_{Ih} of $\check{e}_k = e_k/\pi_{Fk}$ and where e_k is given by (18) . This variance estimator is obtained by replacing E_{1k} and $\hat{E}_k$ in (19) and (20) by e_{1k} and e_k, respectively.

4.2 Calibration for nonresponse

When auxiliary information exists at both the sample level and the population level the calibration technique, when used under nonresponse, seeks a weight system for $k \in r$, where r is the response set, that satisfies the calibration equation $\sum_r v_k \check{\mathbf{x}}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_s \check{\mathbf{x}}_{2k} \end{pmatrix}$. In Särndal and Lundström (2005) a single-step nonresponse calibration estimator for domains is suggested. Using this estimator and the definition of y_k from (4) and the auxiliary information from (1) and (3), we obtain the estimator

$$\hat{t}_{yU_I cal} = \sum\nolimits_{r_I} v_{Ik}\check{y}_k \tag{23}$$

where

$$v_{Ik} = 1 + \left(\ddot{\mathbf{t}}_{\mathbf{x}U_I} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{24}$$

for every $k \in r_I$, with $\ddot{\mathbf{t}}_{\mathbf{x}U_I} = \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix}$.

Remark 2 *An alternative to* v_{Ik} *would be*

$$v_{Ikz} = 1 + \left(\ddot{\mathbf{t}}_{\mathbf{x}U_I} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_r \mathbf{z}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{z}_k$$

Here $\boldsymbol{z}_k$ *is an instrument vector that can be any vector value specified for* $k \in s_I$ *of the same dimension as* $\boldsymbol{x}_k$*. The vector* $\boldsymbol{z}_k$ *can be a specified function of* $\boldsymbol{x}_k$*, e.g.* $\boldsymbol{z}_k = c_k \mathbf{x}_k$*. Henceforth we will assume the standard specification,* $\mathbf{z}_k = \mathbf{x}_k$*.*

The bias of $\hat{t}_{yU_I cal}$ with respect to t_{yU_I}, for large response sets, is a function of y_k, $\mathbf{x}_k$ and θ_k of all $k \in U_I$ and it is given approximately by

$$B\left(\hat{t}_{yU_I cal}\right) \approx -\sum\nolimits_{U_I} \left(1 - \theta_k\right) E_{\theta k} \tag{25}$$

where $E_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_{U_I \theta}$ with

$$\mathbf{B}_{U_I \theta} = \left(\sum\nolimits_{U_I} \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum\nolimits_{U_I} \theta_k \mathbf{x}_k y_k$$

The proof is given in appendix A.2. If there exists a vector $\boldsymbol{\lambda}$ such that $\dfrac{1}{\theta_k} = 1 + \boldsymbol{\lambda}' \mathbf{x}_k$ the approximate bias in (25) is equal to zero.

Using the principles in Särndal and Lundström a variance estimator of $\hat{t}_{yU_I cal}$ is

$$\hat{V}\left(\hat{t}_{yU_I cal}\right) = \hat{V}_{SAM}\left(\hat{t}_{yU_I cal}\right) + \hat{V}_{NR}\left(\hat{t}_{yU_I cal}\right) \tag{26}$$

where

$$\begin{aligned} \hat{V}_{SAM}\left(\hat{t}_{yU_I cal}\right) = & \sum\sum\nolimits_{r_I} \left(\frac{1}{\pi_{Fk}\pi_{Fl}} - \frac{1}{\pi_{Fkl}}\right) \left(v_{Ik} e_{1v_I k}\right) \left(v_{Il} e_{1v_I l}\right) \\ & - \sum\nolimits_{r_I} \frac{1}{\pi_{Fk}} \left(\frac{1}{\pi_{Fk}} - 1\right) v_{Ik} \left(v_{Ik} - 1\right) \left(e_{1v_I k}\right)^2 \end{aligned} \tag{27}$$

and

$$\hat{V}_{NR}\left(\hat{t}_{yU_I cal}\right) = \sum\nolimits_{r_I} v_{Ik} \left(v_{Ik} - 1\right) \left(\check{e}_{v_I k}\right)^2 \tag{28}$$

with v_{Ik} given by (24) . The residuals are given by

$$e_{1v_I k} = y_k - \mathbf{x}_{1k}' \widehat{\mathbf{B}}_{v_I r_I}^{(1)}$$

and

$$e_{v_I k} = y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_{v_I r_I}$$

where

$$\widehat{\mathbf{B}}_{v_I r_I} = \begin{pmatrix} \widehat{\mathbf{B}}^{(1)}_{v_I r_I} \\ \widehat{\mathbf{B}}^{(2)}_{v_I r_I} \end{pmatrix} = \left(\sum\nolimits_{r_I} v_{Ik} \mathbf{x}_k \check{\mathbf{x}}_k' \right)^{-1} \sum\nolimits_{r_I} v_{Ik} \mathbf{x}_k \check{y}_k$$

with v_{Ik} given by (24) .

Another suggestion would be to use

$$v_{Ik} = v_{s_I k} = 1 + \left(\sum\nolimits_{s_I} \check{\mathbf{x}}_k - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k' \right)^{-1} \mathbf{x}_k \qquad (29)$$

in the variance estimator.

5 A perfect current register exists - estimation of t_{yU}

In this section we will present two approaches in estimation of t_{yU}:

(i) estimation of t_{yU} in two steps and

(ii) direct estimation of t_{yU}.

The first step in *approach (i)* is to find an (at least approximately) unbiased estimator of t_{yU_I}. In step two we must find a guesstimate of $t_{yU_{UC}}$. One way of reasoning is as follows.

Let $\mathbf{u}_k$, a column vector with P components, denote an auxiliary vector value known in the current register. The vector is denoted $\mathbf{u}$ in order to distinguish it from the auxiliary vector $\mathbf{x}$ appearing in the estimator for t_{yU_I}, since the two vectors are not necessarily identical. Assume that a strong linear relationship between y and $\mathbf{u}$ in U_{UC} exists, such that $y_k \approx \mathbf{u}_k' \mathbf{B}_{\mathbf{u}U_{UC}}$, with

$$\mathbf{B}_{\mathbf{u}U_{UC}} = \left(\sum\nolimits_{U_{UC}} \frac{\mathbf{u}_k \mathbf{u}_k'}{\sigma^2_{uk}} \right)^{-1} \sum\nolimits_{U_{UC}} \frac{\mathbf{u}_k y_k}{\sigma^2_{uk}} \qquad (30)$$

where σ^2_{uk} is a suitably chosen constant, e.g. capturing an assumed heteroscedasticity in the linear relationship between y and $\mathbf{u}$. If $\mathbf{B}_{\mathbf{u}U_{UC}}$ was

known we would have an approximation of the total in the undercoverage set, $t_{yU_{UC}}$ using

$$t_{yU_{UC}}^{app} = \sum\nolimits_{U_{UC}} \mathbf{u}_k' \mathbf{B}_{\mathbf{u}U_{UC}} \tag{31}$$

However, $\mathbf{B}_{\mathbf{u}U_{UC}}$ is unknown. And, since no observations on y exist in U_{UC}, it is not possible to find a proper estimator for $\mathbf{B}_{\mathbf{u}U_{UC}}$. However, if $\mathbf{B}_{\mathbf{u}U_{UC}} \approx \mathbf{B}_{\mathbf{u}U_I}$, i.e. the linear relationship between y and $\mathbf{u}$ in U_{UC} is fairly well approximated by the linear relationship between y and $\mathbf{u}$ in U_I, we could use

$$\widetilde{t_{yU_{UC}}^{app}} = \sum\nolimits_{U_{UC}} \mathbf{u}_k' \mathbf{B}_{\mathbf{u}U_I}$$

Now, $\mathbf{B}_{\mathbf{u}U_I}$ may be estimated from the sample and we obtain a guesstimate of $t_{yU_{UC}}$,

$$\hat{\widetilde{t}}_{yU_{UC}}^{app} = \sum\nolimits_{U_{UC}} \mathbf{u}_k' \widehat{\mathbf{B}}_{\mathbf{u}r_I}$$

where $\widehat{\mathbf{B}}_{\mathbf{u}r_I}$ is an estimator for $\mathbf{B}_{\mathbf{u}U_I}$ based on the set r_I.

Remark 3 *If we have reason to believe that the linear relationship between y and $\boldsymbol{u}$ in the undercoverage set is better described by using a subgroup $U_{Ig} \subseteq U_I$, we could use $\widehat{\mathbf{B}}_{\mathbf{u}r_{Ig}}$ to estimate this relationship.*

Example 5.1 Consider a survey of enterprises. If, by comparing the frame with the current register, it is noticed that the undercoverage set mainly consists of small enterprises, we would choose the group g in U_I to represent these enterprises. Attention should be paid to the fact that this group must contain enough observations to avoid an unstable estimate $\widehat{\mathbf{B}}_{\mathbf{u}r_{Ig}}$.

Now, any guesstimate of $t_{yU_{UC}}$ may be used in combination with any approximately unbiased estimator of t_{yU_I}. Approach (i) is considered in section 5.1 where we consider the situation when both t_{yU_I} and $\mathbf{B}_{\mathbf{u}U_I}$ (in $\widetilde{t_{yU_{UC}}^{app}}$) are estimated using the RHG approach.

Note that, when using approach (i) it is possible to separate the effect of adjustment of nonresponse from the effect of adjustment of frame error since the two steps in this approach applies to these two adjustments.

In *approach (ii)* direct estimation of t_{yU} is considered. In section 5.2 we will use calibration as a method of direct estimation. When the calibration approach is used to adjust for nonresponse and frame imperfections it is not immediately possible to separate the effect of the nonresponse bias adjustment from the effect of the adjustment of frame error bias, since the adjustment for these two errors are done in one single step.

5.1 Regression estimator under the response homogeneity groups model

As discussed in the previous section, a guesstimate of $t_{yU_{UC}}$ could be obtained by

$$\hat{t}_{yU_{UC}}^{\widetilde{app}} = \sum\nolimits_{U_{UC}} \mathbf{u}_k' \widehat{\mathbf{B}}_{\mathbf{u}r_I}$$

where $\widehat{\mathbf{B}}_{\mathbf{u}r_I}$ is an estimator for $\mathbf{B}_{\mathbf{u}U_I}$ based on the set r_I.

One approximately unbiased estimator of $\mathbf{B}_{\mathbf{u}U_I}$ is given by

$$\begin{aligned}
\widehat{\mathbf{B}}_{\mathbf{u}r_I}^{RHG} &= \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}} \right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{u}yr_{Ih}} \\
&= \left(\sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{u}_k \mathbf{u}_k'}{\sigma_{uk}^2 \pi_{Fk}} \right)^{-1} \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{u}_k y_k}{\sigma_{uk}^2 \pi_{Fk}} \qquad (32)
\end{aligned}$$

where σ_{uk}^2 is a suitably chosen constant, e.g. capturing an assumed heteroscedasticity in the linear relationship between y and $\mathbf{u}$. Using this estimator, a guesstimate of $t_{yU_{UC}}$ would be

$$\hat{t}_{yU_{UC}}^{\widetilde{app}} = \sum\nolimits_{U_{UC}} \mathbf{u}_k' \widehat{\mathbf{B}}_{\mathbf{u}r_I}^{RHG} \qquad (33)$$

The (approximate) bias of $\hat{t}_{yU_{UC}}^{\widetilde{app}}$ with respect to $t_{yU_{UC}}$ is given by

$$\begin{aligned}
B\left(\hat{t}_{yU_{UC}}^{\widetilde{app}}\right) &= t_{yU_{UC}}^{\widetilde{app}} - t_{yU_{UC}} \\
&= \sum\nolimits_{U_{UC}} \mathbf{u}_k' \mathbf{B}_{\mathbf{u}U_I} - t_{yU_{UC}} \qquad (34)
\end{aligned}$$

i.e. the approximation error of $t_{yU_{UC}}^{\widetilde{app}}$.

Using $\hat{t}_{yU_I cr}$ from (7) and $\hat{t}_{yU_{UC}}^{\widetilde{app}}$ from (33) to build an estimator for t_{yU} we obtain

$$\begin{aligned}
\hat{t}_{yU_{cr}}^{\widetilde{app}} &= \hat{t}_{yU_I cr} + \hat{t}_{yU_{UC}}^{\widetilde{app}} \\
&= \sum\nolimits_{U_I} \hat{y}_{1k} + \sum_{h=1}^{H_{s_F}} \left(\sum\nolimits_{s_{Ih}} \frac{(\hat{y}_k - \hat{y}_{1k})}{\pi_{Fk}} + f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{(y_k - \hat{y}_k)}{\pi_{Fk}} \right) \\
&\quad + \sum\nolimits_{U_{UC}} \mathbf{u}_k' \widehat{\mathbf{B}}_{\mathbf{u}r_I}^{RHG} \qquad (35)
\end{aligned}$$

It follows that, for large samples, the approximate expectation of $\hat{t}_{yUcr}^{\widetilde{app}}$ is given by

$$\begin{aligned} E_p\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) &= E_p\left(E_{RD}\left(\hat{t}_{yUcr}^{\widetilde{app}}\,|s_F\right)\right) \\ &= E_p\left(E_{\mathbf{n}_{r_F}} E_{RD}\left(\hat{t}_{yUcr}^{\widetilde{app}}\,|s_F, \mathbf{n}_{r_F}\right)\right) \\ &\approx E_p\left(\hat{t}_{yUreg}^{*}\right) \\ &\approx t_{yU_I} + t_{yU_{UC}}^{\widetilde{app}} \\ &= t_{yU} + \left(t_{yU_{UC}}^{\widetilde{app}} - t_{yU_{UC}}\right) \end{aligned} \tag{36}$$

where

$$\hat{t}_{yUreg}^{*} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{1k} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{1k}\right)' \widehat{\mathbf{B}}_{1s_I} + \sum\nolimits_{U_{UC}} \mathbf{u}_k' \widehat{\mathbf{B}}_{\mathbf{u}s_I} \tag{37}$$

with $\widehat{\mathbf{B}}_{1s_I}$ defined by (12) and

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{u}s_I} &= \widehat{\mathbf{T}}_{\mathbf{u}s_I}^{-1} \widehat{\mathbf{t}}_{\mathbf{u}ys_I} \\ &= \left(\sum\nolimits_{s_I} \frac{\mathbf{u}_k \mathbf{u}_k'}{\sigma_{uk}^2 \pi_{Fk}}\right)^{-1} \sum\nolimits_{s_I} \frac{\mathbf{u}_k y_k}{\sigma_{uk}^2 \pi_{Fk}} \end{aligned} \tag{38}$$

The estimator $\hat{t}_{yUreg}^{*}$ is an estimator that could have been used in case of frame error and full response (Ängsved, 2004).

