WORKING PAPER SERIES

WORKING PAPER NO 3, 2004

ESI

Estimating the finite population total under Frame imperfections

by

Marianne Ängsved

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

SE-701 82 Örebro
SWEDEN

ISSN 1403-0586
ISBN 91-7668-401-6

Estimating the finite population total under frame imperfections

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 confront the problem of estimation in the face of non-negligible frame imperfections, e.g. overcoverage and undercoverage. In this paper we discuss different approaches to deal with this problem. In particular, we address the situation when there exists a new up-to-date current register and the improvement this brings along. The paper is part of a project which eventually aims at developing approaches for handling the estimation problem when nonresponse and frame imperfection occur simultaneously.

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

[a] Department of Statistics (ESI), Örebro University, SE-701 82 Örebro, Sweden.

Contents

1 The problem

1.1 Introduction

Survey errors are generally divided into two major types, sampling errors and nonsampling errors. Sampling errors stem from the fact that a sample, not the entire population, is observed. A well-developed theory exists to facilitate for the survey statistician to deal with this kind of error. Nonsampling errors encompass all other errors that contribute to survey error, i.e. frame errors, measurement errors, coding and editing errors, imputation errors and nonresponse errors.

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 is a perfect match to the target population, i.e. to the population the statistician wishes to study, and equipped with good auxiliary information. This is far from always the case, and the statistician has to accept the fact that the sampling frame is more or less imperfect with respect to matching the target population, with more or less useful auxiliary information.

There are many reasons why frame imperfection may occur. One is that it may be difficult to find an appropriate register for the target population, forcing the statistician to utilize a second-rate one. Another is difficulties in obtaining updated information for an already existing, otherwise acceptable, frame.

A particular survey setup which gives rise to a gradually deteriorating frame situation is the following. Suppose that a monthly or quarterly survey due to practical considerations is based on one and the same sample during a period of one year and that the target population changes over the year. Even if the frame is perfect for the first monthly (or quarterly) survey, this will not be the case for later months because over time some elements cease to exist, while new elements are "born". This may cause a non-ignorable amount of frame error.

In this paper we will discuss estimation when the frame suffers from imperfections. We assume that the frame enables *direct element sampling,* i.e. population elements are directly identifiable using information in the frame. The paper is part of a project which eventually aims at developing approaches for handling the estimation problem when nonresponse and frame imperfection occur simultaneously.

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 now define two different populations, viz., target population and frame population. 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.

In the ideal situation the two populations 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 has 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.

Finally, 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 An illustration

The Business Register (BR) at Statistics Sweden is frequently used as the standard frame for business surveys. The aim of the register is to contain all businesses in Sweden which are running some economic activity. Variables in the register are e.g. address, number of employees and Swedish Standard Industrial Classification (SE-SIC 92). The information in the BR is updated through different administrative sources and through surveys made by Statistics Sweden, at different times during the year.

The following table illustrates annual change of the number of businesses in the BR.

Table 1.1 *Annual change in the BR. Number of enterprises in November 2000 versus November 2001*

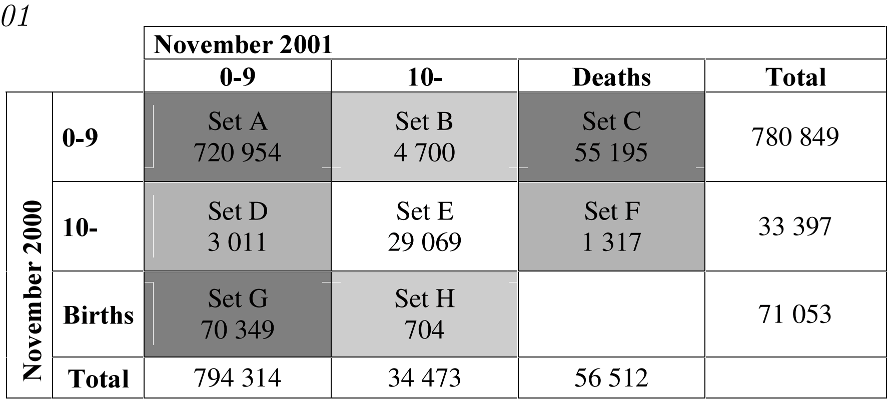
	November 2001	**Deaths**	**Total**
November 2000	757 734	56 512	814 246
Births	71 053		
Total	828 787		

As can be seen from table 1.1 there is a substantial amount of births and deaths registered in the BR during a year. Suppose a survey is to be taken at November 2001 using the BR at November 2000 as frame. This frame suffers from both undercoverage (births) and overcoverage (deaths). The undercoverage rate is 8.6% ($N_{UC}/N =$ 71 053/828 787) and the overcoverage rate is 6.9% ($N_{OC}/N_F =$ 56 512/814 246).

A variable in the BR often used in business surveys for defining cut-off limits is *size group* (number of employees). If the survey has a cut-off limit at 10 employees this means that businesses with less than 10 employees are deliberately excluded from the survey, i.e. the target population is now restricted to businesses with 10 employees or more. The analysis of the rates of change in the BR becomes even more complex adding this variable.

Suppose that a monthly or a quarterly survey were to use a sample drawn from the November 2000 version of the BR and then keep this sample for a whole year. Suppose furthermore that this survey uses a cut-off limit at 10 employees. Then, by the end of the period November 2000 to November 2001, the undercoverage consists not only of the births of enterprises with 10 or more employees but also of the enterprises that in November 2000 had less than 10 employees but in November 2001 have 10 employees or more, which the following table illustrates.

Table 1.2 *Annual change in the BR. Number of enterprises per number of employees (cut-off limit at 10 employees) in November 2000 versus November 2001*



		November 2001			
		0-9	10-	Deaths	Total
November 2000	0-9	Set A 720 954	Set B 4 700	Set C 55 195	780 849
	10-	Set D 3 011	Set E 29 069	Set F 1 317	33 397
	Births	Set G 70 349	Set H 704		71 053
	Total	794 314	34 473	56 512	

The sets A, B and C in table 1.2 are all deliberately excluded from the survey since businesses belonging to these sets have less then 10 employees at the sampling stage. However, set B consists of enterprises that has increased the number of employees during the year, and this set, by the end of the period, actually ought to be included in the survey. Thus, these 4,700 businesses belong to the undercoverage by November 2001. Sets D and F both belong to the overcoverage set, since these businesses had 10 or more employees by November 2000 and thus was included in the frame at the time of sampling, but have either decreased their staff to less then 10 employees or "died" during the period. Set E is the set of enterprises that correctly has been included in the survey. Finally, the two sets of businesses that have been "born" during the period are sets G and H. Both are excluded from the survey since these businesses would not be included in the frame at the sampling stage. However, by the end of the period the businesses in set H belong to the target population and thus, this set, together with set B, constitutes the undercoverage set. Using table 1.2 and the fact that, in this situation, $U_{OC} = D \cup F$, $U_{UC} = B \cup H$ and $U_I = E$ we get the following table, which corresponds to table 1.1.

Table 1.3 *Annual change in the BR. Number of enterprises with 10 employees or more in November 2000 versus November 2001*

	November 2001	Deaths	Total
November 2000	29 069	4 328	33 397
Births	5 404		
Total	34 473		

The use of a cutoff limit at 10 employees drastically changes both the frame population and the target population as table 1.3 shows. This also affects the size of the undercoverage and the overcoverage. The undercoverage rate is now 15.7% and the overcoverage rate is 13.0%.

The variable *size group* is also often used in business surveys for defining domains or strata. Table 1.4 shows the complexity of the births, deaths and changes of size group during the period from November 2000 to November 2001.

Table 1.4 *Annual change in the BR: Number of enterprises per size group in November 2000 versus November 2001*

		November 2001							
		0	**1-4**	**5-9**	**10-19**	**20-49**	**50-**	**Deaths**	**Total**
November 2000	**0**	532 494	20 199	1 059	293	122	56	47 114	601 337
	1-4	17 004	114 958	7 172	458	88	6	6 778	146 464
	5-9	900	5 283	21 885	3 500	166	11	1 303	33 048
	10-19	245	272	2 246	12 303	1 708	29	623	17 426
	20-49	93	60	62	879	7 811	596	409	9 910
	50-	25	4	4	15	272	5 456	285	6 061
	Births	62 383	6 979	987	315	214	175	-	71 053
	Total	613 144	147 755	33 415	17 763	10 381	6 329	56 512	

The main diagonal contains those businesses that to Statistics Sweden have reported numbers of employees in November 2000 and November 2001 that do not call for a change of size group. Excluding the row with births and the column with deaths, we see that below the main diagonal are those businesses that have moved to a smaller size group during the time period, while we above the main diagonal find those businesses that have moved to a larger size group.

Tables 1.1 to 1.4 illustrate the BR at two specific time points, November 2000 and November 2001. Obviously, events like births, deaths etc. in the real world and in the BR occur more or less continuously over the year.

Further analysis of the BR shows that most of the changes of size group that took place during the whole year between November 2000 and November 2001 seem to have occurred between March and May. Of course, this is not likely to give a fair picture of the way things happen in reality, but rather a consequence of the way information enters the BR. The major source of size is an administrative source which enters the BR in April/May every year. This is why it seems like most such changes occur in the spring. It is very likely that changes of size is a more continuous process in reality. By contrast, births and deaths seem to be registered in a more continuous process in the BR. (see table B.1.A-D in appendix)

Thus, even if the survey is taken very soon after sampling from a fresh BR version, there is still likely to be a non-ignorable amount of frame error. This is due to the fact that it always takes some time before events (births, deaths etc.) in the real world are registered in the BR.

Table 1.5 below illustrates the relative frequency distribution by size group. Suppose a survey is to be taken at November 2001 using the BR at November 2000 as frame. Thus the BR at November 2000 is the frame population and the BR at November 2001 is the target population. Also the deaths during the period constitutes the overcoverage set, U_{OC}, the births during the period constitutes the undercoverage set, U_{UC}, and finally the businesses in both the frame population and the target population constitutes the intersection set, U_I.

The table shows that the small businesses contribute with most of the changes in the BR. We can also see that the relative frequency distributions of the variable size in U_I and U are essentially equal.

Table 1.5A *Relative frequency distribution: Number of enterprises in U_F, and U_{OC}, by size group at November 2000*

	Nov 2000 (%), U_F		**Deaths during Nov 2000 – Nov 2001 (%), U_{OC}**	
0	601 337	(73.9)	47 114	(83.4)
1-4	146 464	(18.0)	6 778	(12.0)
5-9	33 048	(4.1)	1 303	(2.3)
10-19	17 426	(2.1)	623	(1.1)
20-49	9 910	(1.2)	409	(0.7)
50-	6 061	(0.7)	285	(0.5)
Total	814 246	(100)	56 512	(100)

Table 1.5B *Relative frequency distribution: Number of enterprises in U, U_{UC}, and U_I by size group at November 2001*

	Nov 2001 (%), U	**Births during Nov 2000 – Nov 2001 (%),** U_{UC}	**Intersection set (%),** U_I
0	613 144 (74.0)	62 383 (87.8)	550 761 (72.7)
1-4	147 755 (17.8)	6 979 (9.8)	140 776 (18.6)
5-9	33 415 (4.0)	987 (1.4)	32 428 (4.3)
10-19	17 763 (2.1)	315 (0.4)	17 448 (2.3)
20-49	10 381 (1.3)	214 (0.3)	10 167 (1.3)
50-	6 329 (0.8)	175 (0.2)	6 154 (0.8)
Total	828 787 (100)	71 053 (100)	757 734 (100)

Remark 1 *This section has illustrated the coverage problem when using the BR as frame for business surveys. It should be noted that the results presented in the following not only apply to business surveys, but also to other kinds of surveys.*

1.4 Notations and definitions

The structure of the frame, the information it contains, and the quality of that information will determine the type of sampling designs and estimators that can be used in a survey. If the frame contains auxiliary information this information can be used for (1) special sampling techniques, such as stratification and probability-proportional-to-size sample selections, and/or for (2) special estimation techniques, such as ratio or regression estimation. As already mentioned in section 1.1, we assume that the frame enables *direct element sampling.*

If there is a one-to-one relationship between the elements in U and the elements in U_F the frame is perfect for the target population in the sense that it will be possible to give every element in U a positive probability of inclusion in the sample to be drawn - a necessary condition for unbiased estimation.

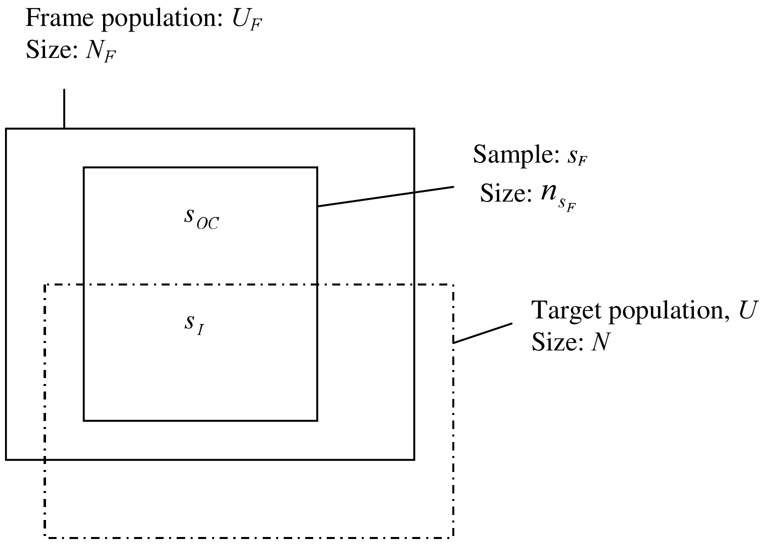
We assume that there are Q auxiliary variables in the frame. Let $\mathbf{x}_{Fk}$ denote the value of the auxiliary variable $\mathbf{x}_F$ associated with frame population element $k \in U_F$, i.e. $\mathbf{x}_F$ is a column vector of Q components. In the situation when U_F is a perfect match to U we will use $\mathbf{x}$ to denote the auxiliary variable. Associated with each element $k \in U$ is a fixed but unknown value y_k, for the study variable y.

Let s_F denote a sample of size n_{s_F} drawn from U_F. Furthermore, let $\pi_{Fk} = P(k \in s_F)$ and $\pi_{Fkl} = P(k, l \in s_F)$ denote first- and second-order inclusion probabilities. To simplify expressions derived in subsequent sections, let $\Delta_{Fkl} = \pi_{Fkl} - \pi_{Fk}\pi_{Fl}$. For a single element, let the symbol $\check{}$ symbolize division by π_{Fk}, i.e. $\check{y}_k = y_k/\pi_{Fk}$. Also, for pairs of elements, let $\check{}$ symbolize division by π_{Fkl}, i.e. $\check{\Delta}_{Fkl} = \Delta_{Fkl}/\pi_{Fkl}$.

Remark 2 *In the special case when $U_F = U$ we drop the sub-index F, i.e. denote the sample s, its size n_s, and the inclusion probabilities π_k, π_l and π_{kl} respectively.*

Figure 1 illustrates the structural relationship between a target population and an imperfect frame from which a sample is drawn.

Figure 1 *Structural relationship between a target population and an imperfect frame from which a sample is drawn*



The target population is dotted to stress the fact that both the undercoverage set and the intersection set are unknown. In this figure both undercoverage and overcoverage are present. There are two basic subsets of the selected sample s_F:

$$s_I = s_F \cap U_I \quad \begin{cases} \text{sample elements that belong to} \\ \text{the target population.} \end{cases}$$

$$s_{OC} = s_F \cap U_{OC} \quad \begin{cases} \text{sample elements that belong to} \\ \text{the overcoverage population.} \end{cases}$$

The (random) number of elements in s_I and s_{OC} are denoted n_{s_I} and $n_{s_{OC}}$ respectively.

The parameter of interest is the population total of y,

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

To simplify, we assume that all y_k are positive.

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 are able to 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 guesstimate $t_{yU_{UC}}$.

1.5 Outline of the paper

The paper has the following structure. In section 2 we give a short introduction to estimation under ideal survey conditions. The standard estimation setup in the presence of imperfect frames is discussed in section 3. In section 4 the concept of the register population and the up-to-date current register is presented and we discuss the improved estimation setup this will entail. Finally, some notes concerning future work are given in section 5.

2 Estimation under ideal survey conditions

Nonsampling errors are normal features of any survey. However, to fix ideas and to give some results that later will be extended to deal with more realistic survey setups, we briefly introduce estimation under the assumption that there are no nonsampling errors. In this ideal situation a sample s is drawn from $U_F = U$ according to a sampling design $p\,(s)$, with first- and second-order inclusion probabilities π_k and π_{kl}. We assume that there are J auxiliary varibles, denoted by $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$.

A widely used estimator in the ideal estimation situation is the generalized regression estimator $(GREG)$, which is introduced in section 2.1.

We will also introduce estimation for domains in section 2.2.

2.1 The generalized regression estimator

The generalized regression estimator $(GREG)$ is defined as

$$\hat{t}_{yUgreg} = \sum\nolimits_s \check{y}_k + \left(\sum\nolimits_U \mathbf{x}_k - \sum\nolimits_s \check{\mathbf{x}}_k \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}} \tag{1}$$

where

$$\begin{aligned} \widehat{\mathbf{B}} &= \left(\widehat{B}_1, \ldots, \widehat{B}_Q\right)' &(2)\\ &= \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s}\widehat{\mathbf{t}}_{\mathbf{x}ys} &(3)\end{aligned}$$

with $\widehat{\mathbf{T}}_{\mathbf{xx}s} = \sum_s c_k \mathbf{x}_k \check{\mathbf{x}}_k'$ and $\widehat{\mathbf{t}}_{\mathbf{x}ys} = \sum_s c_k \mathbf{x}_k \check{y}_k$. The factor c_k is a suitably chosen weight assigned to all $k \in U$. In a model assisted approach it can for example be chosen as $c_k = 1/\sigma_k^2$, where σ_k^2 expresses the statistician's best opinion of the residual variability of y in a linear relationship with $\mathbf{x}_k$. For details, see Särndal, Swensson, and Wretman (1992).

We can view $\hat{t}_{yUgreg}$ as an attempt to improve over the basic π estimator $\sum_s \check{y}_k = \hat{t}_{yU\pi}$. Explicitly, the regression estimator is equal to the π estimator plus an adjustment term.

Remark 3 *The* GREG *can be expressed in terms of g-weights,*

$$\hat{t}_{yUgreg} = \sum\nolimits_s d_k g_{ks} y_k \tag{4}$$

where $d_k = 1/\pi_k$ *and* $g_{ks} = 1 + (\sum_U \mathbf{x}_k - \sum_s \check{\mathbf{x}}_k)' \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s} c_k \mathbf{x}_k$.

The *GREG* can be seen as a special case of calibration. Calibration is a technique that, with starting point from the basic π estimator $\hat{t}_{yU\pi} = \sum_s d_k y_k$, creates a new estimator $\hat{t}_{yUcal} = \sum_s w_k y_k$. This estimator has new weights w_k that lie as close as possible to the original sampling weights $d_k = 1/\pi_k$, subject to the calibration constraint $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. When minimizing the Generalized Least Squares (GLS) distance function

$$\sum\nolimits_s c_k \left(w_k - d_k\right)^2 / d_k,$$

the calibrated weights are given by $w_k = d_k g_{ks}$ (see Deville and Särndal, 1992). Thus we may express the calibration estimator as

$$\begin{aligned} \hat{t}_{yUcal} &= \sum\nolimits_s w_k y_k \\ &= \sum\nolimits_s \check{y}_k + \left(\sum\nolimits_U \mathbf{x}_k - \sum\nolimits_s \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}} \end{aligned} \tag{5}$$

Hence, using the same c_k in $\hat{t}_{yUgreg}$ and $\hat{t}_{yUcal}$, the two estimators coincide.

Using first order Taylor linearization, it can be shown that $\hat{t}_{yUgreg}$ is approximately unbiased for $t_{yU} = \sum_U y_k$ with the approximate variance given by

$$AV\left(\hat{t}_{yUgreg}\right) = \sum\sum\nolimits_U \Delta_{kl} \check{E}_k \check{E}_l \tag{6}$$

where $\check{E}_k = E_k/\pi_k = \left(y_k - \mathbf{x}_k'\mathbf{B}\right)/\pi_k$ and where

$$\mathbf{B} = \mathbf{T}_{\mathbf{xx}U}^{-1} \mathbf{t}_{\mathbf{x}yU} = \left(\sum\nolimits_U c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum\nolimits_U c_k \mathbf{x}_k y_k.$$

(For a detailed proof, see Särndal et al., 1992) A variance estimator is given by

$$\hat{V}\left(\hat{t}_{yUgreg}\right) = \sum\sum\nolimits_s \check{\Delta}_{kl} \left(g_{ks}\check{e}_{ks}\right)\left(g_{ls}\check{e}_{ls}\right) \tag{7}$$

with $\check{e}_{ks} = e_{ks}/\pi_k = \left(y_k - \mathbf{x}_k'\widehat{\mathbf{B}}\right)/\pi_k$.

2.2 Estimation for domains

Consider a partitioning of the population U into D domains, denoted $U_1, \ldots, U_d, \ldots, U_D$. Let N_d be the size of U_d. As before a sample s is drawn from U. Let $s_d = s \cap U_d$ denote the part of s that happens to fall in U_d. Also let

$$y_{dk} = \begin{cases} y_k & \text{if } k \in U_d \\ 0 & \text{otherwise} \end{cases}$$

The objective is to estimate the domain totals $t_d = \sum_U y_{dk} = \sum_{U_d} y_k$, $d = 1, \ldots, D$.

2.2.1 Basic estimators

A simple estimator for the domain total when N_d is unknown is the domain π estimator

$$\hat{t}_{yU_d\pi} = \sum\nolimits_s \check{y}_{dk} = \sum\nolimits_{s_d} \check{y}_k = \sum\nolimits_{s_d} y_k/\pi_k \tag{8}$$

If the domain set U_d is small the estimator will have poor precision.

When N_d is known, an alternative estimator is

$$\tilde{t}_{yU_d} = N_d \tilde{y}_{s_d} = N_d \hat{t}_{yU_d\pi}/\hat{N}_d \tag{9}$$

where $\hat{N}_d = \sum_{s_d} 1/\pi_k$. This estimator would be preferable to $\hat{t}_{yU_d\pi}$ since the variance of $\tilde{t}_{yU_I}$ ordinarily is smaller (see Särndal et al., 1992, p. 391).

2.2.2 Regression estimators

When auxiliary information is available improved domain estimators may be obtained by using the regression approach.