From Ängsved (2004) we have, with some minor modifications, an approximate sampling variance

$$AV_{SAM}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) = \sum\sum\nolimits_{U_I} \left(\pi_{Fkl} - \pi_{Fk}\pi_{Fl}\right) \check{F}_k \check{F}_l \tag{39}$$

where

$$F_k = E_{1k} + \frac{\sum_{U_{UC}} \mathbf{u}_k' \mathbf{T}_{\mathbf{u}U_I}^{-1} \mathbf{u}_k E_{uk}}{\sigma_{uk}^2}$$

with E_{1k} given by (13) and $E_{uk} = y_k - \mathbf{u}_k' \mathbf{B}_{\mathbf{u}U_I}$ with

$$\mathbf{B}_{\mathbf{u}U_I} = \mathbf{T}_{\mathbf{u}U_I}^{-1} \mathbf{t}_{\mathbf{u}yU_I} = \left(\sum\nolimits_{U_I} \frac{\mathbf{u}_k \mathbf{u}_k'}{\sigma_{uk}^2}\right)^{-1} \sum\nolimits_{U_I} \frac{\mathbf{u}_k y_k}{\sigma_{uk}^2}$$

Using the Taylor linearization technique we obtain an expression for the approximate nonresponse variance

$$AV_{NR}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) = E_p\left[\sum_{h=1}^{H_{s_F}} n_{s_{Ih}}^2 \frac{1-f_{Ih}}{n_{r_{Ih}}} S_{\check{F}_{xus_{Ih}}}^2\right] \tag{40}$$

where $S_{\check{F}_{xus_{Ih}}}^2$ is the variance in the set s_{Ih} of F_{xuk}/π_{Fk} with

$$F_{xuk} = \frac{\left(\sum_{U_I}\mathbf{x}_{1k} - \sum_{s_I}\check{\mathbf{x}}_{1k}\right)'\widehat{\mathbf{T}}_{\mathbf{x}_1 s_I}^{-1}\mathbf{x}_{1k}\hat{E}_{1k}}{\sigma_{1k}^2} + \hat{E}_k + \frac{\sum_{U_{UC}}\mathbf{u}_k'\widehat{\mathbf{T}}_{\mathbf{u}s_I}^{-1}\mathbf{u}_k\hat{E}_{uk}}{\sigma_{uk}^2}$$

with $\hat{E}_{1k} = y_k - \mathbf{x}_{1k}'\widehat{\mathbf{B}}_{1s_I}$, $\hat{E}_k$ given by (15) and $\hat{E}_{uk} = y_k - \mathbf{u}_k'\widehat{\mathbf{B}}_{\mathbf{u}s_I}$ respectively. (For details on the Taylor linearization, see appendix A.1)

A variance estimator of $\hat{t}_{yUcr}^{\widetilde{app}}$ is given by

$$\hat{V}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) = \hat{V}_{SAM}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) + \hat{V}_{NR}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right)$$

An estimated sampling variance component would be

$$\hat{V}_{SAM}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right) = \sum\sum_{r_I}\frac{\pi_{Fkl}-\pi_{Fk}\pi_{Fl}}{\pi_{Fkl}\pi_{kl}\big|_{s_F,\mathbf{n}_{r_F}}}\check{f}_k\check{f}_l \tag{41}$$

where

$$f_k = e_{1k} + \frac{\sum_{U_{UC}}\mathbf{u}_k'\widehat{\mathbf{T}}_{\mathbf{u}s_I}^{-1}\mathbf{u}_k e_{uk}}{\sigma_{uk}^2}$$

with e_{1k} from (17) and

$$e_{uk} = y_k - \mathbf{u}_k'\widehat{\mathbf{B}}_{\mathbf{u}r_I}^{RHG} \tag{42}$$

Remark 4 *An alternative to f_k would be*

$$f_{k,alt} = e_{1k} + \frac{\sum_{U_{UC}}\mathbf{u}_k'\left(\sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}\right)^{-1}\mathbf{u}_k e_{uk}}{\sigma_{uk}^2}$$

with $\left(\sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}\right)^{-1}$ defined in (32) *At present it is not clear which choice is the better and some further work is needed in deciding whether to use f_k or $f_{k,alt}$ in this variance estimator.*

A variance estimator for the nonresponse variance is given by

$$\hat{V}_{NR}\left(\hat{t}_{yUcr}^{\widetilde{app}}\right)=\sum_{h=1}^{H_{s_F}} n_{s_{Ih}}^2 \frac{1-f_{Ih}}{n_{r_{Ih}}} S_{\check{f}_{xur_{Ih}}}^2 \tag{43}$$

where $S_{\check{f}_{xur_{Ih}}}^2$ is the variance in the set r_{Ih} of f_{xuk}/π_{Fk} where

$$f_{xuk}=\frac{\left(\sum_{U_I}\mathbf{x}_{1k}-\sum_{s_I}\check{\mathbf{x}}_{1k}\right)'\widehat{\mathbf{T}}_{\mathbf{x}_1 s_I}^{-1}\mathbf{x}_{1k}e_{1k}}{\sigma_{1k}^2}+e_k+\frac{\sum_{U_{UC}}\mathbf{u}_k'\widehat{\mathbf{T}}_{\mathbf{u}s_I}^{-1}\mathbf{u}_k e_{uk}}{\sigma_{uk}^2}$$

with e_{1k}, e_k and e_{uk} given by (17), (18) and (42) respectively.

Remark 5 *An alternative to f_{xuk} would be*

$$\begin{aligned} f_{xuk,alt} &= \frac{\left(\sum_{U_I}\mathbf{x}_{1k}-\sum_{h=1}^{H_{s_F}}\hat{\mathbf{t}}_{x_1 r_{Ih}}\right)'\left(\sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}}\right)^{-1}\mathbf{x}_{1k}e_{1k}}{\sigma_{1k}^2} \\ &\quad +e_k+\frac{\sum_{U_{UC}}\mathbf{u}_k'\left(\sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}\right)^{-1}\mathbf{u}_k e_{uk}}{\sigma_{uk}^2} \end{aligned}$$

where

$$\sum_{h=1}^{H_{s_F}}\hat{\mathbf{t}}_{x_1 r_{Ih}}=\sum_{h=1}^{H_{s_F}} f_{Ih}^{-1}\sum\nolimits_{r_{Ih}}\check{\mathbf{x}}_{1k} \quad \textit{and} \quad \sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}}=\sum_{h=1}^{H_{s_F}} f_{Ih}^{-1}\sum\nolimits_{r_{Ih}}\frac{\mathbf{x}_{1k}\mathbf{x}_{1k}'}{\pi_{Fk}\sigma_{1k}^2}$$

($\sum_{h=1}^{H_{s_F}}\widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}$ follows analogously). At present it is not clear which choice is the better and some further work is needed in deciding whether to use f_{xuk} or $f_{xuk,alt}$ in this variance estimator.

A special case is when $\mathbf{u}_k=\mathbf{x}_{1k}$, the updated auxiliary information from the current register. Using $\mathbf{u}_k=\mathbf{x}_{1k}$ and $\sigma_{uk}^2=\sigma_{1k}^2$ estimator $\hat{t}_{yUcr}^{\widetilde{app}}$ simplifies and we obtain the special case

$$\hat{t}_{yUcr}^{\widetilde{app}}=\sum\nolimits_U\hat{y}_{1k}+\sum_{h=1}^{H_{s_F}}\left(\sum\nolimits_{s_{Ih}}\frac{(\hat{y}_k-\hat{y}_{1k})}{\pi_{Fk}}+f_{Ih}^{-1}\sum\nolimits_{r_{Ih}}\frac{(y_k-\hat{y}_k)}{\pi_{Fk}}\right) \tag{44}$$

5.2 Calibration for nonresponse and frame imperfections

In Särndal and Lundström (2005) the following estimator is proposed as an estimator for t_{yU} in case of both frame imperfections and nonresponse,

$$\hat{t}_{yUcal,1} = \sum\nolimits_{r_I} \tilde{v}_{1k}\check{y}_k \tag{45}$$

with $\tilde{v}_{1k} = 1 + \left(\tilde{\mathbf{t}}_{x_1U} - \sum_{r_I}\check{\mathbf{x}}_{1k}\right)' \left(\sum_{r_I}\mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1}\mathbf{x}_{1k}$ where $\tilde{\mathbf{t}}_{x_1U}$ should be a close approximation of $\sum_U \mathbf{x}_{1k}$. In the situation when a perfect current register exists, $\sum_U \mathbf{x}_{1k}$ is known and we obtain

$$\hat{t}_{yUcal,1} = \sum\nolimits_{r_I} v_{1k}\check{y}_k \tag{46}$$

with $v_{1k} = 1 + \left(\sum_U \mathbf{x}_{1k} - \sum_{r_I}\check{\mathbf{x}}_{1k}\right)' \left(\sum_{r_I}\mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1}\mathbf{x}_{1k}$.

The difference $\hat{t}_{yUcal,1} - \hat{t}_{yU_Ical,1}$, the latter from (23) when using $\mathbf{x}_k = \mathbf{x}_{1k}$, indicates the additional contribution of $\hat{t}_{yUcal,1}$ compared to $\hat{t}_{yU_Ical,1}$. We have

$$\begin{aligned}\hat{t}_{yUcal,1} - \hat{t}_{yU_Ical,1} &= \left(\sum\nolimits_U \mathbf{x}_{1k} - \sum\nolimits_{U_I}\mathbf{x}_{1k}\right)'\widehat{\mathbf{B}}_{1r_I} \\ &= \sum\nolimits_{U_{UC}} \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1r_I}\end{aligned} \tag{47}$$

where $\widehat{\mathbf{B}}_{1r_I} = \left(\sum_{r_I}\mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1}\sum_{r_I}\mathbf{x}_{1k}\check{y}_k$.

Using the principles in Särndal and Lundström we propose the following variance estimator of $\hat{t}_{yUcal,1}$

$$\hat{V}\left(\hat{t}_{yUcal,1}\right) = \hat{V}_{SAM}\left(\hat{t}_{yUcal,1}\right) + \hat{V}_{NR}\left(\hat{t}_{yUcal,1}\right) \tag{48}$$

where

$$\begin{aligned}\hat{V}_{SAM}\left(\hat{t}_{yUcal,1}\right) = &\sum\sum\nolimits_{r_I}\left(\frac{1}{\pi_{Fk}\pi_{Fl}} - \frac{1}{\pi_{Fkl}}\right)\left(\varphi_k e_{\varphi k}\right)\left(\varphi_l e_{\varphi l}\right) \\ &- \sum\nolimits_{r_I}\frac{1}{\pi_{Fk}}\left(\frac{1}{\pi_{Fk}} - 1\right)\varphi_k\left(\varphi_k - 1\right)\left(e_{\varphi k}\right)^2\end{aligned} \tag{49}$$

and

$$\hat{V}_{NR}\left(\hat{t}_{yUcal,1}\right) = \sum\nolimits_{r_I}\varphi_k\left(\varphi_k - 1\right)\left(\check{e}_{\varphi k}\right)^2 \tag{50}$$

where

$$e_{\varphi k} = y_k - \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1\varphi r_I} \tag{51}$$

with

$$\widehat{\mathbf{B}}_{1\varphi r_I} = \left(\sum\nolimits_{r_I} \varphi_k \mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1} \sum\nolimits_{r_I} \varphi_k \mathbf{x}_{1k}\check{y}_k \tag{52}$$

Still following Särndal and Lundström, and using the fact that we can identify the set o_I from o_F we propose the use of

$$\varphi_k = \varphi_{s_I k} = 1 + \left(\sum\nolimits_{s_I} \check{\mathbf{x}}_{1k} - \sum\nolimits_{r_I} \check{\mathbf{x}}_{1k}\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1} \mathbf{x}_{1k} \tag{53}$$

in the variance estimator.