Four alternative regression estimators for domains are as follows:

(i)

$$\hat{t}^{(i)}_{yU_dreg} = \frac{N_d}{\hat{N}_d} \sum\nolimits_{s_d} \check{y}_k + \left(\sum\nolimits_{U_d} \mathbf{x}_k - \frac{N_d}{\hat{N}_d} \sum\nolimits_{s_d} \check{\mathbf{x}}_k \right)' \widehat{\mathbf{B}}_s \tag{10}$$

where $\hat{N}_d = \sum_{s_d} 1/\pi_k$, N_d is known and

$$\begin{aligned} \widehat{\mathbf{B}}_s &= \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s} \widehat{\mathbf{t}}_{\mathbf{x}ys} \\ &= \left(\sum\nolimits_s c_k \mathbf{x}_k \check{\mathbf{x}}'_k \right)^{-1} \sum\nolimits_s c_k \mathbf{x}_k \check{y}_k \end{aligned}$$

(ii)

$$\hat{t}_{yU_d reg}^{(ii)} = \sum\nolimits_{s_d} \check{y}_k + \left(\sum\nolimits_{U_d} \mathbf{x}_k - \sum\nolimits_{s_d} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_s \tag{11}$$

(iii)

$$\hat{t}_{yU_d reg}^{(iii)} = \sum\nolimits_{s_d} \check{y}_k + \left(\sum\nolimits_{U} \mathbf{x}_k - \sum\nolimits_{s} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{s,s_d} \tag{12}$$

where

$$\begin{aligned} \widehat{\mathbf{B}}_{s,s_d} &= \widehat{\mathbf{T}}_{\mathbf{xx}s}^{-1} \widehat{\mathbf{t}}_{\mathbf{x}ys_d} \\ &= \left(\sum\nolimits_{s} c_k \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \sum\nolimits_{s_d} c_k \mathbf{x}_k \check{y}_k \end{aligned}$$

(iv)

$$\hat{t}_{yU_d reg}^{(iv)} = \sum\nolimits_{s_d} \check{y}_k + \left(\sum\nolimits_{U_d} \mathbf{x}_k - \sum\nolimits_{s_d} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{s_d} \tag{13}$$

where

$$\begin{aligned} \widehat{\mathbf{B}}_{s_d} &= \widehat{\mathbf{T}}_{\mathbf{xx}s_d}^{-1} \widehat{\mathbf{t}}_{\mathbf{x}ys_d} \\ &= \left(\sum\nolimits_{s_d} c_k \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \sum\nolimits_{s_d} c_k \mathbf{x}_k \check{y}_k \end{aligned} \tag{14}$$

The estimators (i) and (ii) are suggested in Särndal et al. (1992). Both estimators require known auxiliary totals $\sum_{U_d} \mathbf{x}_k$ for the domain. If $\sum_{s_d} \left(y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_s\right)$ is identically equal to zero the estimators (i) and (ii) agree. If not, then (i) is usually preferred to (ii) when N_d is known. The reason is that the size of s_d is random and the term $\left(N_d/\hat{N}_d\right) \sum_{s_d} \left(y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_s\right)/\pi_k$ in (i) tends to be less variable than the corresponding term $\sum_{s_d} \left(y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_s\right)/\pi_k$ in (ii). However, if a main objective is additivity, i.e. the estimates in different subpopulations should add up to the estimate made for the population as a whole, then (ii) should be used despite some loss in efficiency compared to (i).

Estimators (iii) and (iv) are discussed in Estevao, Hidiroglou, and Särndal (1995). The domain estimator (iii) will produce little or no gain in precision due to regression. The reason for this is that the fit of the regression of the domain variable y_{dk} on $\mathbf{x}$ through the model ξ

$$y_{dk} = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k \qquad \text{for} \quad k \in U \tag{15}$$

where $\mathrm{E}_{\xi}\left(\varepsilon_k\right) = 0$, $\mathrm{Var}_{\xi}\left(\varepsilon_k\right) = c_k \sigma^2$ and $\mathrm{Cov}_{\xi}\left(\varepsilon_k, \varepsilon_l\right) = 0$ for all $k \neq l$, will sometimes be mediocre because of the special nature of y_{dk}, which equals y_k inside the domain but is always equal to zero outside. The important properties of (iii) is that: (1) the g-factors produce additive domain estimates, and (2) the g-factors are unchanged from one domain to another. The estimator (iv) requires known auxiliary totals for the domain itself. It also requires that each domain should contain enough observations to avoid unstable slope estimates $\widehat{\mathbf{B}}_{s_d}$.

Remark 4 *A variant of* $\hat{t}^{(iv)}_{yU_dreg}$ *in eq (13) would be*

$$\hat{t}^{(iv,alt)}_{yU_dreg} = \frac{N_d}{\hat{N}_d} \sum\nolimits_{s_d} \check{y}_k + \left(\sum\nolimits_{U_d} \mathbf{x}_k - \frac{N_d}{\hat{N}_d} \sum\nolimits_{s_d} \check{\mathbf{x}}_k \right)' \widehat{\mathbf{B}}_{s_d} \tag{16}$$

which would be less variable than (13). However, if the domain is very large the possible gain in efficiency will be modest.

3 Imperfect frames - standard estimation setup

In the presence of frame error, i.e. when $U_F \neq U$, a sample s_F is drawn from the frame according to a sampling design $p(s_F)$, with first- and second-order inclusion probabilities π_{Fk} and π_{Fkl} respectively. The frame contains an auxiliary vector, i.e. we have for every $k \in U_F$:

$$\mathbf{x}_{Fk} = (x_{F1k}, \ldots, x_{Fqk}, \ldots x_{FQk})'$$

Furthermore, we have the known frame population total

$$\mathbf{t}_{x_F} = \left(t_{x_{F1}}, \ldots, t_{x_{Fq}}, \ldots, t_{x_{FQ}}\right)'$$

where $t_{x_{Fq}} = \sum_{U_F} x_{Fqk}$.

Using this setup we will *at most* have information on target population membership/non-membership for every $k \in s_F$ at the estimation stage of the survey. Furthermore, the extent of the undercoverage will be unknown. In this section we will illustrate two different cases that might occur in the presence of frame error, (i) information on target population membership/non-membership is *unknown* for every $k \in s_F$, and (ii) information on target population membership/non-membership is *known* for every $k \in s_F$.

Firstly suppose that information on target population membership/non-membership is missing for every $k \in s_F$. Thus, it is not possible to identify the two subsets of s_F, i.e. s_I and s_{OC}.

Assume that there exists a y-value for every $k \in U_F$, i.e. for every $k \in U_I$ as well as for every $k \in U_{OC}$. If so, a y-value will exist for every $k \in s_F$. Since information on target population membership/non-membership is missing for every $k \in s_F$ the statistician is apt to use an estimator $\hat{t}_{yU_F}$ for t_{yU_F}. If $E\left(\hat{t}_{yU_F}\right) = t_{yU_F}$ the unknown bias of $\hat{t}_{yU_F}$ with respect to t_{yU} is given by

$$\begin{aligned} B\left(\hat{t}_{yU_F}\right) &= t_{yU_F} - t_{yU} \\ &= \left(t_{yU_I} + t_{yU_{OC}}\right) - \left(t_{yU_I} + t_{yU_{UC}}\right) \\ &= t_{yU_{OC}} - t_{yU_{UC}} \\ &= N_{OC}\,\bar{y}_{OC} - N_{UC}\,\bar{y}_{UC} \qquad (17) \end{aligned}$$

Obviously, if the difference between the mean values of y in U_{OC} and U_{UC} is modest, the size of the bias depends on (1) the difference of the number of elements in the overcoverage set and undercoverage set and (2) the magnitude of the mean values $\bar{y}_{OC} \approx \bar{y}_{UC}$.

Remark 5 *The assumption that we have information on y_k for every $k \in s_F$ may be somewhat unrealistic. The situation when the value on y_k is missing for some $k \in s_{OC}$, due to the fact that element k has ceased to exist, is more likely to appear.*

The situation changes if information on target population membership/non-membership is available for every $k \in s_F$. This information may be at hand if the sample data give information on target population membership for every sample element. In this situation we assume that information on y_k is at hand at least for every $k \in s_I$. There may be y_k-values available for some $k \in s_{OC}$, but not for all, since some elements may have ceased to exist.

The given information makes it possible to identify the subsets s_I and s_{OC}, and thus n_{s_I} and $n_{s_{OC}}$ are known.

Furthermore, we have access to the sample totals

$$\sum\nolimits_{s_I} y_k \text{ and } \sum\nolimits_{s_I} \mathbf{x}_{Fk}$$

In the described situation we are able to use the methods of domain estimation in order to estimate t_{yU_I}. However, using an (at least approximately) unbiased estimator $\hat{t}_{yU_I}$ of t_{yU_I} in order to estimate t_{yU} will lead to negative bias since $y_k > 0$. The bias and relative bias are given by

$$B(\hat{t}_{yU_I}) = -t_{yU_{UC}} \tag{18}$$

and

$$RB(\hat{t}_{yU_I}) = -\frac{t_{yU_{UC}}}{t_{yU}} = -\frac{1}{1 + t_{yU_I}/t_{yU_{UC}}} \tag{19}$$

respectively. Hence, the bias will be substantial unless $t_{yU_{UC}}$ is very small compared to t_{yU_I}, and we ought to find a way to compensate for this presumed bias. Realizing that no information on y_k exist for the elements $k \in U_{UC}$ we have to depend on more or less speculative approaches in order to adjust for undercoverage.

Example 3.1 As a simple illustration of the possible size of the relative bias, let us first rewrite the relative bias as

$$RB\left(\hat{t}_{yU_I}\right) = -\frac{1}{1 + \dfrac{N_I}{N_{UC}} \dfrac{\bar{y}_{U_I}}{\bar{y}_{U_{UC}}}}$$

If $\bar{y}_{U_I} = \bar{y}_{U_{UC}}$ the relative bias is given by

$$-\frac{N_{UC}}{N}$$

Using data from table 1.1 in section 1.3 we get $RB(\hat{t}_{yU_I}) = -\dfrac{71053}{828787} \approx -0.086$, i.e. the use of $\hat{t}_{yU_I}$ will in this case lead to a negative relative bias of 8.6 %. Furthermore, if $\bar{y}_{U_I} > \bar{y}_{U_{UC}}$, as table 1.5B indicates, the absolute relative bias might be even larger. □

Under the assumptions that (1) information on target population membership/non-membership is available for every $k \in s_F$ and (2) information on y_k is at hand for every $k \in s_I$, we will first present two approaches for estimating t_{yU_I} as a basis for the more difficult problem of estimating t_{yU}.

3.1 Estimation of t_{yU_I}

Using techniques for estimation of a domain total, the most simple estimator (corresponding to $\hat{t}_{yU_d\pi}$) for estimating the total of U_I is the domain π estimator

$$\hat{t}_{yU_I\pi} = \sum\nolimits_{s_I} \check{y}_k = \sum\nolimits_{s_I} y_k/\pi_{Fk} \tag{20}$$

An alternative estimator, when N_I is known, is

$$\tilde{t}_{yU_I} = N_I \tilde{y}_{s_I} = N_I \hat{t}_{yU_I\pi}/\hat{N}_I \tag{21}$$

where $\hat{N}_I = \sum_{s_I} 1/\pi_{Fk}$. This estimator would be preferable to $\hat{t}_{yU_I\pi}$ since the variance of $\tilde{t}_{yU_I}$ ordinarily is smaller than that of $\hat{t}_{yU_I\pi}$. But since N_I is unknown, $\tilde{t}_{yU_I}$ is inapplicable in this case.

Remark 6 *As stated in section 2.2.1 the estimator $\hat{t}_{yU_I\pi}$ will have poor precision if the "domain" U_I is small. However, this will not be the case in the present context, since it would mean that the frame would be useless for the survey.*

We now turn to regression estimators for t_{yU_I}. Recall that in section 2.2.2 four alternative regression estimators for domains proposed by different authors were given. We will, under the assumptions made in this setup, examine whether any of these four estimators are possible candidates for estimation of t_{yU_I}.

In order to facilitate the discussion in this and following sections we begin by rewriting these estimators using a notation that better fits the situation where the domain of interest is U_I of size N_I and where the auxiliary vector is $\mathbf{x}_F$. This gives the following expressions:

(i)

$$\hat{t}^{(i)}_{yU_Ireg} = \frac{N_I}{\hat{N}_I}\sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{Fk} - \frac{N_I}{\hat{N}_I}\sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_F} \tag{22}$$

where N_I is known, $\hat{N}_I = \sum_{s_I} 1/\pi_{Fk}$, and

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}_F s_F} &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_F s_F}\widehat{\mathbf{t}}_{\mathbf{x}_F y s_F} \\ &= \left(\sum\nolimits_{s_F} c_k \mathbf{x}_{Fk}\check{\mathbf{x}}'_{Fk}\right)^{-1} \sum\nolimits_{s_F} c_k \mathbf{x}_{Fk}\check{y}_k \end{aligned} \tag{23}$$

is the estimated regression coefficient in U_F.

(ii)

$$\hat{t}^{(ii)}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{Fk} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_F} \tag{24}$$

(iii)

$$\hat{t}^{(iii)}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_F} \mathbf{x}_{Fk} - \sum\nolimits_{s_F} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_F, s_I} \tag{25}$$

where

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}_F s_F, s_I} &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_F s_F}\widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} \\ &= \left(\sum\nolimits_{s_F} c_k \mathbf{x}_{Fk}\check{\mathbf{x}}'_{Fk}\right)^{-1} \sum\nolimits_{s_I} c_k \mathbf{x}_{Fk}\check{y}_k \end{aligned} \tag{26}$$

(iv)

$$\hat{t}^{(iv)}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{Fk} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{27}$$

where

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_F s_I}\widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} \\ &= \left(\sum\nolimits_{s_I} c_k \mathbf{x}_{Fk}\check{\mathbf{x}}'_{Fk}\right)^{-1} \sum\nolimits_{s_I} c_k \mathbf{x}_{Fk}\check{y}_k \end{aligned} \tag{28}$$

is the estimated regression coefficient in U_I.

Remark 7 *We use the $\widehat{\mathbf{B}}$ index $\mathbf{x}_F$ to underline the difference between the regression coefficient under the standard estimation setup and under the ideal situation.*

It is immediately clear that none of the first two estimators can be used, since they both require knowledge of the domain total of the auxiliary vector $\sum_{U_I} \mathbf{x}_{Fk}$ and since their common $\widehat{\mathbf{B}}_{\mathbf{x}_F s_F}$ requires y_k-values for every $k \in s_F$. The first estimator also requires knowledge of N_I. None of these requirements are fulfilled.

The third estimator is a potential candidate, since it is based on available y information only. However, for reasons given in section 3.2, it will only produce little or no gain in precision due to regression, and hence we do not see this estimator as a useful alternative to the simple domain π estimator, $\hat{t}_{yU_I\pi}$.

Hence, the only estimator that might be a potential candidate is the fourth estimator. However, like the first two estimators, it requires knowledge that is not at hand, although not to the same extent. The only information we now lack is the value of $\sum_{U_I} \mathbf{x}_{Fk} = \mathbf{t}_{x_F U_I}$. This suggests an estimator that might be used in some surveys.

Suppose that we, using external information, can come up with a reasonably close approximation to $\mathbf{t}_{x_F U_I}$, say $\tilde{\mathbf{t}}_{x_F U_I}$. Using this approximation instead of $\sum_{U_I} \mathbf{x}_{Fk} = \mathbf{t}_{x_F U_I}$ in the fourth estimator gives the following alternative estimator for t_{yU_I}:

$$\tilde{t}_{yU_I reg} = \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{29}$$

Remark 8 *An alternative expression of $\tilde{t}_{yU_I reg}$ is*

$$\tilde{t}_{yU_I reg} = \sum\nolimits_{s_I} \tilde{w}_{Ik} y_k = \sum\nolimits_{s_I} \tilde{v}_{Ik} \check{y}_k \tag{30}$$

where $\tilde{w}_{Ik} = \tilde{v}_{Ik}/\pi_{Fk}$ and

$$\tilde{v}_{Ik} = 1 + c_k \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \left(\sum\nolimits_{s_I} c_k \mathbf{x}_{Fk} \check{\mathbf{x}}'_{Fk}\right)^{-1} \mathbf{x}_{Fk} \tag{31}$$

The expected value of $\tilde{t}_{yU_I reg}$ is given by (for a proof, using Taylor linearization, see appendix A.1)

$$E\left(\tilde{t}_{yU_I reg}\right) \doteq t_{yU_I} + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} \tag{32}$$

where $\mathbf{B}_{\mathbf{x}_F U_I} = \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} \mathbf{t}_{\mathbf{x}_F y U_I} = \left(\sum_{U_I} c_k \mathbf{x}_{Fk} \mathbf{x}'_{Fk} \right)^{-1} \sum_{U_I} c_k \mathbf{x}_{Fk} y_k$. Hence, its bias is given by

$$B\left(\tilde{t}_{yU_I reg}\right) \doteq \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} \tag{33}$$

The approximate variance of $\tilde{t}_{yU_I reg}$ is given by (again, for a proof, see appendix A.1)

$$AV\left(\tilde{t}_{yU_I reg}\right) = \sum\sum\nolimits_{U_I} \Delta_{Fkl} \check{E}^{\alpha_I}_{Fk} \check{E}^{\alpha_I}_{Fl} \tag{34}$$

where $\Delta_{Fkl} = \pi_{Fkl} - \pi_{Fk}\pi_{Fl}$ and

$$E^{\alpha_I}_{Fk} = \left(1 + \boldsymbol{\alpha}'_I c_k \mathbf{x}_{Fk}\right) E_{Fk}$$

with $\boldsymbol{\alpha}'_I = \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I}$ and $E_{Fk} = y_k - \mathbf{x}'_{Fk} \mathbf{B}_{\mathbf{x}_F U_I}$.

A variance estimator would be

$$\hat{V}\left(\tilde{t}_{yU_I reg}\right) = \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl} \check{e}^{\hat{\alpha}_I}_{Fk} \check{e}^{\hat{\alpha}_I}_{Fl} \tag{35}$$

where $e^{\hat{\alpha}_I}_{Fk} = \left(1 + \hat{\boldsymbol{\alpha}}'_I c_k \mathbf{x}_{Fk}\right) e_{Fk}$ with $\hat{\boldsymbol{\alpha}}'_I = \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I}$ and $e_{Fk} = y_k - \mathbf{x}'_{Fk} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I}$.

Remark 9 *An alternative to $e^{\hat{\alpha}_I}_{Fk}$ would be $e^{\alpha_I}_{Fk}$. At present it is not clear which choice is the better and some further work is needed in deciding whether to use $\boldsymbol{\alpha}'_I$ or $\hat{\boldsymbol{\alpha}}'_I$ in the variance estimator.*

Now, even if we have managed to come up with an estimator $\tilde{t}_{yU_I reg}$ that in some circumstances might be good for t_{yU_I} (if $\tilde{\mathbf{t}}_{x_F U_I}$ is a close approximation to $\mathbf{t}_{x_F U_I}$ and if there is a strong linear relationship between y and $\mathbf{x}_F$) we are not finished, since what we are looking for is a good estimator for $t_{yU} = t_{yU_I} + t_{yU_{UC}}$. Hence, using $\tilde{t}_{yU_I reg}$ for t_{yU} is likely to lead to underestimation (since $y_k > 0$), which will be substantial unless $t_{yU_{UC}}$ is very small compared to t_{yU_I}. Since we have no information on the elements $k \in U_{UC}$, neither from the frame, nor from the sample, we have to resort to more or less speculative approaches. For example, looking at the estimator $\tilde{t}_{yU_I reg}$ above as an estimator for $t_{yU} = t_{yU_I} + t_{yU_{UC}}$, its bias is approximately given by

$$\left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} - t_{yU_{UC}}$$

and, hence, we should not try to find an approximation $\tilde{\mathbf{t}}_{x_F U_I}$ close to $\mathbf{t}_{x_F U_I}$. Instead we should try to find a vector $\tilde{\mathbf{t}}$ such that

$$\left(\tilde{\mathbf{t}} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} \approx t_{y U_{UC}}$$

If such a vector $\tilde{\mathbf{t}}$ could be found, we might as an estimator for t_{yU} take

$$\tilde{t}_{yU} = \sum\nolimits_{s_I} \breve{y}_k + \left(\tilde{\mathbf{t}} - \sum\nolimits_{s_I} \breve{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{36}$$

In the following section we will discuss two such more or less speculative approaches.

3.2 Estimation of t_{yU}

3.2.1 A simplistic approach

One simple approach in adjusting for the undercoverage is based on the following reasoning. Suppose hypothetically that we know the relation between the population total t_{yU} and the domain total t_{yU_I} and that it may be expressed as $t_{yU} = \delta_y t_{yU_I}$, where δ_y is a known constant.

Remark 10 *Since $t_{yU} = t_{yU_I} + t_{yU_{UC}}$ we can alternatively express this relation as $t_{yU_{UC}} = (\delta_y - 1)\, t_{yU_I}$.*

In this hypothetical situation we could as an estimator for t_{yU} use $\delta_y \hat{t}_{yU_I}$, where $\hat{t}_{yU_I}$ is an estimator for t_{yU_I}. Obviously, since δ_y is unknown, it is not possible to use $\delta_y \hat{t}_{yU_I}$. But suppose that the statistician, using prior information, has access to a good approximation of δ_y, say $\tilde{\delta}_y$. This would make it possible to use the estimator

$$\hat{t}_{yU\tilde{\delta}_y} = \tilde{\delta}_y \hat{t}_{yU_I} \tag{37}$$

If $\hat{t}_{yU_I}$ is (approximately) unbiased for the total t_{yU_I}, the (approximate) relative bias of $\hat{t}_{yU\tilde{\delta}_y}$ is

$$\begin{aligned} RB\left(\hat{t}_{yU\tilde{\delta}_y}\right) &= \frac{\tilde{\delta}_y t_{yU_I} - t_{yU}}{t_{yU}} = \frac{\frac{\tilde{\delta}_y}{\delta_y}\delta_y t_{yU_I} - t_{yU}}{t_{yU}} \\ &= \frac{\tilde{\delta}_y}{\delta_y} - 1 \end{aligned} \tag{38}$$

while the approximate variance is given by

$$AV\left(\hat{t}_{yU\tilde{\delta}_y}\right) = \tilde{\delta}_y^2 AV\left(\hat{t}_{yU_I}\right) \tag{39}$$

where $AV\left(\hat{t}_{yU_I}\right)$ depends on the choice of $\hat{t}_{yU_I}$.

Any of the estimators proposed above for t_{yU_I} are possible to use in $\hat{t}_{yU\tilde{\delta}_y}$. However, note that $\tilde{t}_{yU_Ireg}$ will be biased for t_{yU_I} if $\tilde{\mathbf{t}}_{x_FU_I}$ is not close to $\mathbf{t}_{x_FU_I}$, leading to a more complex expression for the relative bias in (38).