However, the approach above does not take full advantage of all possible information since only one level of auxiliary information is involved in the estimator. Using the auxiliary information from both the sample level and the population level we obtain

$$\hat{t}_{yUcal} = \sum\nolimits_{r_I} v_k \check{y}_k \tag{54}$$

with

$$v_k = 1 + \left[\begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right]' \left(\sum\nolimits_{r_I} \mathbf{x}_k\check{\mathbf{x}}'_k\right)^{-1} \mathbf{x}_k \tag{55}$$

Looking at the difference $\hat{t}_{yUcal} - \hat{t}_{yU_I cal}$ (with the latter from (23)) we notice that $\hat{t}_{yUcal}$ adds the term

$$\begin{aligned} & \left[\begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix}\right]' \widehat{\mathbf{B}}_{r_I} \\ = \; & \left(\sum\nolimits_U \mathbf{x}_{1k} - \sum\nolimits_{U_I} \mathbf{x}_{1k}\right)' \widehat{\mathbf{B}}^{(1)}_{r_I} \end{aligned} \tag{56}$$

where

$$\widehat{\mathbf{B}}_{r_I} = \begin{bmatrix} \widehat{\mathbf{B}}^{(1)}_{r_I} \\ \widehat{\mathbf{B}}^{(2)}_{r_I} \end{bmatrix} = \left(\sum\nolimits_{r_I} \mathbf{x}_k\check{\mathbf{x}}'_k\right)^{-1} \sum\nolimits_{r_I} \mathbf{x}_k\check{y}_k \tag{57}$$

to $\hat{t}_{yU_I cal}$.

For large response sets, the approximate bias of $\hat{t}_{yUcal}$ is

$$B\left(\hat{t}_{yUcal}\right) \approx -\sum\nolimits_{U_I}\left(1-\theta_k\right)E_{\theta k}+\sum\nolimits_{U_{UC}}\mathbf{x}'_{1k}\mathbf{B}^{(1)}_{\boldsymbol{\theta} U_I}-t_{yU_{UC}} \tag{58}$$

where θ_k are the unknown response probabilities and $E_{\theta k}=y_k-\mathbf{x}'_k\mathbf{B}_{\boldsymbol{\theta} U_I}$ with

$$\mathbf{B}_{\boldsymbol{\theta} U_I}=\left[\begin{array}{c}\mathbf{B}^{(1)}_{\boldsymbol{\theta} U_I}\\ \mathbf{B}^{(2)}_{\boldsymbol{\theta} U_I}\end{array}\right]=\left(\sum\nolimits_{U_I}\theta_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\sum\nolimits_{U_I}\theta_k\mathbf{x}_k y_k$$

The derivation is given in appendix A.3. The approximate bias is a function of both the unknown response probabilities and the unknown elements in the undercoverage.

Again using the principles in Särndal and Lundström we propose the following variance estimator of $\hat{t}_{yUcal}$

$$\hat{V}\left(\hat{t}_{yUcal}\right)=\hat{V}_{SAM}\left(\hat{t}_{yUcal}\right)+\hat{V}_{NR}\left(\hat{t}_{yUcal}\right) \tag{59}$$

where

$$\begin{aligned}\hat{V}_{SAM}\left(\hat{t}_{yUcal}\right)=&\sum\sum\nolimits_{r_I}\left(\frac{1}{\pi_{Fk}\pi_{Fl}}-\frac{1}{\pi_{Fkl}}\right)\left(v_k e_{1vk}\right)\left(v_l e_{1vl}\right)\\ &-\sum\nolimits_{r_I}\frac{1}{\pi_{Fk}}\left(\frac{1}{\pi_{Fk}}-1\right)v_k\left(v_k-1\right)\left(e_{1vk}\right)^2\end{aligned} \tag{60}$$

and

$$\hat{V}_{NR}\left(\hat{t}_{yUcal}\right)=\sum\nolimits_{r_I}v_k\left(v_k-1\right)\left(\check{e}_{vk}\right)^2 \tag{61}$$

where

$$e_{1vk}=y_k-\mathbf{x}'_{1k}\widehat{\mathbf{B}}^{(1)}_{vr_I} \tag{62}$$

$$e_{vk}=y_k-\mathbf{x}'_k\widehat{\mathbf{B}}_{vr_I} \tag{63}$$

with

$$\widehat{\mathbf{B}}_{vr_I}=\left[\begin{array}{c}\widehat{\mathbf{B}}^{(1)}_{vr_I}\\ \widehat{\mathbf{B}}^{(2)}_{vr_I}\end{array}\right]=\left(\sum\nolimits_{r_I}v_k\mathbf{x}_k\check{\mathbf{x}}'_k\right)^{-1}\sum\nolimits_{r_I}v_k\mathbf{x}_k\check{y}_k \tag{64}$$

with v_k given by (55).

Remark 6 *When auxiliary information exists at the population level only two alternative variance estimators may be used,* $\hat{V}\left(\hat{t}_{yUcal,1}\right)$ *from* (48) *using*

$$\varphi_k = \varphi_{s_I k} = 1 + \left(\sum\nolimits_{s_I} \check{\mathbf{x}}_{1k} - \sum\nolimits_{r_I} \check{\mathbf{x}}_{1k}\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1} \mathbf{x}_{1k} \tag{65}$$

or $\hat{V}\left(\hat{t}_{yUcal}\right)$ *from* (59) *using*

$$v_k = v_{Uk} = 1 + \left(\sum\nolimits_{U} \mathbf{x}_{1k} - \sum\nolimits_{r_I} \check{\mathbf{x}}_{1k}\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_{1k}\check{\mathbf{x}}'_{1k}\right)^{-1} \mathbf{x}_{1k} \tag{66}$$

6 A small simulation study

A limited Monte Carlo study was performed in order to get some empirical insight into the behaviour of the suggested point estimators for t_{yU} as well as of their corresponding variance estimators.

The starting point was to create the population sets U_F and U, and their subsets U_{UC}, U_{OC} and U_I. The objective was that these sets to some extent should resemble an actual real life population. In order to do this we used information from an illustration on changes in the Business Register at Statistics Sweden (Ängsved, 2004).

6.1 Population sets and study variable

An artificial population, denoted *MU20000*, of 20 000 elements was created from the population *MU281* according to Axelson (2000). The population *MU281* consists of the 281 smallest municipalities in Sweden in 1982, see appendix B in Särndal, Swensson, and Wretman (1992), and *MU20000* was created in order to mimic *MU281*. From *MU20000* the different population sets, i.e. the target population, the overcoverage population and the undercoverage population, to be used in the Monte Carlo study, were created. Using simple random sampling, *MU20000* was divided into two subsets. The first part, denoted U_F (of size $N_F = 10\ 000$) was used as frame population in the simulation study and from this set we created the overcoverage and the intersection sets. From the second part of *MU20000*, denoted $U_{"newborn"}$ the undercoverage set was created.

In order to create the overcoverage set the variable corresponding to *P75* in U_F was used to form five size groups. The overcoverage was determined

by drawing a Bernoulli sample from each group with inclusion probabilities, P ("*death*"), given in table 6.1.

Table 6.1 *P("death") in five size groups (the groups created from P75)*

P75	0-30	30-50	50-70	70-100	100-300
P("*death*")	0.07835	0.04628	0.03943	0.03575	0.04702

Using these Bernoulli samples we created the overcoverage set U_{OC} of size $N_{OC} = 666$ and the intersection set U_I of size $N_I = 9\ 334$. Table 6.2 illustrates the frequency distribution by size group in the intersection set and the overcoverage set respectively.

Table 6.2 *Population by size group (the groups created from P75) in the intersection set and the overcoverage set*

P75	0-30	30-50	50-70	70-100	100-300	Total
U_I (%)	6 632 (71.1)	1 586 (17.0)	629 (6.7)	341 (3.7)	146 (1.6)	9 334 (100)
U_{OC} (%)	473 (71.0)	111 (16.7)	49 (4.0)	26 (4.0)	7 (1.0)	666 (100)

The next step was to create the undercoverage set. This was done from the set $U_{"newborn"}$ which was divided into five size groups using the variable corresponding to *P85*. From each group a Bernoulli sample was drawn with inclusion probabilities, P ("*born*") , given in table 6.3.

Table 6.3 *P("born") in five size groups (the groups created from P85)*

P85	0-30	30-50	50-70	70-100	100-300
P ("*born*")	0.10174	0.04723	0.02954	0.01773	0.02765

This gave us the undercoverage set U_{UC} of size $N_{UC} = 850$. Table 6.4 illustrates the frequency distribution by size group in the undercoverage set.

Table 6.4 *Population by size group (the groups created from P85) in the intersection set and the undercoverage set*

P85	0-30	30-50	50-70	70-100	100-300	Total
U_I (%)	6 491 (69.5)	1 636 (17.5)	655 (7.0)	382 (4.1)	170 (1.8)	9 334 (100)
U_{UC} (%)	722 (84.9)	88 (10.4)	28 (3.3)	8 (0.9)	4 (0.5)	850 (100)

From these created sets we can form $U = U_I \cup U_{UC}$, of size $N = 10\ 184$.

The variable in U corresponding to *REV84* was chosen as the study variable y in the Monte Carlo study. From the created population sets we are able to calculate the values of $t_{yU}, t_{yU_{UC}}$ and t_{yU_I}.

Table 6.5 *The total of y and the number of elements in* U, U_{UC} *and* U_I.

	t_y	*Number of elements*
U	27 778 011	10 184
U_{UC}	1 817 452	850
U_I	25 960 559	9 334

Thus we have created a situation with an undercoverage rate of 8.3% ($N_{UC}/N = 850/10\,184$) and an overcoverage rate of 6.7% ($N_{OC}/N_F = 666/10\,000$). Furthermore, we see that $t_{yU_{UC}}/t_{yU} = 1\,817\,452/27\,778\,011 \approx 6.5\%$, i.e. t_{yU_I} underapproximates t_{yU} by roughly 6.5%.

6.2 Auxiliary information, response mechanism and point estimators

Estimators for the total of y in both U_I and U, denoted $\hat{t}_{yU_I}$ and $\hat{t}_{yU}$ in general, were included in the Monte Carlo study.

Auxiliary information at the population level only were considered in the study. Two specifications of the auxiliary vector $\mathbf{x}_k$ were considered, $\mathbf{x}_k = x_{1k}$ and $\mathbf{x}_k = (x_{1k}, x_{2k})'$ respectively, where $x_{1k} = 1$ for all k in U and x_{2k} is a quantitative auxiliary variable. Furthermore, two different choices of x_{2k} were considered, the variables in U corresponding to *CS82* and *ME84* respectively. Henceforth they are referred to as *Case CS* and *Case ME* respectively, while *Case* x_1 will refer to the use of $\mathbf{x}_k = x_{1k} = 1$.

The three cases of auxiliary information represent different strength of association with the study variable. One measure of this strength is the multiple coefficient of determination. Table 6.6 presents the multiple coefficient of determination between y and the three cases of auxiliary vector for different population sets.

Table 6.6 *Multiple coefficient of determination in the sets* U, U_I *and* U_{UC}.

	U	U_I	U_{UC}
Case x_1	0	0	0
Case CS	0.34	0.34	0.29
Case ME	0.82	082	0.80

The population regression lines in U_{UC} and U_I are presented in figures 6.1 and 6.2 for *Case CS* and *Case ME*. The solid line represent regression in U_{UC} and the dotted line regression in U_I, respectively. The lines are printed over the range of x_2 values in U_{UC} and U_I, respectively.

Figure 6.1 *Case CS*

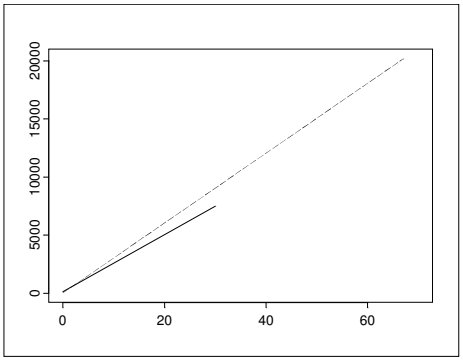
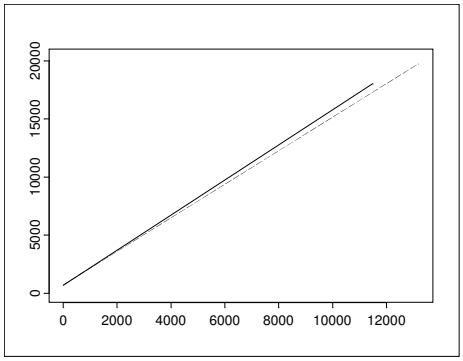


Figure 6.2 *Case ME*



The regression lines show that the approximation $\mathbf{B}_{\mathbf{x}U_{UC}} \approx \mathbf{B}_{\mathbf{x}U_I}$ is better in *Case ME* than in *Case CS*. Since $\mathbf{B}_{\mathbf{x}U_I}$ is estimated approximately unbiased this means that the guesstimation $\hat{\widetilde{t}}^{\widetilde{app}}_{yU_{UC}} = \sum_{U_{UC}} \mathbf{x}'_k \widehat{\mathbf{B}}^{RHG}_{\mathbf{x}r_I}$ will be better in *Case ME* than in *Case CS*. We note that the range of the x_2 values in *Case CS* in U_{UC} is between 0 and 30 and in U_I between 0 and 70. This fact could actually be a reason to use the subgroup U_{Ig}, $k \in g$ if $0 < x_{2k} < 30$, in the approximation of $\mathbf{B}_{\mathbf{x}U_{UC}}$, i.e. $\mathbf{B}_{\mathbf{x}U_{UC}} \approx \mathbf{B}_{\mathbf{x}U_{Ig}}$. However, this is not considered in the study.