3.2.2 A more elaborate approach

Recall from the end of section 3.1 that with a suitable choice of the vector $\tilde{\mathbf{t}}$ in (36), i.e. in

$$\tilde{t}_{yU} = \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I}$$

this estimator will be a good candidate estimator for t_{yU}. Could such a vector be found? One way of reasoning for finding an answer is as follows.

1. The estimator $\tilde{t}_{yU_Ireg}$ is partly based on the approximate linear relation

   $$y_k \approx \mathbf{x}'_{Fk}\mathbf{B}_{x_FU_I} \quad \text{for} \quad k \in U_I$$

   Suppose, hypothetically, that the same type of auxiliary $\mathbf{x}_F$-information as the one used for $\tilde{t}_{yU_Ireg}$ were available for every $k \in U_{UC}$ as well, and let

   $$y_k \approx \mathbf{x}'_{Fk}\mathbf{B}_{x_FU_{UC}} \quad \text{for} \quad k \in U_{UC}$$

2. Furthermore suppose, still hypothetically, that $\mathbf{B}_{x_FU_{UC}} = \mathbf{B}_{x_FU_I}$, i.e. that

   $$y_k \approx \mathbf{x}'_{Fk}\mathbf{B}_{x_FU_I} \quad \text{for} \quad k \in U_{UC}$$

If both these, most hypothetical, assumptions were correct, it would be reasonable to use $\hat{y}_k = \mathbf{x}'_{Fk}\widehat{\mathbf{B}}_{\mathbf{x}_F s_I}$ $(k \in U_{UC})$ as predictions for the unknown y_k-values. Hence, as an estimator for $t_{yU_{UC}} = \sum_{U_{UC}} y_k$, we might use

$$\begin{aligned} \hat{t}_{yU_{UC}} &= \sum\nolimits_{U_{UC}} \hat{y}_k = \sum\nolimits_{U_{UC}} \mathbf{x}'_{Fk}\widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \\ &= \mathbf{t}'_{x_FU_{UC}}\widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \end{aligned} \tag{40}$$

Now, unfortunately, we do not have access to $\mathbf{t}_{x_F U_C}$, and hence $\hat{t}_{yU_C}$ cannot be used.

3. However, suppose that we, again using some external data, can find a close approximation to $\mathbf{t}_{x_F U_C}$, say $\tilde{\mathbf{t}}_{x_F U_C}$. If so, a possible estimator for t_{yU_C} would be

$$\tilde{t}_{yU_C} = \tilde{\mathbf{t}}'_{x_F U_C} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{41}$$

The above reasoning, finally, suggests the following estimator for t_{yU}

$$\begin{aligned}
\tilde{t}_{yUreg} &= \tilde{t}_{yU_I reg} + \tilde{t}_{yU_C} \\
&= \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} + \tilde{\mathbf{t}}'_{x_F U_C} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \\
&= \sum\nolimits_{s_I} \check{y}_k + \left(\left(\tilde{\mathbf{t}}_{x_F U_I} + \tilde{\mathbf{t}}_{x_F U_C}\right) - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \\
&= \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_F U} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \qquad (42)
\end{aligned}$$

where $\tilde{\mathbf{t}}_{x_F U}$ is an approximation, using external information, to the fictitious $\mathbf{t}_{x_F U}$.

Remark 11 *An alternative way of reasoning about the hypothetical linear relationship between y and $\mathbf{x}_F$ in U_{UC} would be the following: suppose, hypothetically, that the linear population relationship between y and $\mathbf{x}_F$ were the same for elements in U_{UC} as for elements in U_I,* but for the factor Λ_F, *where Λ_F is a $Q \times Q$ diagonal matrix with values $\lambda_{F1}, \ldots, \lambda_{FQ}$ on the main diagonal. I.e. we assume, hypothetically, that*

$$\begin{aligned}
y_k &\approx \mathbf{x}'_{Fk} \mathbf{B}_{\mathbf{x}_F U_{UC}} \\
&= \mathbf{x}'_{Fk} \Lambda_F \mathbf{B}_{\mathbf{x}_F U_I} \quad \textit{for} \quad k \in U_{UC}
\end{aligned}$$

The assumption in item 2 above is a special case of this reasoning, i.e. when $\Lambda_F = \mathbf{I}_F$, the identity matrix.

Comparing the expressions for $\tilde{t}_{yUreg}$ and $\tilde{t}_{yU}$, we see that the only difference is that $\tilde{\mathbf{t}}$ has been replaced by $\tilde{\mathbf{t}}_{x_F U}$, and we have thus given an example of what might be needed to arrive at a useful $\tilde{\mathbf{t}}$ vector. Admittedly, the suggested estimator $\tilde{t}_{yUreg}$ relies heavily on several assumptions. However, as will be seen later in section 4, there are survey setups where we can get rid of some of these assumptions.

Remark 12 *An alternative expression for $\tilde{t}_{yUreg}$ is*

$$\tilde{t}_{yUreg} = \sum\nolimits_{s_I} \tilde{w}_k y_k = \sum\nolimits_{s_I} \tilde{v}_k \check{y}_k \tag{43}$$

where $\tilde{w}_k = \tilde{v}_k / \pi_{Fk}$ and

$$\tilde{v}_k = 1 + c_k \left(\tilde{\mathbf{t}}_{x_F U} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk} \right)' \left(\sum\nolimits_{s_I} c_k \mathbf{x}_{Fk} \check{\mathbf{x}}'_{Fk} \right)^{-1} \mathbf{x}_{Fk} \tag{44}$$

for $k \in s_I$

The bias of $\tilde{t}_{yUreg}$ with respect to t_{yU} is given by

$$B\left(\tilde{t}_{yUreg}\right) \doteq \left(\tilde{\mathbf{t}}_{x_F U} - \mathbf{t}_{x_F U_I} \right)' \mathbf{B}_{\mathbf{x}_F U_I} - t_{yU_{UC}} \tag{45}$$

Furthermore, the approximate variance is given by

$$AV\left(\tilde{t}_{yUreg}\right) = \sum \sum\nolimits_{U_I} \Delta_{Fkl} \check{E}^{\alpha}_{Fk} \check{E}^{\alpha}_{Fl} \tag{46}$$

where

$$E^{\alpha}_{Fk} = \left(1 + \boldsymbol{\alpha}' c_k \mathbf{x}_{Fk}\right) E_{Fk}$$

with $\boldsymbol{\alpha}' = \left(\tilde{\mathbf{t}}_{x_F U} - \mathbf{t}_{x_F U_I} \right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I}$ and $E_{Fk} = y_k - \mathbf{x}'_{Fk} \mathbf{B}_{\mathbf{x}_F U_I}$. Both the bias and the approximate variance of $\tilde{t}_{yUreg}$ follows from Taylor linearization of $\tilde{t}_{yUreg}$. These results follows easily from the proof of the Taylor linearization of $\tilde{t}_{yU_I reg}$ which is given in appendix A.1.

A variance estimator would be

$$\hat{V}\left(\tilde{t}_{yUreg}\right) = \sum \sum\nolimits_{s_I} \check{\Delta}_{Fkl} \check{e}^{\hat{\alpha}}_{Fk} \check{e}^{\hat{\alpha}}_{Fl} \tag{47}$$

where $e^{\hat{\alpha}}_{Fk} = \left(1 + \hat{\boldsymbol{\alpha}}' c_k \mathbf{x}_{Fk}\right) e_{Fk}$ with $\hat{\boldsymbol{\alpha}}' = \left(\tilde{\mathbf{t}}_{x_F U} - \sum_{s_I} \check{\mathbf{x}}_{Fk} \right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I}$ and $e_{Fk} = y_k - \mathbf{x}'_{Fk} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I}$.

Remark 13 *An alternative to $e^{\hat{\alpha}}_{Fk}$ in $\hat{V}\left(\tilde{t}_{yUreg}\right)$ would be e^{α}_{Fk}. At present it is not clear which choice is the better and some further work is needed in deciding whether to use $\boldsymbol{\alpha}'$ or $\hat{\boldsymbol{\alpha}}'$ in this variance estimator.*

In a few, probably rare, cases it may be reasonable to assume that the known $\mathbf{t}_{x_F U_F}$ could serve as approximation to the fictitious $\mathbf{t}_{x_F U}$. In these cases a variant of the more elaborate approach would be to use the known $\mathbf{t}_{x_F U_F}$ as an approximation to the unknown $\mathbf{t}_{x_F U}$. Thus, for equation (42) this will result in

$$\hat{t}_{yUreg}^{alt} = \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{x_F U_F} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{48}$$

As an example, suppose that $\mathbf{x}_{Fk}$ is scalar, i.e. $\mathbf{x}_{Fk} = x_{Fk}$, and $c_k = 1$. Then

$$\begin{aligned}\hat{t}_{yUreg}^{alt} &= \sum\nolimits_{s_I} \check{y}_k + \left(t_{x_F U_F} - \sum\nolimits_{s_I} \check{x}_{Fk}\right) \frac{\sum_{s_I} \check{y}_k}{\sum_{s_I} \check{x}_{Fk}} \\ &= \frac{t_{x_F U_F}}{\sum_{s_I} \check{x}_{Fk}} \sum\nolimits_{s_I} \check{y}_k\end{aligned}$$

which is biased with bias given by

$$B\left(\hat{t}_{yUreg}^{alt}\right) \doteq t_{x_F U_F} \frac{t_{yU_I}}{t_{x_F U_I}} - t_{yU}$$

For the special case $x_{Fk} = 1$ for every $k \in U_F$, the bias is given by

$$B\left(\hat{t}_{yUreg}^{alt}\right) \doteq N_F \frac{t_{yU_I}}{N_I} - t_{yU} = N_F \bar{y}_U \left(\frac{\bar{y}_{U_I}}{\bar{y}_U} - \frac{N}{N_F}\right)$$

Obviously, the bias can be positive as well as negative. If, for example, $\bar{y}_{U_I} < \bar{y}_U$, it will always be negative if $N > N_F$, while if $\bar{y}_{U_I} > \bar{y}_U$, it will always be positive if $N < N_F$.

3.3 Conclusion

In the standard estimation setup there exists no information on the undercoverage. In cases where it seems reasonable to assume that the undercoverage is negligible it may suffice to use $\hat{t}_{yU_I}$ as an estimator of t_{yU}. However, if this is not the case, the statistician faces a most delicate situation, where he/she has to find a way to adjust for the negative bias the undercoverage brings about. Since no information exists on the undercoverage this adjustment will rely on more or less speculative reasoning which may have to be applied to many study variables separately.

4 Imperfect frames - improved estimation setup using an up-to-date current register

4.1 Introduction and notation

Now we introduce the concept of the *register population*, denoted U_R. The register population is the set of all elements that can be reached via an up-to-date current register. 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 completely different from the original frame, i.e. a newly developed register. The register population accessible from the current register matches the target population better than the frame population. Henceforth we assume that the current register is perfect in the sense that the register population equals the target population, i.e., $U_R = U$.

Example 4.1 A survey conducted at Statistics Sweden is the "Kortperiodisk industrienkät". The aim of this survey is to measure variables such as order intake and order deliveries every month of the year. The target population consists of industrial enterprises ("industriföretag") and the Business Register(BR) is used as frame. There is a cut-off limit at 10 employees, i.e. enterprises with less than 10 employees are deliberately excluded from the sample selection. The sample is drawn in March and the questionnaire is sent out to these enterprises once a month in March to July. A new sample is drawn in August and the questionnaire is sent out to these new enterprises in August to February. Although the sample is renewed during the year it is still likely to be a certain amount of "births" and "deaths" in for example the period between August and February. However, since the BR is updated during this period there is a possibility that the updated BR is a better match to the target population then the BR version used for sampling frame. The updated BR is thus a potential candidate to serve as a current register in this survey. □

We assume that the current register contains auxiliary information, i.e. associated with every $k \in U$ is a vector $\mathbf{x}_k$, where

$$\mathbf{x}_k = (x_{1k}, \ldots, x_{pk}, \ldots x_{Jk})'$$

and

$$\mathbf{t}_x = \left(t_{x_1}, \ldots, t_{x_j}, \ldots, t_{x_J}\right)'$$

where $t_{x_j} = \sum_U x_{jk}$. The J variables in $\mathbf{x}_k$ may consist of updated values of the variables in $\mathbf{x}_{Fk}$. In this case $J = Q$.

We now have an improved setup for making inference to the target population. Firstly, it would be possible to draw a probability sample from U_{UC}. 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. But the current register also provides information on the undercoverage. We now have an auxiliary vector $\mathbf{x}_k$, not only for the elements in U_I, but also for every $k \in U_{UC}$. This information may be used with the regression estimator in order to get better estimates. Besides, for the elements $k \in U_I$ the vector $\mathbf{x}_k$ contains more up-to-date information than does $\mathbf{x}_{Fk}$. The current register also enables the identification of the subsets

$$U_I, U_{OC} \text{ and } U_{UC}$$

and thus

$$N_I, N_{OC} \text{ and } N_{UC}$$

are known quantities.

Finally, the information at hand also implies that it is possible to identify the subsets s_I and s_{OC} from s_F irrespective of the information from the sample (see previous section).

In the rest of this section, we will study in what way the access to a current register could improve inference as compared to the standard estimation setup.

4.2 Estimation of t_{yU_I}

Recall that in the standard estimation setup two basic estimators were presented, the simple domain π estimator

$$\hat{t}_{yU_I\pi} = \sum\nolimits_{s_I} \check{y}_k = \sum\nolimits_{s_I} \check{y}_k/\pi_{Fk} \tag{49}$$

and

$$\tilde{t}_{yU_I} = N_I \tilde{y}_{s_I} = N_I \hat{t}_{yU_I\pi}/\hat{N}_I \tag{50}$$

In the present setup the total N_I is known and it is thus possible to use $\tilde{t}_{yU_I}$. This is an improvement compared to the standard estimation setup, since the variance of $\tilde{t}_{yU_I}$ ordinarily is smaller than the variance of $\hat{t}_{yU_I\pi}$.

In section 3.1 we presented four different regression estimators for domains. For different reasons we rejected three of them, i.e. $\hat{t}^{(i)}_{yU_Ireg}, \hat{t}^{(ii)}_{yU_Ireg}$ and $\hat{t}^{(iii)}_{yU_Ireg}$. Now, these three estimators are, for partially the same reasons, still rejected, whereas we can, using the assumptions under the improved estimation setup, modify the fourth estimator, $\hat{t}^{(iv)}_{yU_Ireg}$. In the standard setup we had

$$\hat{t}^{(iv)}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\sum\nolimits_{U_I} \mathbf{x}_{Fk} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{51}$$

with

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_F s_I} \widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} \\ &= \left(\sum\nolimits_{s_I} c_k \mathbf{x}_{Fk} \check{\mathbf{x}}'_{Fk}\right)^{-1} \sum\nolimits_{s_I} c_k \mathbf{x}_{Fk} \check{y}_k \end{aligned} \tag{52}$$

and the unknown total

$$\mathbf{t}_{x_F U_I} = \sum\nolimits_{U_I} \mathbf{x}_{Fk} \tag{53}$$

We suggested the use of

$$\tilde{t}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I} \tag{54}$$

which required the statistician to come up with the approximation $\tilde{\mathbf{t}}_{x_F U_I}$.

Using the known total $\mathbf{t}_{xU_I} = \sum_{U_I} \mathbf{x}_k$ instead of the unknown $\mathbf{t}_{x_F U_I}$ and the vector $\mathbf{x}_k$ instead of $\mathbf{x}_{Fk}$ we now can use the estimator

$$\hat{t}^{new}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I} \tag{55}$$

where

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}s_I} &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}\mathbf{x}s_I} \widehat{\mathbf{t}}_{\mathbf{x}y s_I} \\ &= \left(\sum\nolimits_{s_I} c_k \mathbf{x}_k \check{\mathbf{x}}'_k\right)^{-1} \sum\nolimits_{s_I} c_k \mathbf{x}_k \check{y}_k \end{aligned} \tag{56}$$

Obviously, this is an improvement from the standard setup, since (1) we have the known total $\mathbf{t}_{xU_I} = \sum_{U_I} \mathbf{x}_k$ instead of the more ore less correct conjecture $\tilde{\mathbf{t}}_{x_F U_I}$ of $\mathbf{t}_{x_F U_I}$ and (2) we have the updated vector $\mathbf{x}$ instead of $\mathbf{x}_F$.

Remark 14 *An alternative to (55) would be to modify (16) to fit this situation, i.e.we would have*

$$\hat{t}_{yU_I reg}^{new,alt} = \frac{N_I}{\hat{N}_I}\sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU_I} - \frac{N_I}{\hat{N}_I}\sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I}$$

However, in this paper we consider mainly situations when the overcoverage is limited and the possible gain in efficiency over (55) will be modest.

An alternative expression of $\hat{t}_{yU_I reg}^{new}$ is

$$\hat{t}_{yU_I reg}^{new} = \sum\nolimits_{s_I} w_{Ik} y_k = \sum\nolimits_{s_I} v_{Ik} \check{y}_k \tag{57}$$

where $w_{Ik} = v_{Ik}/\pi_{Fk}$ and

$$v_{Ik} = 1 + c_k \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{s_I} c_k \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{58}$$

By Taylor linearization the estimator $\hat{t}_{yU_I reg}^{new}$ is approximated by

$$\hat{t}_{yU_I reg}^{new} \doteq \mathbf{t}_{xU_I}' \mathbf{B}_{\mathbf{x}U_I} + \sum\nolimits_{s_I} \check{E}_k \tag{59}$$

where $\mathbf{B}_{\mathbf{x}U_I} = \mathbf{T}_{\mathbf{xx}U_I}^{-1} \mathbf{t}_{\mathbf{x}yU_I} = \left(\sum_{U_I} c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{U_I} c_k \mathbf{x}_k y_k$ and $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}U_I}$. The proof of this linearization follows easily from the proof of the Taylor linearization of $\tilde{t}_{yU_I reg}$ which is given in appendix A.1. The expected value of $\hat{t}_{yU_I reg}^{new}$ is given by

$$\begin{aligned} E\left(\hat{t}_{yU_I reg}^{new}\right) &\doteq \mathbf{t}_{xU_I}' \mathbf{B}_{\mathbf{x}U_I} + E\left(\sum\nolimits_{s_I} \check{E}_k\right) \\ &= t_{yU_I} \end{aligned} \tag{60}$$

which indicates that this estimator is approximately unbiased. The approximate variance for $\hat{t}_{yU_I reg}^{new}$ is

$$AV\left(\hat{t}_{yU_I reg}^{new}\right) = \sum\sum\nolimits_{U_I} \Delta_{Fkl} \check{E}_k \check{E}_l \tag{61}$$

where $\Delta_{Fkl} = \pi_{Fkl} - \pi_{Fk}\pi_{Fl}$. A variance estimator would be

$$\hat{V}\left(\hat{t}_{yU_I reg}^{new}\right) = \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl} v_{Ik} \check{e}_k v_{Il} \check{e}_l \tag{62}$$

where v_{Ik} is given by (58) and $e_k = y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_{\mathbf{x}s_I}$.

Obviously, even if this is a good estimator for t_{yU_I}, we still have the problem that we lack information on y for every $k \in U_{UC}$.

4.3 Estimation of t_{yU} in analogy with the standard setup

Two approaches were suggested under the standard estimation setup. Obviously, corresponding approaches may be used also under this setup.

4.3.1 The simplistic approach

Recall that the hypothetical reasoning in the simplistic approach in section 3.2.1 was, in short:

1. The relation between the population total t_{yU} and the domain total t_{yU_I} was expressed as $t_{yU} = \delta_y t_{yU_I}$, where δ_y is unknown.
2. A good approximation, $\tilde{\delta}_y$, of δ_y was at hand.
3. As an estimator for t_{yU}, use $\hat{t}_{yU\tilde{\delta}_y} = \tilde{\delta}_y \hat{t}_{yU_I}$, where $\hat{t}_{yU_I}$ is an (approximately) unbiased estimator for t_{yU_I}.

Obviously, under this improved estimation setup we will, using this hypothetical reasoning, get the same general results for $\hat{t}_{yU\tilde{\delta}_y}$, i.e. if $\hat{t}_{yU_I}$ is (approximately) unbiased for the total t_{yU_I}, the (approximate) relative bias for $\hat{t}_{yU\tilde{\delta}_y}$ is

$$\begin{aligned} RB\left(\hat{t}_{yU\tilde{\delta}_y}\right) &= \frac{\tilde{\delta}_y t_{yU_I} - t_{yU}}{t_{yU}} = \frac{\frac{\tilde{\delta}_y}{\delta_y}\delta_y t_{yU_I} - t_{yU}}{t_{yU}} \\ &= \frac{\tilde{\delta}_y}{\delta_y} - 1 \end{aligned} \tag{63}$$

and the (approximate) variance

$$AV\left(\hat{t}_{yU\tilde{\delta}_y}\right) = \tilde{\delta}_y^2 AV\left(\hat{t}_{yU_I}\right) \tag{64}$$

where $AV\left(\hat{t}_{yU\tilde{\delta}_y}\right)$ depends on the choice of $\hat{t}_{yU_I}$. However, the difference now from the standard estimation setup is the new possibilities with improved estimators for t_{yU_I}, i.e. (approximately) unbiased estimators with smaller variances, viz. $\tilde{t}_{yU_I}$ and $\hat{t}^{new}_{yU_Ireg}$.

4.3.2 The more elaborate approach

Recall from the standard setup, that the adjustment for undercoverage using the more elaborate approach relied heavily on several hypothetical constructs:

1. A good approximation $\tilde{\mathbf{t}}_{x_F U_I}$ to the unknown $\mathbf{t}_{x_F U_I}$.
2. A close approximation $\tilde{\mathbf{t}}_{x_F U_{UC}}$ to the hypothetical $\mathbf{t}_{x_F U_{UC}}$.
3. An assumption saying that, for $k \in U_{UC}$, the linear relationship between y and a hypothetical $\mathbf{x}_F$ is the same as for $k \in U_I$, i.e., $y_k \approx \mathbf{x}'_{Fk}\mathbf{B}_{\mathbf{x}_F U_I}$.