We also note that under *Case* x_1 we have $\mathbf{B}_{\mathbf{x}U_{UC}} = B_{xU_{UC}} = \bar{y}_{U_{UC}}$ and $\mathbf{B}_{\mathbf{x}U_I} = B_{xU_I} = \bar{y}_{U_I}$. From table 6.5 we calculate $\bar{y}_{U_{UC}} \approx 2138$ and $\bar{y}_{U_I} \approx 2781$. Clearly, since $\bar{y}_{U_{UC}} \neq \bar{y}_{U_I}$ the guesstimation under *Case* x_1 will be poor.

Six population response groups, U_{Fh}, $h = 0, 1, 2, 3, 4, 5$, were formed where the group U_{F0} corresponds to U_{OC}. The variable in U_F corresponding to *S82* was chosen for this grouping. For $h = 1, \ldots, 5$ all elements in the same group were assigned the same response probability value, denoted θ_h.

Table 6.7 *Response probabilities in five RHG groups*

Population group	N_h	θ_h
U_{I1}	1 921	0.624
U_{I2}	1 795	0.668
U_{I3}	1 833	0.693
U_{I4}	2 271	0.728
U_{I5}	1 514	0.813

Using auxiliary information at population level only, regression based estimators, i.e. $\hat{t}_{yU_I cr}$ from (7) and $\hat{\widetilde{t}}^{\widetilde{app}}_{yU_{cr}}$ from (44), as well as calibration based estimators, i.e. $\hat{t}_{yU_I cal}$ from (23) and $\hat{t}_{yU cal}$ from (54), were used in

the study. Note that we use the special case $\sigma_{1k}^2 = 1$ in $\hat{t}_{yU_I cr}$ and $\hat{t}_{yUcr}^{\widetilde{app}}$. The five groups from table 6.7 were used as RHG groups in the regression based estimators. Furthermore, two different versions of $\hat{t}_{yU_I cr}$ and $\hat{t}_{yUcr}^{\widetilde{app}}$ were considered:

D1: $h = H_{s_F} = 1$ and $f_h^{-1} = f^{-1} = \dfrac{n_{s_I}}{n_{r_I}}$

D2: $h = 1, \ldots, H_{s_F}$ and $f_h^{-1} = \dfrac{n_{s_{Ih}}}{n_{r_{Ih}}}$

Since the regression based estimator using RHG groups explicitly models the nonresponse (in version *D2*) we wanted to create a situation where the RHG estimator and the calibration based estimator used the same level of information. Thus, we created the auxiliary variable $\mathbf{x}_{3k} = (\gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k}, \gamma_{5k})'$, where $\gamma_{1k} = 1$ for all k in group 1, $\gamma_{2k} = 1$ for all k in group 2 and so on. These five groups represent the RHG groups in table 6.7.

Therefore, two versions of $\hat{t}_{yU_I cal}$ and $\hat{t}_{yUcal}$ were considered:

D3: *not* including $\mathbf{x}_{3k}$ in the auxiliary vector
D4: including $\mathbf{x}_{3k}$ in the auxiliary vector

The superscripts *(reg)* and *(reg, RHG)* are used to distinguish estimators using versions *D1* and *D2*, respectively, and superscripts *(cal)* and *(cal,$\boldsymbol{x}_3$)* to distinguish estimators using versions *D3* and *D4*.

It should be noted that we have created a situation favourable for the regression based estimators using the RHG groups in table 6.7 in the sense that these estimators have a negligible nonresponse bias.

Within each case (*Case* x_1, *Case CS* and *Case ME)* the estimators $\hat{t}_{yU_I}^{(reg,RHG)}$ and $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$, as well as $\hat{t}_{yU}^{(reg,RHG)}$ and $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$, are comparable since information on the response behaviour are involved, although in $\hat{t}_{yU_I}^{(reg,RHG)}$ and $\hat{t}_{yU}^{(reg,RHG)}$ the variable $\mathbf{x}_3$ enters indirectly in the RHG groups. Similarly, the estimators $\hat{t}_{yU_I}^{(reg)}$ and $\hat{t}_{yU_I}^{(cal)}$ as well as $\hat{t}_{yU}^{(reg)}$ and $\hat{t}_{yU}^{(cal)}$ are comparable since no information on the response behaviour is involved.

6.3 Analyses of the simulation runs

A simple random sample s_F of size $n_{s_F} = 2\ 000$ was drawn and a response set r_F of size n_{r_F} was generated using independent Bernoulli trials with the

preassigned response probabilities. This was repeated to obtain $M = 20\ 000$ samples and their associated response sets.

We denote by $\hat{t}_y$ any estimator presented above and $\hat{V}$ its estimated variance, and by $\hat{t}_{yU_I}$ an estimator for the total of U_I and by $\hat{t}_{yU}$ an estimator for the total of U. For the mth simulation run, the estimators $\hat{t}_y$ were computed for all cases, *Case* x_1, *Case CS* and *Case ME*, and definitions *D1* to *D4*, as well as their estimated variances. Let $\hat{t}_{y(m)}$ denote the mth value of an arbitrary estimator $\hat{t}_y$ and $\hat{V}_{(m)}$ its estimated variance.

The relative Monte Carlo bias is defined as

$$RB_{MC}\left(\hat{t}_{yU_I}\right) = \left(E_{MC}\left(\hat{t}_{yU_I}\right) - t_{yU_I}\right)/t_{yU_I}$$

and

$$RB_{MC}\left(\hat{t}_{yU}\right) = \left(E_{MC}\left(\hat{t}_{yU}\right) - t_{yU}\right)/t_{yU}$$

where

$$E_{MC}\left(\hat{t}_{yU_I}\right) = \sum_{i=1}^{M} \hat{t}_{yU_I(m)}/M$$

and

$$E_{MC}\left(\hat{t}_{yU}\right) = \sum_{i=1}^{M} \hat{t}_{yU(m)}/M$$

respectively. Furthermore, we define

$$RB_{MC}\left(\hat{t}_{yU_I}, t_{yU}\right) = \left(E_{MC}\left(\hat{t}_{yU_I}\right) - t_{yU}\right)/t_{yU}$$

i.e. the relative Monte Carlo bias using $\hat{t}_{yU_I}$ as an estimator for t_{yU}.

We define the Monte Carlo variance of $\hat{t}_y$

$$V_{MC}\left(\hat{t}_y\right) = \sum_{i=1}^{M} \left(\hat{t}_{y(m)} - E_{MC}\left(\hat{t}_y\right)\right)^2 / (M-1)$$

and the Monte Carlo expectation of $\hat{V}$

$$E_{MC}\left(\hat{V}_{(m)}\right) = \sum_{i=1}^{M} \hat{V}_{(m)}/M$$

Confidence intervals were constructed as

$$CI\left(\hat{t}_{y(m)}\right) = \left[\hat{t}_{y(m)} - 1.96\sqrt{\hat{V}_{(m)}}, \hat{t}_{y(m)} + 1.96\sqrt{\hat{V}_{(m)}}\right]$$

The empirical coverage rate (in %) of the interval $CI\left(\hat{t}_{y(m)}\right)$ is defined as

$$CR_{MC}\left(\hat{t}_y\right) = \frac{100}{M}\sum_{i=1}^{m} I\left(\hat{t}_{y(m)}, t_y\right)$$

where $I\left(\hat{t}_{y(m)}, t_y\right) = \begin{cases} 1 & \text{if } CI\left(\hat{t}_{y(m)}\right) \ni t_y \\ 0 & \text{otherwise} \end{cases}$.

Finally, in order to evaluate the adjustment for undercoverage we define, for the regression based estimators,

$$E_{MC}\left(\hat{t}_{yU_{UC}}^{\widetilde{app}}\right) = \sum_{i=1}^{M} \hat{t}_{yU_{UC}(m)}^{\widetilde{app}}/M$$

where $\hat{t}_{yU_{UC}(m)}^{\widetilde{app}} = \sum_{U_{UC}} \mathbf{x}_k' \widehat{\mathbf{B}}_{\mathbf{x}r_I(m)}^{RHG}$.

Remark 7 *The formulations of definitions* D1 *and* D3 *are such that* $\sum_{r_I} \check{e}_k = 0$, *where* $e_k = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}_{\mathbf{x}r_I}$ *with* $\widehat{\mathbf{B}}_{\mathbf{x}r_I} = \left(\widehat{\mathbf{T}}_{\mathbf{x}r_I}\right)^{-1}\hat{\mathbf{t}}_{\mathbf{x}yr_I} = \left(\sum_{r_I}\mathbf{x}_k\check{\mathbf{x}}_k'\right)^{-1}\sum_{r_I}\mathbf{x}_k\check{y}_k$, *so estimators* $\hat{t}_{yU_I}^{(reg)}$ *and* $\hat{t}_{yU_I}^{(cal)}$ *agree and we have* $\hat{t}_{yU_I}^{(reg)} = \hat{t}_{yU_I}^{(cal)} = \mathbf{t}_{\mathbf{x}U_I}'\widehat{\mathbf{B}}_{\mathbf{x}r_I}$. *Furthermore, we have* $\hat{t}_{yU}^{(reg)} = \hat{t}_{yU}^{(cal)} = \mathbf{t}_{\mathbf{x}U}'\widehat{\mathbf{B}}_{\mathbf{x}r_I}$.

6.4 Results

The Monte Carlo sampling distributions of point estimators are presented in appendix B.

6.4.1 $\hat{t}_{yU_I}$ as an estimator of t_{yU_I}

We start by analysing the properties of $\hat{t}_{yU_I}$ with respect to estimating t_{yU_I}. This means that in this case the only nonsampling error is due to nonresponse.

From section 4.1 we know that, for large samples, $\hat{t}_{yU_I}^{(reg,RHG)}$ has a negligible bias and that the variance estimator of $\hat{t}_{yU_I}^{(reg,RHG)}$ has small bias. We also

have that $\hat{t}_{yU_I}^{(cal)}$ and $\hat{t}_{yU_I}^{(reg)}$ coincide for all cases in this study and therefore have the same (approximate) bias and variance. However, the variance estimators need not be the same. The (approximate) bias of $\hat{t}_{yU_I}^{(cal)}(=\hat{t}_{yU_I}^{(reg)})$ and $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$ may be calculated from (25) for all cases.

Table 6.8 *Relative approximate bias of* $\hat{t}_{yU_I}^{(cal)}(=\hat{t}_{yU_I}^{(reg)})$, $\hat{t}_{yU_I}^{(reg,RHG)}$ *and* $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}, \hat{t}_{yU_I}^{(cal)}$	0.0223	0.0124	0.0005
$\hat{t}_{yU_I}^{(reg,RHG)}$	appr. 0	appr. 0	appr. 0
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$	appr. 0	appr. 0	appr. 0

We compare these results with the results for the relative Monte Carlo bias.

Table 6.9 *The relative Monte Carlo bias of* $\hat{t}_{yU_I}, RB_{MC}\left(\hat{t}_{yU_I}\right)$, *with the standard error of* $RB_{MC}\left(\hat{t}_{yU_I}\right)$ *within parentheses*

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}, \hat{t}_{yU_I}^{(cal)}$	0.0536	0.0288	0.0019
	(0.00016)	(0.00013)	(0.00007)
$\hat{t}_{yU_I}^{(reg,RHG)}$	0.0003	0.0002	0.0001
	(0.00013)	(0.00012)	0.00006)
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$	0.0001	0.0001	0.0001
	(0.00011)	(0.00010)	0.00007)

Since the standard error is negligible the difference in results between table 6.8 and table 6.9 for $\hat{t}_{yU_I}^{(cal)}$ and $\hat{t}_{yU_I}^{(reg)}$ is not due to simulation effects. As it seems, the approximate bias from (25) underestimates the bias for these estimators. From (74) in appendix A.2 we have that the error of $\hat{t}_{yU_I cal}$ involves the term

$$\left(\sum\nolimits_{r_I} \check{\mathbf{x}}_k - \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix}\right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I}\right)$$

which in our setup with auxiliary information $\mathbf{x}_k$ at population level only, is

$$-\left(\sum\nolimits_{r_I} \check{\mathbf{x}}_k - \sum\nolimits_{U_I} \mathbf{x}_k\right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I}\right) \tag{67}$$

Calculating this term for the 20 000 simulation runs, summing them and dividing by $(20000 \cdot t_{yU_I})$ we have, for the three cases, 0.0311, 0.0162 and 0.0014, respectively. Adding these to the values for $\hat{t}_{yU_I}^{(cal)}$ and $\hat{t}_{yU_I}^{(reg)}$in table 6.8 the bias now is at the same level as the relative Monte Carlo bias.

Table 6.10 *Adjusted relative approximate bias of* $\hat{t}_{yU_I}^{(cal)} (= \hat{t}_{yU_I}^{(reg)})$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}, \hat{t}_{yU_I}^{(cal)}$	0.0534	0.0286	0.0019

Thus, in our particular setup the bias approximation is poor for these estimators.

Remark 8 *The approximate bias in* (25) *may need to be adjusted. However, this is not considered any further here.*

From table 6.9 we see that the estimators using information on the response behaviour are, not surprisingly, approximately unbiased. However, when no information on the response behaviour exist the estimators using an auxiliary vector with strong association with the study variable also perform well as the results from $\hat{t}_{yU_I}^{(reg)}$ and $\hat{t}_{yU_I}^{(cal)}$ under *Case ME* indicate. This is further illustrated by figure B.1 to B.12 in appendix B.

For the calibration based estimators we used two alternative variance estimators in this study, i.e. $\hat{V}\left(\hat{t}_{yU_I cal}\right)$ from (26) using

$$v_{Ik} = v_{U_I k} = 1 + \left(\sum\nolimits_{U_I} \mathbf{x}_k - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{68}$$

as well as

$$v_{Ik} = v_{s_I k} = 1 + \left(\sum\nolimits_{s_I} \check{\mathbf{x}}_k - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{69}$$

As we noted before, the variance estimator of $\hat{t}_{yU_I}^{(reg,RHG)}$ has small bias, which is confirmed by the results in table 6.11.

Table 6.11 $E_{MC}\left(V_{(m)}\right)/V_{MC}\left(\hat{t}_{yU_I}\right)$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}$	1.0015	1.0802	1.1544
$\hat{t}_{yU_I}^{(reg,RHG)}$	0.9992	1.0027	0.9914
$\hat{t}_{yU_I}^{(cal)}\left(v_{U_I k}\right)$	1.1649	1.1691	1.1678
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}\left(v_{U_I k}\right)$	1.1586	1.1652	1.1619
$\hat{t}_{yU_I}^{(cal)}\left(v_{s_I k}\right)$	1.1649	1.1684	1.1667
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}\left(v_{s_I k}\right)$	1.1568	1.1624	1.1587

The variance estimator of the calibration estimators $\hat{t}_{yU_I}^{(cal)}$ and $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$ overestimate the variance by approximately 16% in all cases. The same thing may be said of the variance estimator for $\hat{t}_{yU_I}^{(reg)}$ in the case of strong auxiliary vector information.

The results for the empirical coverage rate is a reflection of the results for the relative Monte Carlo bias and the Monte Carlo expectation of $\hat{V}\left(\hat{t}_{yU_I}\right)$. An estimator with high relative Monte Carlo bias shows a poor empirical coverage rate. When the variance is overestimated this somewhat increases the empirical coverage rate (compare CR_{MC} of $\hat{t}_{yU_I}^{(reg)}$ and CR_{MC} of $\hat{t}_{yU_I}^{(cal)}$, estimators with the same RB_{MC}).

Table 6.12 *The empirical coverage rate of* $\hat{t}_{yU_I}, CR_{MC}\left(\hat{t}_{yU_I}\right)$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}$	32.4	65.7	96.2
$\hat{t}_{yU_I}^{(reg,RHG)}$	94.9	94.9	95.0
$\hat{t}_{yU_I}^{(cal)}\left(v_{U_I k}\right)$	38.0	68.8	96.2
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}\left(v_{U_I k}\right)$	96.3	96.5	96.5
$\hat{t}_{yU_I}^{(cal)}\left(v_{s_I k}\right)$	38.0	68.8	96.5
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}\left(v_{s_I k}\right)$	96.3	96.5	96.4

6.4.2 $\hat{t}_{yU_I}$ as an estimator of t_{yU}

In our particular setup we have $t_{yU_{UC}}/t_{yU} \approx 6.5\%$. This means that if we use an (approximately) unbiased estimator for t_{yU_I} in order to estimate t_{yU} the

bias, with respect to t_{yU}, would be (approximately) -0.065.

Table 6.13 shows the relative Monte Carlo bias using $\hat{t}_{yU_I}$ as an estimator for t_{yU}.

Table 6.13 *The relative Monte Carlo bias using* $\hat{t}_{yU_I}$ *for estimating* t_{yU}, $RB_{MC}\left(\hat{t}_{yU_I}, t_{yU}\right)$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU_I}^{(reg)}, \hat{t}_{yU_I}^{(cal)}$	-0.0153	-0.0385	-0.0636
$\hat{t}_{yU_I}^{(reg,RHG)}$	-0.0652	-0.0652	-0.0653
$\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$	-0.0653	-0.0653	-0.0654

For all estimators and all three cases of auxiliary information the relative Monte Carlo bias is negative. However, the results in table 6.13 are implied already by the results in table 6.9 together with the fact that $t_{yU_{UC}}/t_{yU} \approx 6.5\%$. An approximately unbiased estimator (with respect to t_{yU_I}) will obviously be biased when it comes to estimate t_{yU}. From table 6.13 we notice that estimators $\hat{t}_{yU_I}^{(reg)}$ and $\hat{t}_{yU_I}^{(cal)}$ under *Case* x_1 have the least bias with respect to t_{yU}. This should not be interpreted as if these estimators perform well in estimating t_{yU}. Rather our results are a consequence of the nonresponse bias that affects these estimators. Another set of response probabilities in the Monte Carlo study would have given another result.

6.4.3 $\hat{t}_{yU}$ as an estimator of t_{yU}

In order to analyse the properties of $\hat{t}_{yU}$ we start by looking at the guesstimation of $t_{yU_{UC}}$. Recall that we want

$$t_{yU_{UC}}^{app} = \sum\nolimits_{U_{UC}} \mathbf{x}_k' \mathbf{B}_{\mathbf{x}U_{UC}}$$

to be a good approximation for $t_{yU_{UC}}$ in order to find a good adjustment for the undercoverage. Moreover, since we cannot estimate $\mathbf{B}_{\mathbf{x}U_{UC}}$ we make the assumption that $\mathbf{B}_{\mathbf{x}U_{UC}} \approx \mathbf{B}_{\mathbf{x}U_I}$ and instead we use

$$\widetilde{t_{yU_{UC}}^{app}} = \sum\nolimits_{U_{UC}} \mathbf{x}_k' \mathbf{B}_{\mathbf{x}U_I}$$

since it is possible to find an estimate of $\widetilde{t_{yU_{UC}}^{app}}$ from the response set.

In table 6.14 we show $\left(t_{yU_{UC}}^{app} - t_{yU_{UC}}\right)/t_{yU_{UC}}$ and $\left(t_{yU_{UC}}^{\widetilde{app}} - t_{yU_{UC}}\right)/t_{yU_{UC}}$, respectively, where $t_{yU_{UC}}^{\widetilde{app}} - t_{yU_{UC}}$ is the approximation error of $t_{yU_{UC}}^{\widetilde{app}}$ from (34) .

Table 6.14 $\left(t_{yU_{UC}}^{app} - t_{yU_{UC}}\right)/t_{yU_{UC}}$ *and* $\left(t_{yU_{UC}}^{\widetilde{app}} - t_{yU_{UC}}\right)/t_{yU_{UC}}$

	Case x_1	*Case CS*	*Case ME*
$t_{yU_{UC}}^{app}$	0.0	0.0	0.0
$t_{yU_{UC}}^{\widetilde{app}}$	0.3	0.2	0.0

We notice that, for our artificial population $t_{yU_{UC}}^{app}$ works well for all cases of auxiliary information. This is due to the definition of the regression coefficient which results in $t_{yU_{UC}}^{app} = t_{yU_{UC}}$. However, for $t_{yU_{UC}}^{\widetilde{app}}$ *Case ME* produces the best approximation. This is implied already in figure 6.1 and 6.2 (see section 6.2) where the regression lines in U_{UC} and U_I displayed that the approximation $\mathbf{B}_{\mathbf{x}U_{UC}} \approx \mathbf{B}_{\mathbf{x}U_I}$ is better in *Case ME* than in *Case CS*.

The next step is to analyse how the actual guesstimate of $t_{yU_{UC}}$ for the regression based estimators, i.e. how

$$\hat{t}_{yU_{UC}}^{\widetilde{app}} = \sum\nolimits_{U_{UC}} \mathbf{x}_k' \widehat{\mathbf{B}}_{\mathbf{x}r_I}^{RHG}$$

performed. From the study we get the following results:

Table 6.15 $\left(E_{MC}\left(\hat{t}_{yU_{UC}}^{\widetilde{app}}\right) - t_{yU_{UC}}\right)/t_{yU_{UC}}$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}$	0.3705	0.2145	0.0067
$\hat{t}_{yU}^{(reg,RHG)}$	0.3011	0.1793	0.0028

Since $\mathbf{B}_{\mathbf{x}U_I}$ is estimated approximately unbiasedly by $\widehat{\mathbf{B}}_{\mathbf{x}r_I}^{RHG}$ the use of an auxiliary variable where the approximation $\mathbf{B}_{\mathbf{x}U_{UC}} \approx \mathbf{B}_{\mathbf{x}U_I}$ works well also makes the guesstimation less biased. However, the use of information on the response mechanism from the *RHG* groups has modest effect on the guesstimation in the present particular simulation setup.

The (approximate) bias of $\hat{t}_{yU}^{(cal)} (= \hat{t}_{yU}^{(reg)})$, $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ and $\hat{t}_{yU}^{(reg,RHG)}$ may be calculated from (36) and (58) for all cases.

Table 6.16 *Relative approximate bias of* $\hat{t}_{yU}^{(cal)}(=\hat{t}_{yU}^{(reg)})$, $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ *and* $\hat{t}_{yU}^{(reg,RHG)}$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}, \hat{t}_{yU}^{(cal)}$	0.0362	0.0211	0.0004
$\hat{t}_{yU}^{(reg,RHG)}$	0.0197	0.0117	0.0002
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$	0.0055	0.0048	0.0002

We compare these results with the results for the relative Monte Carlo bias.

Table 6.17 *The relative Monte Carlo bias of* $\hat{t}_{yU}, RB_{MC}\left(\hat{t}_{yU}\right)$*,with the standard error of* $RB_{MC}\left(\hat{t}_{yU}\right)$ *within parentheses*

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}, \hat{t}_{yU}^{(cal)}$	0.0744	0.0409	0.0022
	(0.00016)	(0.00014)	(0.00007)
$\hat{t}_{yU}^{(reg,RHG)}$	0.0200	0.0119	0.0003
	(0.00014)	(0.00012)	(0.00007)
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$	0.0057	0.0049	0.0003
	(0.00011)	(0.00010)	(0.00007)

The bias approximation works well for estimators $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ and $\hat{t}_{yU}^{(reg,RHG)}$. Analysing the approximate bias of $\hat{t}_{yU}^{(cal)}(=\hat{t}_{yU}^{(reg)})$, we have, from (76) in appendix *A*.3, the term

$$-\left(\sum\nolimits_{r_I}\check{\mathbf{x}}_k-\begin{pmatrix}\sum_{U_I}\mathbf{x}_{1k}\\ \sum_{s_I}\check{\mathbf{x}}_{2k}\end{pmatrix}\right)'\left(\widehat{\mathbf{B}}_{r_I}-\mathbf{B}_{\boldsymbol{\theta}U_I}\right)+\sum\nolimits_{U_{UC}}\mathbf{x}_{1k}'\left(\widehat{\mathbf{B}}_{r_I}^{(1)}-\mathbf{B}_{\boldsymbol{\theta}U_I}^{(1)}\right)$$

which, in our setup with auxiliary information $\mathbf{x}_k$ at population level only, is

$$-\left(\sum\nolimits_{r_I}\check{\mathbf{x}}_k-\sum\nolimits_{U_I}\mathbf{x}_k\right)'\left(\widehat{\mathbf{B}}_{r_I}-\mathbf{B}_{\boldsymbol{\theta}U_I}\right)+\sum\nolimits_{U_{UC}}\mathbf{x}_k'\left(\widehat{\mathbf{B}}_{r_I}-\mathbf{B}_{\boldsymbol{\theta}U_I}\right) \qquad (70)$$

(compare (67)). Starting with summation of the first term in (70) and dividing by $(20000\cdot t_{yU})$ we have, for the three cases 0.0291, 0.0151 and 0.0013, respectively. Adding these to the values for $\hat{t}_{yU}^{(cal)}$ and $\hat{t}_{yU}^{(reg)}$ in table 6.16 the relative bias gets closer to the level of the relative Monte Carlo bias.

Table 6.18 *Adjusted relative approximate bias of* $\hat{t}_{yU}^{(cal)} (= \hat{t}_{yU}^{(reg)})$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}, \hat{t}_{yU}^{(cal)}$	0.0653	0.0362	0.0017

Using the same procedure for the second term of (70) and adding to the values in table 6.18 we get 0.0742, 0.0408 and 0.0022, even closer to the level of the relative Monte Carlo bias. Thus, in our particular setup the approximation is poor for these estimators.

Remark 9 *The approximate bias in* (58) *may need to be adjusted. However, this is not considered any further here.*

Looking at table 6.13 and specifically at the estimators with a relative bias (with respect to t_{yU}) of approximately -6.5%, i.e. $\hat{t}_{yU_I}^{(reg,RHG)}$ and $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$, and comparing with their counterparts from table 6.17 the latter show a bias reduction (with respect to t_{yU}).

For the calibration based estimators we used two alternative variance estimators in the study, i.e. $\hat{V}\left(\hat{t}_{yUcal,1}\right)$ from (48) using

$$\varphi_k = \varphi_{s_I k} = 1 + \left(\sum\nolimits_{s_I} \check{\mathbf{x}}_k - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{71}$$

and $\hat{V}\left(\hat{t}_{yUcal}\right)$ from (59) using

$$v_k = v_{Uk} = 1 + \left(\sum\nolimits_{U} \mathbf{x}_k - \sum\nolimits_{r_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{72}$$

Table 6.19 $E_{MC}\left(V_{(m)}\right)/V_{MC}\left(\hat{t}_{yU}\right)$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}$	0.9058	1.0753	1.1390
$\hat{t}_{yU}^{(reg,RHG)}$	0.8639	1.0029	0.9918
$\hat{t}_{yU}^{(cal)}\left(\varphi_{s_I k}\right)$	0.9785	1.0015	1.0278
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}\left(\varphi_{s_I k}\right)$	1.0132	1.0195	1.0238
$\hat{t}_{yU}^{(cal)}\left(v_{Uk}\right)$	1.1541	1.1585	1.1569
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}\left(v_{Uk}\right)$	1.1485	1.1547	1.1517

Using v_{Uk} in the variance estimator for the calibration based estimators the variance is overestimated by approximately 15%. The variance estimators using $\varphi_{s_I k}$ performed better in our particular setup. Since, in v_{Uk}, we have $\sum_U \mathbf{x}_k - \sum_{r_I} \check{\mathbf{x}}_k$ where $\sum_{r_I} \check{\mathbf{x}}_k$ underestimates $\sum_U \mathbf{x}_k$ on the average the weight v_{Uk} is, on the average, too large, making the estimated variance (on the average) too large.