Now, since we can identify every $k \in U$, and whether it belongs to U_I or U_{UC}, and have an updated auxiliary vector, $\mathbf{x}$, available for every $k \in U_I$ *and* for every $k \in U_{UC}$, we are in a much better situation: $\mathbf{t}_{xU_I}$ as well as $\mathbf{t}_{xU_{UC}}$ are known. This means that we are not at the mercy of item 1 and item 2 above. However, the probem with finding an estimator for $t_{yU_{UC}}$ still holds. Retaining the assumption on the linear relationship between y and the auxiliary variable ($\mathbf{x}$ in the present setup) a possible new estimator for t_{yU} would be

$$\begin{aligned}
\hat{t}^{new}_{yUreg} &= \hat{t}^{new}_{yU_I reg} + \hat{t}_{yU_{UC}} \\
&= \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I} + \mathbf{t}'_{xU_{UC}}\widehat{\mathbf{B}}_{\mathbf{x}s_I} \\
&= \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I} \qquad (65)
\end{aligned}$$

Remark 15 *If we assume that the linear relationship between y and $\mathbf{x}$ in the undercoverage set is better described by using a subgroup of s_I, say s_{Ig}, we would use $\mathbf{t}'_{xU_{UC}}\widehat{\mathbf{B}}_{\mathbf{x}s_{Ig}} = \mathbf{t}'_{xU_{UC}}\left(\sum_{s_{Ig}} c_k\mathbf{x}_k\check{\mathbf{x}}'_k\right)^{-1}\sum_{s_{Ig}} c_k\mathbf{x}_k\check{y}_k$ to estimate $t_{yU_{UC}}$.*

Remark 16 *An alternative way of reasoning about the hypothetical linear relationship between y and $\mathbf{x}$ in U_{UC} would be the following: suppose, hypothetically, that the linear population relationship between y and $\mathbf{x}$ were the same for elements in U_{UC} as for elements in U_I,* but for the factor Λ, *where Λ is a $J \times J$ diagonal matrix with values $\lambda_1, \ldots, \lambda_J$ on the main diagonal. I.e. we assume, hypothetically, that*

$$\begin{aligned}
y_k &\approx \mathbf{x}'_k\mathbf{B}_{\mathbf{x}U_{UC}} \\
&= \mathbf{x}'_k\Lambda\mathbf{B}_{\mathbf{x}U_I} \quad \textit{for} \quad k \in U_{UC}
\end{aligned}$$

The assumption in item 3 above is a special case of this reasoning, i.e. when $\Lambda = \mathbf{I}$, *the identity matrix.*

Comparing this estimator with the variant of the elaborate approach in the standard estimation setup, i.e. with

$$\hat{t}_{yUreg}^{alt} = \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{x_F U_F} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F s_I}$$

we see that the situation has improved. We have an updated auxiliary vector, $\mathbf{x}$, and we can use the known target population total $\mathbf{t}_{xU}$.

Remark 17 *An alternative expression for* $\hat{t}_{yUreg}^{new}$ *is*

$$\hat{t}_{yUreg}^{new} = \sum\nolimits_{s_I} w_k y_k = \sum\nolimits_{s_I} v_k \check{y}_k \tag{66}$$

where $w_k = v_k / \pi_{Fk}$ *and*

$$v_k = 1 + c_k \left(\mathbf{t}_{xU} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{s_I} c_k \mathbf{x}_k \check{\mathbf{x}}_k'\right)^{-1} \mathbf{x}_k \tag{67}$$

for $k \in s_I$

By Taylor linearization the estimator $\hat{t}_{yUreg}^{new}$ is approximated by

$$\begin{aligned} \hat{t}_{yUreg}^{new} &\doteq \mathbf{t}_{xU}' \mathbf{B}_{\mathbf{x}U_I} + \sum\nolimits_{s_I} \left(1 + \mathbf{t}_{xU_{UC}}' \mathbf{T}_{\mathbf{xx}U_I}^{-1} c_k \mathbf{x}_k\right) \check{E}_k \\ &= \mathbf{t}_{xU}' \mathbf{B}_{\mathbf{x}U_I} + \sum\nolimits_{s_I} \check{E}_k^{a_x} \end{aligned} \tag{68}$$

where $E_k^{a_x} = \left(1 + \mathbf{a}_x' c_k \mathbf{x}_k\right) E_k$ with $\mathbf{a}_x' = \mathbf{t}_{xU_{UC}}' \mathbf{T}_{\mathbf{xx}U_I}^{-1}$ and $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}U_I}$. The proof follows easily from the proof of the Taylor linearization of $\tilde{t}_{yU_Ireg}$ which is given in appendix A.1.

The bias of $\hat{t}_{yUreg}^{new}$ is given by

$$\begin{aligned} B\left(\hat{t}_{yUreg}^{new}\right) &\doteq \sum\nolimits_{U_I} y_k + \mathbf{t}_{xU_{UC}}' \mathbf{B}_{\mathbf{x}U_I} - t_{yU} \\ &= \mathbf{t}_{xU_{UC}}' \mathbf{B}_{\mathbf{x}U_I} - t_{yU_{UC}} \end{aligned} \tag{69}$$

The approximate variance of $\hat{t}_{yUreg}^{new}$ is

$$AV\left(\hat{t}_{yUreg}^{new}\right) = \sum\sum\nolimits_{U_I} \Delta_{Fkl} \check{E}_k^{a_x} \check{E}_l^{a_x} \tag{70}$$

where $\Delta_{Fkl} = \pi_{Fkl} - \pi_{Fk}\pi_{Fl}$ and a variance estimator is given by

$$\hat{V}\left(\hat{t}_{yUreg}^{new}\right) = \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl}\check{e}_k^{\hat{a}_x}\check{e}_l^{\hat{a}_x} \tag{71}$$

where $e_k^{\hat{a}_x} = (1 + \hat{\mathbf{a}}_x' c_k \mathbf{x}_k) e_k$ with $\hat{\mathbf{a}}_x' = \mathbf{t}_{xU_{UC}}' \widehat{\mathbf{T}}_{\mathbf{xx}s_I}^{-1}$ and $e_k = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}_{\mathbf{x}s_I}$.

Remark 18 *An alternative to $e_{Fk}^{\hat{\alpha}_x}$ in $\hat{V}\left(\hat{t}_{yUreg}^{new}\right)$ would be $e_{Fk}^{\alpha_x}$. At present it is not clear which choice is the better and some further work is needed in deciding whether to use $\boldsymbol{\alpha}_x'$ or $\hat{\boldsymbol{\alpha}}_x'$ in this variance estimator.*

4.4 Estimation of t_{yU} borrowing techniques from nonresponse treatment

Recall that the sample s_F is drawn from U_F with inclusion probabilities π_{Fk}. Observed values on y for $k \in s_I$ are at hand. In the present improved setup, with access to a perfect up-to-date current register, we could define the inference situation as follows.

Let us look upon $s_T = s_{OC} \cup s_I \cup U_{UC}$ as a sample that has been drawn from $U_T = U_{OC} \cup U_I \cup U_{UC}$ with inclusion probabilities $\pi_{Tk} = \pi_{Fk}$ for $k \in U_F$ and $\pi_{Tk} = 1$ for $k \in U_{UC}$. For this sample we have

$$y_k = \begin{cases} 0 & \text{for} \quad k \in s_{OC} \\ y_k & \text{for} \quad k \in s_I \\ - & \text{for} \quad k \in U_{UC} \end{cases}$$

where $-$ represents missing values. We want to estimate the total

$$\begin{aligned} t_{yU} &= \sum\nolimits_U y_k = \sum\nolimits_{U_I} y_k + \sum\nolimits_{U_{UC}} y_k \\ &= \underbrace{\sum\nolimits_{U_{OC}} y_k}_{0} + \sum\nolimits_{U_I} y_k + \sum\nolimits_{U_{UC}} y_k = t_{yU_T} \end{aligned}$$

Using this point of view, we might try to utilize approaches developed for the treatment of nonresponse, although the missing data now is of a different nature. Two such main approaches can be distinguished, viz. *imputation* and *reweighting*. We will in the following only consider imputation.

4.4.1 Imputation

First, suppose that, besides the observed y_k for $k \in s_I$, we also had access to y-data for every $k \in U_{UC}$, and let $s = s_I \cup U_{UC}$. Since $U \subset U_T$ we could use domain estimation techniques for estimating t_{yU}. A few examples are:

The simple domain "estimator"

$$\begin{aligned} \hat{t}_1 &= \hat{t}_{yU\pi} = \hat{t}_{yU_T\pi} = \sum\nolimits_{s_T} \check{y}_k = \sum\nolimits_{s} \frac{y_k}{\pi_{Tk}} \\ &= \hat{t}_{yU_I\pi} + t_{yU_{UC}} = \sum\nolimits_{s_I} y_k/\pi_{Fk} + \sum\nolimits_{U_{UC}} y_k \end{aligned} \tag{72}$$

or, since N is known, the alternative "estimator"

$$\hat{t}_2 = \tilde{t}_{yU} = N\tilde{y}_{s_T} = \frac{N}{\hat{N}}\hat{t}_{yU\pi} \tag{73}$$

where $\hat{N} = \hat{N}_I + N_{UC} = \sum_{s_I} 1/\pi_{Fk} + N_{UC}$. A third simple alternative "estimator" would be

$$\hat{t}_3 = \tilde{t}_{yU_I} + t_{yU_{UC}} = \frac{N_I}{\hat{N}_I}\hat{t}_{yU_I\pi} + t_{yU_{UC}} \tag{74}$$

There are several potential domain regression "estimators", one of which is

$$\hat{t}_4 = \hat{t}_{yU_Ireg} + t_{yU_{UC}} \tag{75}$$

where

$$\hat{t}_{yU_Ireg} = \sum\nolimits_{s_I} \frac{y_k}{\pi_{Fk}} + \left(\sum\nolimits_{U_I} \mathbf{x}_k - \sum\nolimits_{s_I} \frac{\mathbf{x}_k}{\pi_{Fk}}\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I}$$

and

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathbf{x}s_I} &= \hat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I}\hat{\mathbf{t}}_{\mathbf{x}ys_I} \\ &= \left(\sum\nolimits_{s_I} c_k\mathbf{x}_k\frac{\mathbf{x}'_k}{\pi_{Fk}}\right)^{-1} \sum\nolimits_{s_I} c_k\mathbf{x}_k\frac{y_k}{\pi_{Fk}} \end{aligned} \tag{76}$$

Obviously none of the above "estimators" can be used, since y-data are missing for all $k \in U_{UC}$. One approach now, would be to pick a suitable imputation technique, leading to the following data setup:

$$y_k^* = \begin{cases} 0 & \text{för} \quad k \in U_{OC} \\ y_k & \text{för} \quad k \in U_I \\ \hat{y}_k & \text{för} \quad k \in U_{UC} \end{cases}$$

where $\hat{y}_k$ denotes the imputed value for unit *k,* determined by the adopted imputation technique. There exist many imputation techniques. Some of the more commonly used are: respondent mean imputation, hot deck imputation, nearest neighbour imputation, deterministic and random multiple regression imputation, including the special case ratio imputation. Usually, the different techniques are applied within imputation classes.

We will here only consider the following deterministic multiple regression approach. Let $\mathbf{u}_k$, a column vector with P components, denote the auxiliary vector value used in producing the imputed values $\hat{y}_k$. The vector is denoted $\mathbf{u}_k$ in order to distinguish it from the auxiliary vector $\mathbf{x}_k$ appearing in the regression estimator, since the two vectors not necessarily are identical. Let the imputed value for element k be determined by

$$\hat{y}_k = \mathbf{u}_k' \, \widehat{\mathbf{B}}_{\mathbf{u}s_I} \tag{77}$$

where

$$\widehat{\mathbf{B}}_{\mathbf{u}s_I} = \left(\sum\nolimits_{s_I} h_k \mathbf{u}_k \check{\mathbf{u}}_k' \right)^{-1} \sum\nolimits_{s_I} h_k \mathbf{u}_k \check{y}_k \tag{78}$$

where h_k is a suitably chosen constant, e.g. capturing an assumed heteroscedasticity in the linear relationship between y and $\mathbf{u}$.