The results for the regression estimators exhibit a mixed picture. In particular, *Case* x_1 appears to give variance underestimation, while in *Case ME* (strong auxiliary information) the variance of the simple estimator $\hat{t}_{yU}^{(reg)}$ is overestimated.

The results on variance estimation merit further theoretical and empirical studies, a task that lies outside the scope of the present paper.

Finally, we take a look at the empirical coverage rate of $\hat{t}_{yU}$.

Table 6.20 *The empirical coverage rate of* $\hat{t}_{yU}, CR_{MC}\left(\hat{t}_{yU}\right)$

	Case x_1	*Case CS*	*Case ME*
$\hat{t}_{yU}^{(reg)}$	6.8	38.5	95.8
$\hat{t}_{yU}^{(reg,RHG)}$	79.2	89.5	95.1
$\hat{t}_{yU}^{(cal)}\left(\varphi_{s_I k}\right)$	7.9	35.9	94.7
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}\left(\varphi_{s_I k}\right)$	93.6	94.0	95.3
$\hat{t}_{yU}^{(cal)}\left(v_{Uk}\right)$	10.8	41.6	96.0
$\hat{t}_{yU}^{(cal,\mathbf{x}_3)}\left(v_{Uk}\right)$	95.2	95.5	96.4

These results clearly states that, in our particular setup, the use of information on the response mechanism will result in an empirical coverage rate approximately at the nominal 95% confidence level, with exception of $\hat{t}_{yU}^{(reg,RHG)}$ under *Case* x_1 and *Case CS*. The reason for this exception is that the guesstimation of $t_{yU_{UC}}$ is poor under these cases. The corresponding calibration based estimators show allover good coverage rates. However, if no information on the response mechanism is available the use of auxiliary information highly correlated with y also works well. Leaving out both of these the coverage rate deteriorates drastically.

It should be noted that since the Monte Carlo study is based on one specific population, with specific characteristics for U_I, U_{UC} and U, further studies using populations with other characteristics for the over- and undercoverage is needed to be able to draw more general conclusions. It should

also be noted that only a single study variable and only the estimation of a population total have been considered in this study.

7 Concluding remarks

In the situation where the statistician has access to a perfect current register the prospects of adjusting for frame error bias increases. Information on whether elements in the sample belong to the target population or to the overcoverage set exist and (approximately) unbiased domain estimators may be used in order to estimate the total of U_I. Moreover, we have current auxiliary information for all elements in the target population which provides information on the elements in undercoverage set. This information could be used to find guesstimates of the total in the undercoverage set. In this paper we have presented a way to use this (current) auxiliary information in regression based estimators and calibration based estimators. Along with these suggested estimators we have proposed variance estimators as well as general expressions for the (approximate) bias.

Results from a small Monte Carlo study confirm that doing nothing about the problem with undercoverage leads to negative bias for an always positive study variable. Using information on the response mechanism together with an auxiliary vector with strong association with the study variable the suggested estimators for the total of U show an empirical coverage rate approximately at the nominal 95% confidence level.

Note that the Monte Carlo study is based on one specific population, with specific characteristics for U_I, U_{UC} and U, and further studies using populations with other characteristics for the over- and undercoverage is needed to be able to draw more general conclusions. It should also be noted that only a single study variable and only the estimation of a population total have been considered in this study.

In practice, the estimation situation with access to a perfect current register may be somewhat unrealistic. A situation where an updated (but imperfect) current register is at hand at the estimation stage of a survey is more likely to appear. The following situation is an example on such a situation. Consider a survey setup where the sample is drawn in month m and the questionnaire is sent out to the sampled elements in months m to $m+11$, say. If the frame is updated during this period the updated frame is a potential candidate to serve as a current register. Issues along this line need to be studied further.

References

Ängsved, M. (2004). Estimating the finite population total under frame imperfections. ESI Working Paper Series 2004:3, Örebro University.

Axelson, M. (2000). *On Variance Estimation for the Two-Phase Regression Estimator*. Ph. D. thesis, Department of Statistics, Uppsala University.

Estevao, V. and Särndal, C. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics* **18**, 233–255.

Kish, L. (1979). Populations for survey sampling. *Survey Statistician* **1**, 14–15.

Lessler, J. and Kalsbeek, W. (1992). *Nonsampling Errors in Surveys*. New York: Wiley.

Murthy, M. (1983). A framework for studying incomplete data with a reference to the experience in some countries of asia and the pacific. In W. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys: Volume 3 Proceedings of the Symposium*, New York, pp. 7–24. Academic Press.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley.

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

Tångdahl, S. (2004). Nonresponse bias for some common estimators and its change over time in the data collection process. ESI Working Paper Series 2004:13, Örebro University.

Appendix A: Derivations

A.1 Taylor linearization of $\hat{t}^{\widetilde{app}}_{yUcr}$

We start by rewriting $\hat{t}^{\widetilde{app}}_{yUcr}$,

$$\begin{aligned}
\hat{t}^{\widetilde{app}}_{yUcr} &= \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \breve{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{1k} - \sum_{h=1}^{H_{s_F}} \sum\nolimits_{s_{Ih}} \breve{\mathbf{x}}_{1k} \right)' \widehat{\mathbf{B}}^{RHG}_{1r_I} \\
&\quad + \left(\sum_{h=1}^{H_{s_F}} \sum\nolimits_{s_{Ih}} \breve{\mathbf{x}}_k - \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \breve{\mathbf{x}}_k \right)' \widehat{\mathbf{B}}^{RHG}_{r_I} + \sum\nolimits_{U_{UC}} \mathbf{u}'_k \widehat{\mathbf{B}}^{RHG}_{\mathbf{u}r_I} \\
&= \sum_{h=1}^{H_{s_F}} \hat{t}_{yr_{Ih}} + \left(\mathbf{t}_{x_1 U_I} - \hat{\mathbf{t}}_{x_1 s_I} \right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}} \right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{x}_1 y r_{Ih}} \\
&\quad + \left(\hat{\mathbf{t}}_{x s_I} - \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{x r_{Ih}} \right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x} r_{Ih}} \right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{x} y r_{Ih}} \\
&\quad + \mathbf{t}'_{u U_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u} r_{Ih}} \right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{u} y r_{Ih}} \\
&= f\left(\hat{t}_{yr_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}}, \hat{\mathbf{t}}_{\mathbf{x}_1 y r_{Ih}}, \hat{\mathbf{t}}_{x r_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{x} r_{Ih}}, \hat{\mathbf{t}}_{\mathbf{x} y r_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{u} r_{Ih}}, \hat{\mathbf{t}}_{\mathbf{u} y r_{Ih}}; h = 1, \ldots, H_{s_F} \right)
\end{aligned}$$

We will need the following partial derivatives

$$\frac{\partial f}{\partial \hat{t}_{yr_{Ih}}} = 1$$

$$\frac{\partial f}{\partial \hat{t}_{j_1 j'_1 r_{Ih}}} = -\left(\mathbf{t}_{x_1 U_I} - \hat{\mathbf{t}}_{x_1 s_I} \right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}} \right)^{-1} \Phi_{j_1 j'_1} \widehat{\mathbf{B}}^{RHG}_{1r_I}$$

$$\frac{\partial f}{\partial \hat{t}_{j_1 y r_{Ih}}} = \left(\mathbf{t}_{x_1 U_I} - \hat{\mathbf{t}}_{x_1 s_I} \right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}} \right)^{-1} \phi_{j_1}$$

$$\frac{\partial f}{\partial \hat{t}_{jr_{Ih}}} = -\boldsymbol{\phi}_j' \widehat{\mathbf{B}}_{r_I}^{RHG}$$

$$\frac{\partial f}{\partial \hat{t}_{jj'r_{Ih}}} = -\left(\hat{\mathbf{t}}_{xs_I} - \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{xr_{Ih}}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}r_{Ih}}\right)^{-1} \Phi_{jj'} \widehat{\mathbf{B}}_{r_I}^{RHG}$$

$$\frac{\partial f}{\partial \hat{t}_{jyr_{Ih}}} = \left(\hat{\mathbf{t}}_{xs_I} - \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{xr_{Ih}}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}r_{Ih}}\right)^{-1} \boldsymbol{\phi}_j$$

$$\frac{\partial f}{\partial \hat{t}_{pp'r_{Ih}}} = -\mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}\right)^{-1} \Phi_{pp'} \widehat{\mathbf{B}}_{\mathbf{u}r_I}^{RHG}$$

$$\frac{\partial f}{\partial \hat{t}_{pyr_{Ih}}} = \mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}}\right)^{-1} \boldsymbol{\phi}_p$$

where $\Phi_{jj'}$ is a $J \times J$ matrix with the value 1 in positions (j, j') and (j', j) and the value 0 everywhere else and $\boldsymbol{\phi}_j$ is a j-vector with the jth component equal to one and zeros elsewhere. $\Phi_{pp'}$ and $\boldsymbol{\phi}_p$ follow analogously.

Furthermore, we will need the conditional expected values:

$E_{RD}\left(\hat{t}_{yr_{Ih}} | s_F\right) = E_{RD}\left(f_{Ih}^{-1} \sum_{r_{Ih}} \frac{y_k}{\pi_{Fk}} | s_F\right) = \sum_{s_{Ih}} y_k/\pi_{Fk} = \hat{t}_{ys_{Ih}}$

$E_{RD}\left(\widehat{\mathbf{T}}_{\mathbf{x}_1 r_{Ih}} | s_F\right) = E_{RD}\left(f_{Ih}^{-1} \sum_{r_{Ih}} \frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\pi_{Fk}\sigma_{1k}^2} | s_F\right) = \sum_{s_{Ih}} \frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\pi_{Fk}\sigma_{1k}^2} = \widehat{\mathbf{T}}_{\mathbf{x}_1 s_{Ih}}$

The conditional expected values of $\hat{\mathbf{t}}_{\mathbf{x}_1 y r_{Ih}}, \hat{\mathbf{t}}_{x r_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{x} r_{Ih}}, \hat{\mathbf{t}}_{\mathbf{x} y r_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{u} r_{Ih}}$ and $\hat{\mathbf{t}}_{\mathbf{u} y r_{Ih}}$ follow analogously.

Evaluating the partial derivatives at the expected value point $\left(\hat{t}_{ys_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{x}_1 s_{Ih}}, \hat{\mathbf{t}}_{\mathbf{x}_1 y s_{Ih}}, \hat{\mathbf{t}}_{x s_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{x} s_{Ih}}, \hat{\mathbf{t}}_{\mathbf{x} y s_{Ih}}, \widehat{\mathbf{T}}_{\mathbf{u} s_{Ih}}, \hat{\mathbf{t}}_{\mathbf{u} y s_{Ih}}\right)$ the partial derivatives $\frac{\partial f}{\partial \hat{t}_{jj'r_{Ih}}}$ and $\frac{\partial f}{\partial \hat{t}_{jyr_{Ih}}}$ conveniently vanish and we obtain

$$
\begin{aligned}
\hat{t}_{yUcr}^{\widetilde{app}} \approx & \sum\nolimits_{s_I} \frac{y_k}{\pi_{Fk}} + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \sum_{h=1}^{H_{s_F}} \left(\hat{t}_{yr_{Ih}} - \hat{t}_{ys_{Ih}}\right) \\
& - \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \left(\widehat{\mathbf{T}}_{\mathbf{x}_1r_{Ih}} - \widehat{\mathbf{T}}_{\mathbf{x}_1s_{Ih}}\right) \widehat{\mathbf{B}}_{1s_I} \\
& + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \left(\hat{\mathbf{t}}_{\mathbf{x}_1yr_{Ih}} - \hat{\mathbf{t}}_{\mathbf{x}_1ys_{Ih}}\right) \\
& - \sum_{h=1}^{H_{s_F}} \left(\hat{\mathbf{t}}_{xr_{Ih}} - \hat{\mathbf{t}}_{xs_{Ih}}\right)' \widehat{\mathbf{B}}_{s_I} + \mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \left(\hat{\mathbf{t}}_{\mathbf{u}yr_{Ih}} - \hat{\mathbf{t}}_{\mathbf{u}ys_{Ih}}\right) \\
& - \mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \left(\widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}} - \widehat{\mathbf{T}}_{\mathbf{u}s_{Ih}}\right) \widehat{\mathbf{B}}_{\mathbf{u}s_I}
\end{aligned}
$$

Simplification yields

$$
\begin{aligned}
\hat{t}_{yUcr}^{\widetilde{app}} \approx & \sum_{h=1}^{H_{s_F}} \hat{t}_{yr_{Ih}} + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} \\
& - \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1r_{Ih}} \widehat{\mathbf{B}}_{1s_I} + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} \\
& + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{x}_1s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{x}_1yr_{Ih}} - \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} \\
& - \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}'_{xr_{Ih}} \widehat{\mathbf{B}}_{s_I} + \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}'_{xs_{Ih}} \widehat{\mathbf{B}}_{s_I} - \mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} \\
& + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \mathbf{t}'_{uU_{UC}} \left(\sum_{h=1}^{H_{s_F}} \widehat{\mathbf{T}}_{\mathbf{u}s_{Ih}}\right)^{-1} \sum_{h=1}^{H_{s_F}} \hat{\mathbf{t}}_{\mathbf{u}yr_{Ih}} - \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I}
\end{aligned}
$$

We get

$$\begin{aligned}
\widehat{t_{yUcr}^{app}} &\approx \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \hat{\mathbf{t}}'_{xs_I} \widehat{\mathbf{B}}_{s_I} \\
&\quad + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_1s_I} \sum_{h=1}^{H_{s_F}} \left(\hat{\mathbf{t}}_{\mathbf{x}_1yr_{Ih}} - \widehat{\mathbf{T}}_{\mathbf{x}_1r_{Ih}} \widehat{\mathbf{B}}_{1s_I}\right) \\
&\quad + \sum_{h=1}^{H_{s_F}} \left(\hat{t}_{yr_{Ih}} - \hat{\mathbf{t}}'_{xr_{Ih}} \widehat{\mathbf{B}}_{s_I}\right) + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{T}}^{-1}_{\mathbf{u}s_I} \sum_{h=1}^{H_{s_F}} \left(\hat{\mathbf{t}}_{\mathbf{u}yr_{Ih}} - \widehat{\mathbf{T}}_{\mathbf{u}r_{Ih}} \widehat{\mathbf{B}}_{\mathbf{u}s_I}\right) \\
&= \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \hat{\mathbf{t}}'_{xs_I} \widehat{\mathbf{B}}_{s_I} \\
&\quad + \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_1s_I} \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{x}_{1k}\hat{E}_{1k}}{\pi_{Fk}\sigma^2_{1k}} \\
&\quad + \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{\hat{E}_k}{\pi_{Fk}} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{T}}^{-1}_{\mathbf{u}s_I} \sum_{h=1}^{H_{s_F}} f_h^{-1} \sum\nolimits_{r_{Ih}} \frac{\mathbf{u}_k\hat{E}_{uk}}{\sigma^2_{uk}\pi_{Fk}}
\end{aligned}$$

where $\hat{E}_{1k} = y_k - \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1s_I}$ $\hat{E}_k = y_k - \mathbf{x}'_k\widehat{\mathbf{B}}_{s_I}$ and $\hat{E}_{uk} = y_k - \mathbf{u}'_k\widehat{\mathbf{B}}_{\mathbf{u}s_I}$.
Rewriting again we obtain

$$\begin{aligned}
\widehat{t_{yUcr}^{app}} &\approx \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \hat{\mathbf{t}}'_{xs_I} \widehat{\mathbf{B}}_{s_I} \\
&\quad + \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \left(\frac{\left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_1s_I} \mathbf{x}_{1k}\hat{E}_{1k}}{\pi_{Fk}\sigma^2_{1k}} + \frac{\hat{E}_k}{\pi_{Fk}} + \frac{\mathbf{t}'_{uU_{UC}} \widehat{\mathbf{T}}^{-1}_{\mathbf{u}s_I} \mathbf{u}_k\hat{E}_{uk}}{\sigma^2_{uk}\pi_{Fk}} \right) \\
&= \left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{B}}_{1s_I} + \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{B}}_{\mathbf{u}s_I} + \hat{\mathbf{t}}'_{xs_I} \widehat{\mathbf{B}}_{s_I} \\
&\quad + \sum_{h=1}^{H_{s_F}} f_{Ih}^{-1} \sum\nolimits_{r_{Ih}} \frac{F_{xuk}}{\pi_{Fk}}
\end{aligned}$$

where $F_{xuk} = \dfrac{\left(\mathbf{t}_{x_1U_I} - \hat{\mathbf{t}}_{x_1s_I}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_1s_I} \mathbf{x}_{1k}\hat{E}_{1k}}{\sigma^2_{1k}} + \hat{E}_k + \dfrac{\mathbf{t}'_{uU_{UC}} \widehat{\mathbf{T}}^{-1}_{\mathbf{u}s_I} \mathbf{u}_k\hat{E}_{uk}}{\sigma^2_{uk}}$

A.2 Derivation of approximate bias of $\hat{t}_{yU_I cal}$

The derivation of the approximate bias of $\hat{t}_{yU_I cal}$ follow the approach from Särndal and Lundström (2005).

The calibration estimator $\hat{t}_{yU_I cal}$ can be written as

$$\hat{t}_{yU_I cal} = \sum\nolimits_{r_I} \check{y}_k + \left[\begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \widehat{\mathbf{B}}_{r_I} \tag{73}$$

where

$$\widehat{\mathbf{B}}_{r_I} = \begin{bmatrix} \widehat{\mathbf{B}}_{r_I}^{(1)} \\ \widehat{\mathbf{B}}_{r_I}^{(2)} \end{bmatrix} = \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k' \right)^{-1} \sum\nolimits_{r_I} \mathbf{x}_k \check{y}_k$$

For large response sets, $\widehat{\mathbf{B}}_{r_I}$ is close in probability to the vector

$$\mathbf{B}_{\boldsymbol{\theta} U_I} = \begin{bmatrix} \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \\ \mathbf{B}_{\boldsymbol{\theta} U_I}^{(2)} \end{bmatrix} = \left(\sum\nolimits_{U_I} \theta_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum\nolimits_{U_I} \theta_k \mathbf{x}_k y_k$$

Now we rewrite (73) to obtain

$$\begin{aligned} \hat{t}_{yU_I cal} \;=\; & \sum\nolimits_{r_I} \check{y}_k + \left[\begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \mathbf{B}_{\boldsymbol{\theta} U_I} \\ & + \left[\begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right) \end{aligned}$$

Note that the last term $\left(\sum_{r_I} \check{\mathbf{x}}_k - \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} \right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right)$ is of lower order of importance, compared to the preceeding two terms.

We can write the error of $\hat{t}_{yU_I cal}$ as

$$\begin{aligned} \hat{t}_{yU_I cal} - t_{yU_I} \;=\; & \sum\nolimits_{r_I} \check{E}_{\theta k} - \sum\nolimits_{U_I} E_{\theta k} + \left[\sum\nolimits_{s_I} \check{\mathbf{x}}_{2k} - \sum\nolimits_{U_I} \mathbf{x}_{2k} \right]' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(2)} \\ & - \left(\sum\nolimits_{r_I} \check{\mathbf{x}}_k - \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} \right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right) \end{aligned} \tag{74}$$

where $\check{E}_{\theta k} = \check{y}_k - \check{\mathbf{x}}_k' \mathbf{B}_{\boldsymbol{\theta} U_I}$.

Evaluating the error of $\hat{t}_{yU_I cal}$ we see that the expectation of $\sum_{r_I} \check{E}_{\theta k} - \sum_{U_I} E_{\theta k}$ is $-\sum_{U_I} (1-\theta_k) E_{\theta k}$ and that of $\left[\sum_{s_I} \check{\mathbf{x}}_{2k} - \sum_{U_I} \mathbf{x}_{2k} \right]' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(2)}$ is zero. The expectation of the last term is not exactly zero, but $\left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right)$ tends to zero in probability with increasing size of the response set. Thus, $B\left(\hat{t}_{yU_I cal} \right) \approx -\sum_{U_I} (1-\theta_k) E_{\theta k}$.

A.3 Derivation of approximate bias of $\hat{t}_{yUcal}$

The derivation of the approximate bias of $\hat{t}_{yUcal}$ follow the approach from Särndal and Lundström (2005).

The calibration estimator $\hat{t}_{yUcal}$ can be written as

$$\hat{t}_{yUcal} = \sum\nolimits_{r_I} \check{y}_k + \left[\begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \widehat{\mathbf{B}}_{r_I} \tag{75}$$

where

$$\widehat{\mathbf{B}}_{r_I} = \begin{bmatrix} \widehat{\mathbf{B}}_{r_I}^{(1)} \\ \widehat{\mathbf{B}}_{r_I}^{(2)} \end{bmatrix} = \left(\sum\nolimits_{r_I} \mathbf{x}_k \check{\mathbf{x}}_k' \right)^{-1} \sum\nolimits_{r_I} \mathbf{x}_k \check{y}_k$$

For large response sets, $\widehat{\mathbf{B}}_{r_I}$ is close in probability to the vector

$$\mathbf{B}_{\boldsymbol{\theta} U_I} = \begin{bmatrix} \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \\ \mathbf{B}_{\boldsymbol{\theta} U_I}^{(2)} \end{bmatrix} = \left(\sum\nolimits_{U_I} \theta_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum\nolimits_{U_I} \theta_k \mathbf{x}_k y_k$$

Now we rewrite (75) to obtain

$$\begin{aligned} \hat{t}_{yUcal} &= \sum\nolimits_{r_I} \check{y}_k + \left[\begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \mathbf{B}_{\boldsymbol{\theta} U_I} \\ &\quad + \left[\begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} - \sum\nolimits_{r_I} \check{\mathbf{x}}_k \right]' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right) \end{aligned}$$

Note that the last term $\left(\sum_{r_I} \check{\mathbf{x}}_k - \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} \right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right)$ is of lower order of importance, compared to the preceeding two terms.

We can write the error of $\hat{t}_{yUcal}$ as

$$\begin{aligned} \hat{t}_{yUcal} - t_{yU} &= \sum\nolimits_{r_I} \check{E}_{\theta k} - \sum\nolimits_{U_I} E_{\theta k} + \left[\sum\nolimits_{s_I} \check{\mathbf{x}}_{2k} - \sum\nolimits_{U_I} \mathbf{x}_{2k} \right]' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(2)} \\ &\quad - \left(\sum\nolimits_{r_I} \check{\mathbf{x}}_k - \begin{pmatrix} \sum_{U_I} \mathbf{x}_{1k} \\ \sum_{s_I} \check{\mathbf{x}}_{2k} \end{pmatrix} \right)' \left(\widehat{\mathbf{B}}_{r_I} - \mathbf{B}_{\boldsymbol{\theta} U_I} \right) \\ &\quad + \sum\nolimits_{U_{UC}} \mathbf{x}_{1k}' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} - t_{yU_{UC}} \\ &\quad + \sum\nolimits_{U_{UC}} \mathbf{x}_{1k}' \left(\widehat{\mathbf{B}}_{r_I}^{(1)} - \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \right) \end{aligned} \tag{76}$$

where $\check{E}_{\theta k} = \check{y}_k - \check{\mathbf{x}}_k' \mathbf{B}_{\boldsymbol{\theta} U_I}$.

Thus, the error can be written as

$$\hat{t}_{yUcal} - t_{yU} = \hat{t}_{yU_Ical} - t_{yU_I} + \sum\nolimits_{U_{UC}} \mathbf{x}_{1k}' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} - t_{yU_{UC}} + \sum\nolimits_{U_{UC}} \mathbf{x}_{1k}' \left(\widehat{\mathbf{B}}_{r_I}^{(1)} - \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \right)$$

The expectation of the term $\left(\widehat{\mathbf{B}}_{r_I}^{(1)} - \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \right)$ is not exactly zero, but $\left(\widehat{\mathbf{B}}_{r_I}^{(1)} - \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} \right)$ tends to zero in probability with increasing size of the response set. Thus, $B\left(\hat{t}_{yUcal}\right) \approx -\sum_{U_I} (1 - \theta_k) E_{\theta k} + \sum_{U_{UC}} \mathbf{x}_{1k}' \mathbf{B}_{\boldsymbol{\theta} U_I}^{(1)} - t_{yU_{UC}}$

Appendix B: Monte Carlo sampling distributions

In figure B.1 to B.12 the vertical line marks $t_{yU_I} = 25\ 960\ 559$.

Figure B.1 $\hat{t}_{yU_I}^{(reg)}$ - *Case* x_1

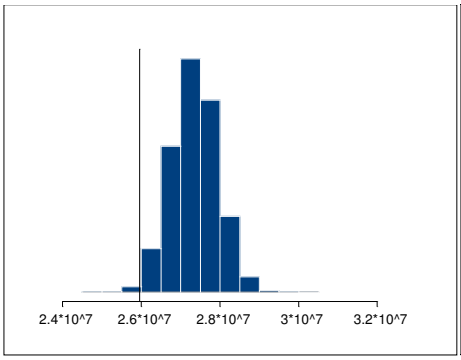


Figure B.2 $\hat{t}_{yU_I}^{(reg)}$ - *Case CS*

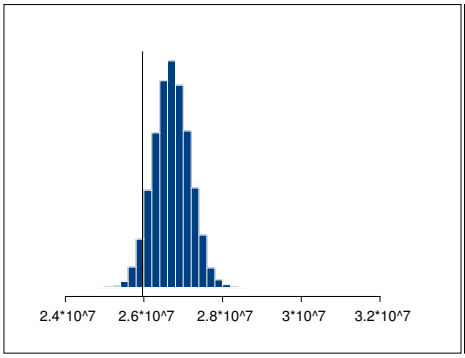


Figure B.3 $\hat{t}_{yU_I}^{(reg)}$ - *Case ME*

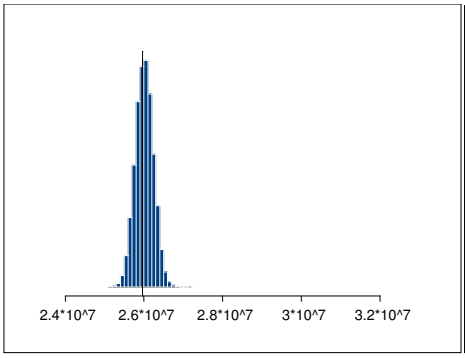


Figure B.4 $\hat{t}_{yU_I}^{(reg,RHG)}$ - *Case* x_1

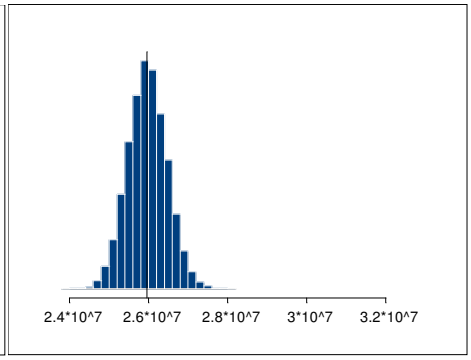


Figure B.5 $\hat{t}_{yU_I}^{(reg,RHG)}$ - *Case CS*

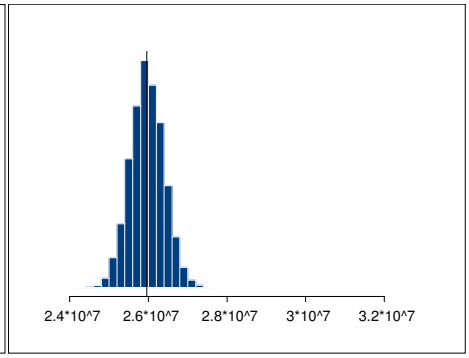


Figure B.6 $\hat{t}_{yU_I}^{(reg,RHG)}$ - *Case ME*

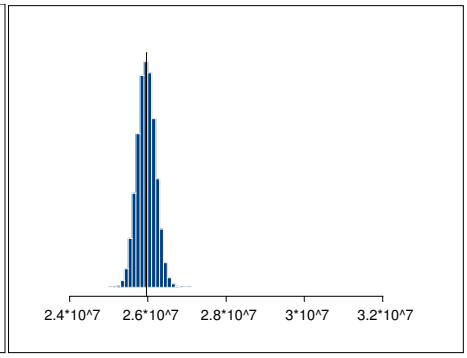


Figure B.7 $\hat{t}_{yU_I}^{(cal)}$ - *Case* x_1

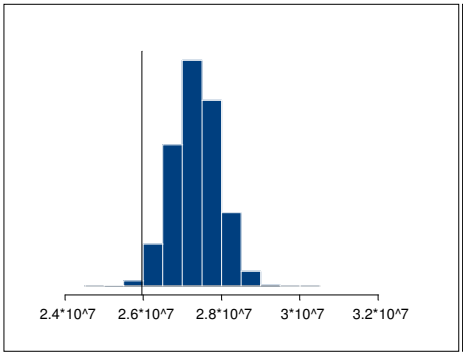


Figure B.8 $\hat{t}_{yU_I}^{(cal)}$ - *Case CS*

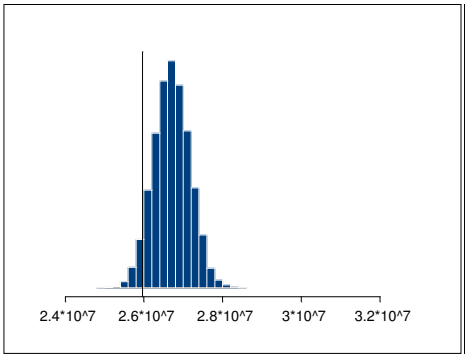


Figure B.9 $\hat{t}_{yU_I}^{(cal)}$ - *Case ME*

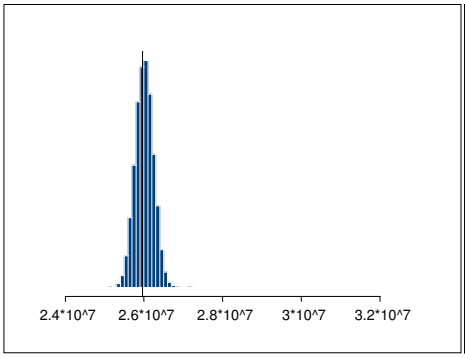


Figure B.10 $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$ - *Case* x_1

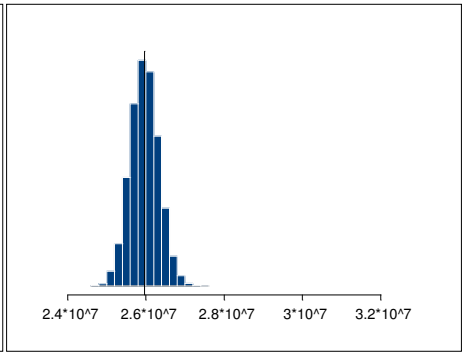


Figure B.11 $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$ - *Case CS*

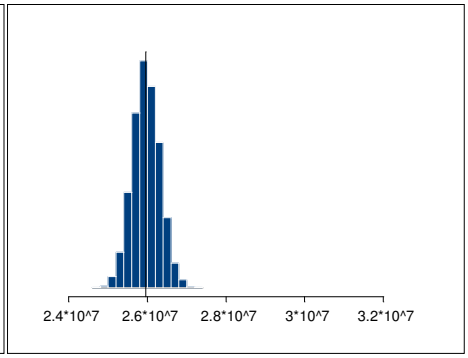
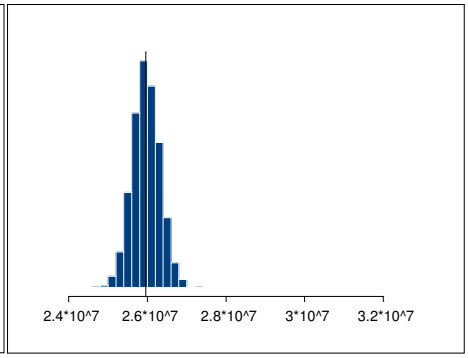


Figure B.12 $\hat{t}_{yU_I}^{(cal,\mathbf{x}_3)}$ - *Case ME*



In figure B.13 to B.24 the vertical line marks $t_{yU} = 27\ 778\ 011$.

Figure B.13 $\hat{t}_{yU}^{(reg)}$ - *Case* x_1

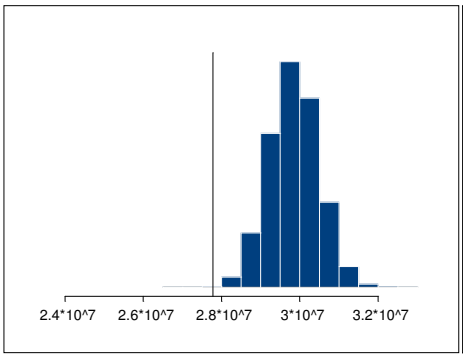


Figure B.14 $\hat{t}_{yU}^{(reg)}$ - *Case CS*

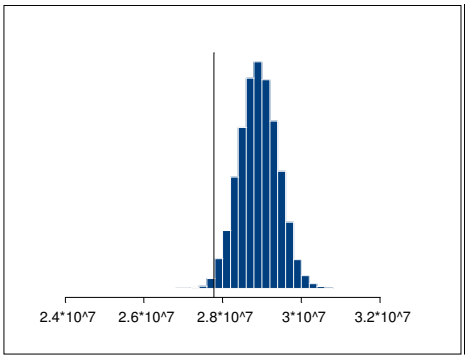


Figure B.15 $\hat{t}_{yU}^{(reg)}$ - *Case ME*

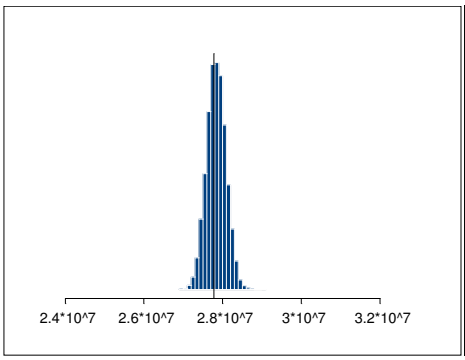


Figure B.16 $\hat{t}_{yU}^{(reg,RHG)}$ - *Case* x_1

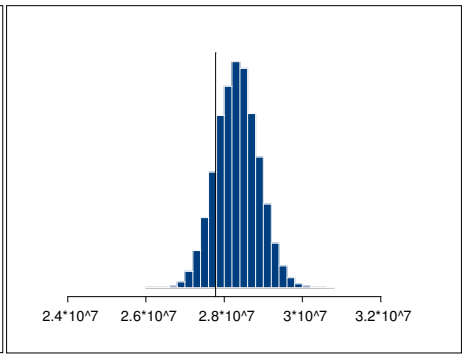


Figure B.17 $\hat{t}_{yU}^{(reg,RHG)}$ - *Case CS*

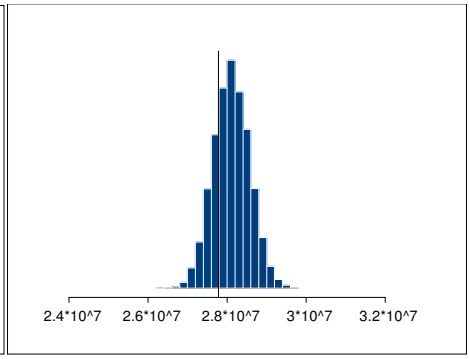


Figure B.18 $\hat{t}_{yU}^{(reg,RHG)}$ - *Case ME*

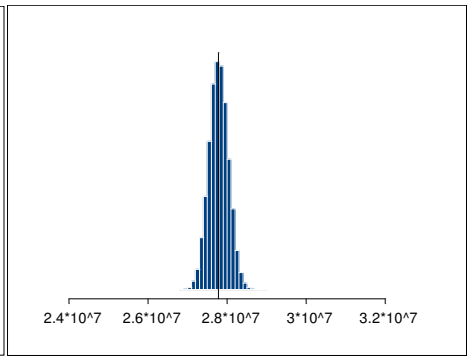


Figure B.19 $\hat{t}_{yU}^{(cal)}$ - *Case* x_1

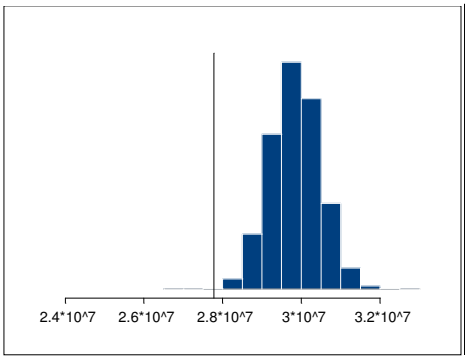


Figure B.20 $\hat{t}_{yU}^{(cal)}$ - *Case CS*

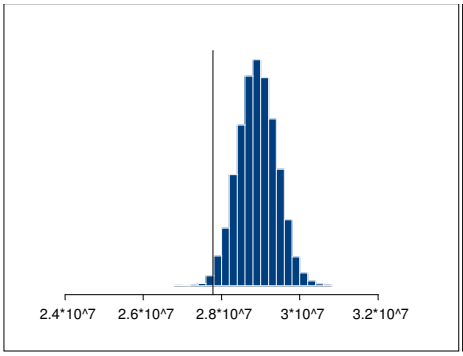


Figure B.21 $\hat{t}_{yU}^{(cal)}$ - *Case ME*

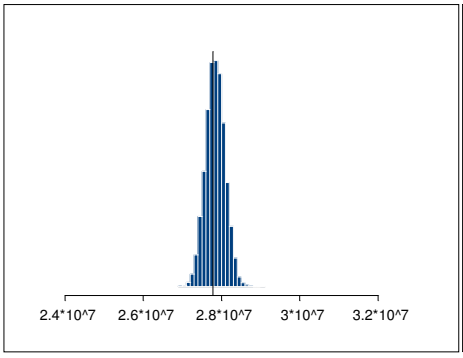


Figure B.22 $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ - *Case* x_1

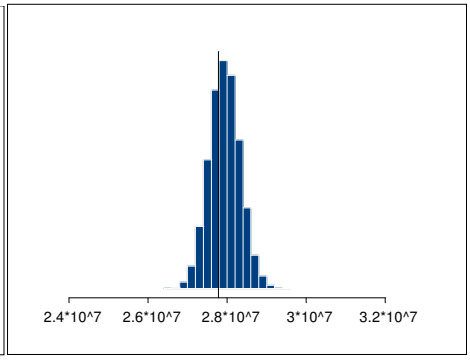


Figure B.23 $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ - *Case CS*

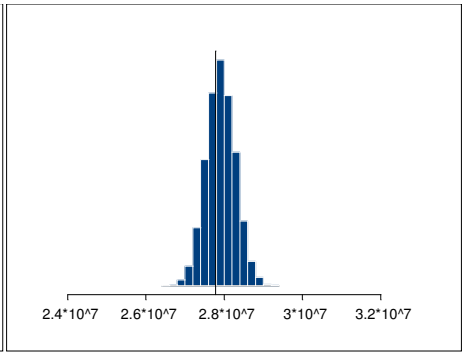


Figure B.24 $\hat{t}_{yU}^{(cal,\mathbf{x}_3)}$ - *Case ME*