Remark 19 *If we assume that the linear relationship between y and $\mathbf{u}$ in the undercoverage set is more well described by using a subgroup of s_I, say s_{Ig}, we would use $\widehat{\mathbf{B}}_{\mathbf{u}s_{Ig}} = \left(\sum_{s_{Ig}} h_k \mathbf{u}_k \check{\mathbf{u}}_k' \right)^{-1} \sum_{s_{Ig}} h_k \mathbf{u}_k \check{y}_k$ instead of $\widehat{\mathbf{B}}_{\mathbf{u}s_I}$.*

Let us consider imputation using two different levels of auxiliary information.

(1) If the frame is "information poor", we may simply have to use $\mathbf{u}_k = 1 = h_k$, which leads to

$$\hat{y}_k = \widehat{B} = \frac{\sum_{s_I} y_k/\pi_{Fk}}{\sum_{s_I} 1/\pi_{Fk}} = \frac{\hat{t}_{yU_I\pi}}{\hat{N}_I} = \tilde{y}_{s_I}$$

Furthermore, we probably will have to insert $\hat{y}_k$ into the "prototype estimators" $\hat{t}_2$ or $\hat{t}_3$. The resulting estimators coincide, and we get the estimator

$$\hat{t}^* = \hat{t}_2^* = \hat{t}_3^* = N\tilde{y}_{s_I} \tag{79}$$

which is biased for t_{yU}, with the bias given by

$$B(\hat{t}^{*}) \doteq N(\bar{y}_{U_I} - \bar{y}_U) = N\bar{y}_U\left(\frac{\bar{y}_{U_I}}{\bar{y}_U} - 1\right) \tag{80}$$

Thus, the sign of the bias here only depends on the relation between $\bar{y}_{U_I}$ and $\bar{y}_U$, while an evaluation of the sign of the bias of $\hat{t}^{alt}_{yUreg}$ in the standard setup, i.e.

$$B\left(\hat{t}^{alt}_{yUreg}\right) \doteq N_F\bar{y}_U\left(\frac{\bar{y}_{U_I}}{\bar{y}_U} - \frac{N}{N_F}\right)$$

also had to include considerations on the relation between N and N_F.

The approximate design variance of $\hat{t}^{*}$ is given by

$$AV\left(\hat{t}^{*}\right) = \left(\frac{N}{N_I}\right)^2 \sum\sum\nolimits_{U_I} \Delta_{Fkl}\left(\frac{y_k - \bar{y}_{U_I}}{\pi_{Fk}}\right)\left(\frac{y_l - \bar{y}_{U_I}}{\pi_{Fl}}\right) \tag{81}$$

while a variance estimator is given by

$$\hat{V}\left(\hat{t}^{*}\right) = \left(\frac{N}{\hat{N}_I}\right)^2 \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl}\left(\frac{y_k - \tilde{y}_{s_I}}{\pi_{Fk}}\right)\left(\frac{y_l - \tilde{y}_{s_I}}{\pi_{Fl}}\right) \tag{82}$$

(see Särndal et al. (1992)).

(2) If the frame is "information rich", we can insert $\hat{y}_k$ into the "prototype estimator" $\hat{t}_4$, which gives

$$\hat{t}^{*}_{yUreg} = \hat{t}_{yU_Ireg} + \left(\sum\nolimits_{U_C}\mathbf{u}'_k\right)\widehat{\mathbf{B}}_{\mathbf{u}s_I} \tag{83}$$

Through Taylor linearization the bias of $\hat{t}^{*}_{yUreg}$ is given by

$$\begin{aligned} B\left(\hat{t}^{*}_{yUreg}\right) &\doteq \sum\nolimits_{U_I} y_k + \sum\nolimits_{U_C}\mathbf{u}'_k\mathbf{B}_{\mathbf{u}U_I} - t_{yU} \\ &= \sum\nolimits_{U_C}\mathbf{u}'_k\mathbf{B}_{\mathbf{u}U_I} - t_{yU_C} \end{aligned} \tag{84}$$

where $\mathbf{B}_{\mathbf{u}U_I} = \left(\sum_{U_I} h_k\mathbf{u}_k\mathbf{u}'_k\right)^{-1}\sum_{U_I} h_k\mathbf{u}_k y_k$. A proof is given in appendix A.2. The approximate design variance is

$$AV\left(\hat{t}^{*}_{yUreg}\right) = \sum\sum\nolimits_{U_I}\Delta_{Fkl}\check{F}_k\check{F}_l \tag{85}$$

where $F_k = E_{xk} + \boldsymbol{\alpha}'_u h_k \mathbf{u}_k E_{uk}$ with $E_{xk} = y_k - \mathbf{x}'_k \mathbf{B}_{\mathbf{x}U_I}$, $\boldsymbol{\alpha}'_u = \mathbf{t}'_{uU_{UC}} \mathbf{T}^{-1}_{\mathbf{uu}U_I}$ and $E_{uk} = y_k - \mathbf{u}'_k \mathbf{B}_{\mathbf{u}U_I}$. A variance estimator is given by

$$\hat{V}\left(\hat{t}^*_{yUreg}\right) = \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl} \check{f}_k \check{f}_l \tag{86}$$

where $f_k = e_{xk} + \hat{\boldsymbol{\alpha}}'_u h_k \mathbf{u}_k e_{uk}$ with $e_{xk} = y_k - \mathbf{x}'_k \widehat{\mathbf{B}}_{\mathbf{x}s_I}$, $\hat{\boldsymbol{\alpha}}'_u = \mathbf{t}'_{uU_{UC}} \widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I}$ and $e_{uk} = y_k - \mathbf{u}'_k \widehat{\mathbf{B}}_{\mathbf{u}s_I}$.

Remark 20 *An alternative to* $\hat{\boldsymbol{\alpha}}'_u$ *in* $\hat{V}\left(\hat{t}^*_{yUreg}\right)$ *would be* $\boldsymbol{\alpha}'_u$*. At present it is not clear which choice is the better and some further work is needed in deciding whether to use* $\boldsymbol{\alpha}'_u$ *or* $\hat{\boldsymbol{\alpha}}'_u$ *in this variance estimator.*

If we find it appropriate to use $\mathbf{u}_k = \mathbf{x}_k$ and $h_k = c_k$, we arrive at

$$\begin{aligned} \hat{t}^*_{yUreg} &= \hat{t}_{yU_Ireg} + \left(\sum\nolimits_{U_{UC}} \mathbf{x}'_k\right) \widehat{\mathbf{B}}_{\mathbf{x}s_I} \\ &= \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{B}}_{\mathbf{x}s_I} \end{aligned}$$

which is the estimator $\hat{t}^{new}_{yUreg}$ from section 4.3.2, from where we get bias, design variance and a variance estimator for $\hat{t}^*_{yUreg}$ in this special case.

Another approach for finding a variance estimator for $\hat{t}^*_{yUreg}$ would be the following.
Consider the imputation model

$$y_k = \mathbf{u}'_k \boldsymbol{\beta} + \varepsilon_k$$

for $k \in s$ (recall that $s = s_I \cup U_{UC}$)

$$\begin{aligned} E_\xi(\varepsilon_k) &= 0 \\ E_\xi\left(\varepsilon_k \varepsilon_l\right) &= \begin{cases} \sigma^2 a_k & \text{for } k = l \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Now,

$$\hat{t}^*_{yUreg} = \hat{t}_{yU_Ireg} + \sum\nolimits_{U_{UC}} \hat{y}_k$$

where

$$\hat{y}_k = \mathbf{u}'_k \hat{\boldsymbol{\beta}}$$

for $k \in U_{UC}$, with

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{U}'\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{A}^{-1}\mathbf{y} = \left(\hat{\beta}_1, \dots, \hat{\beta}_P\right)'$$

where $\mathbf{U}$ is a $(n_{s_I} \times P)$ matrix with typical row element $\mathbf{u}_k'$ and $\mathbf{A}$ is a $n_{s_I} \times n_{s_I}$ diagonal matrix with typical diagonal element $\sigma^2 a_k$.
The error from using $\sum_{U_{UC}} \hat{y}_k$ instead of $\sum_{U_{UC}} y_k$ in $\hat{t}^*_{yUreg}$ is

$$\begin{aligned}
IE &= \textstyle\sum_{U_{UC}} y_k - \sum_{U_{UC}} \hat{y}_k \\
&= \textstyle\sum_{U_{UC}} y_k - \sum_{U_{UC}} \mathbf{u}_k'\hat{\boldsymbol{\beta}} \\
&= \textstyle\sum_{U_{UC}} \left(\mathbf{u}_k'\boldsymbol{\beta} + \varepsilon_k\right) - \sum_{U_{UC}} \mathbf{u}_k' \left(\boldsymbol{\beta} + \left(\mathbf{U}'\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{A}^{-1}\boldsymbol{\varepsilon}\right) \\
&= \textstyle\sum_{U_{UC}} \varepsilon_k - \sum_{U_{UC}} \underbrace{\mathbf{u}_k' \left(\mathbf{U}'\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{A}^{-1}}_{\boldsymbol{\mu}'}\boldsymbol{\varepsilon} \\
&= \textstyle\sum_{U_{UC}} \varepsilon_k - \sum_{U_{UC}} \sum_{s_I} \mu_k \varepsilon_k \\
&= \textstyle\sum_{U_{UC}} \varepsilon_k - N_{U_{UC}} \sum_{s_I} \mu_k \varepsilon_k
\end{aligned}$$

Since

$$E_\xi(IE) = 0$$

we have

$$\begin{aligned}
V_\xi(IE) &= \textstyle\sum_{U_{UC}} \sigma^2 a_k + N_{U_{UC}}^2 \sum_{s_I} \mu_k^2 \sigma^2 a_k \\
&= \left[\sum_{U_{UC}} a_k + N_{U_{UC}}^2 \sum_{s_I} \mu_k^2 a_k\right] \sigma^2 \qquad (87)
\end{aligned}$$

To find an estimator $\hat{V}_\xi(IE)$ we need to estimate σ^2. One such σ^2 estimator is given by

$$\hat{\sigma}^2 = \frac{1}{n_{s_I} - P} \textstyle\sum_{s_I} \left(y_k - \hat{y}_k\right)^2 \qquad (88)$$

Now, $\hat{t}_{yUreg} = \hat{t}_{yU_Ireg} + \sum_{U_{UC}} \hat{y}_k$ and thus a variance estimator for $\hat{t}^*_{yUreg}$ is given by

$$\hat{V}\left(\hat{t}^*_{yUreg}\right) = \hat{V}\left(\hat{t}_{yU_Ireg}\right) + \hat{V}_\xi(IE) \qquad (89)$$

Remark 21 *In the improved estimation setup of the present section 4, we have access to the known total* $\sum_U \mathbf{x}_k = \mathbf{t}_{xU}$*. Using the calibration technique under this setup will result in*

$$\hat{t}_{yUcal} = \sum\nolimits_{s_I} w_k y_k \tag{90}$$

where $w_k = v_k \ / \pi_{Fk}$*, with*

$$v_k = 1 + c_k \left(\mathbf{t}_{xU} - \sum\nolimits_{s_I} c_k \check{\mathbf{x}}_k\right)' \left(\sum\nolimits_{s_I} c_k \mathbf{x}_k \check{\mathbf{x}}_k'\right) \mathbf{x}_k \tag{91}$$

for $k \in s_I$*. Using the same* c_k *in* $\hat{t}_{yUcal}$ *as in* $\hat{t}^*_{yUreg}$*, we see that this estimator corresponds to* $\hat{t}^*_{yUreg}$ *above.*

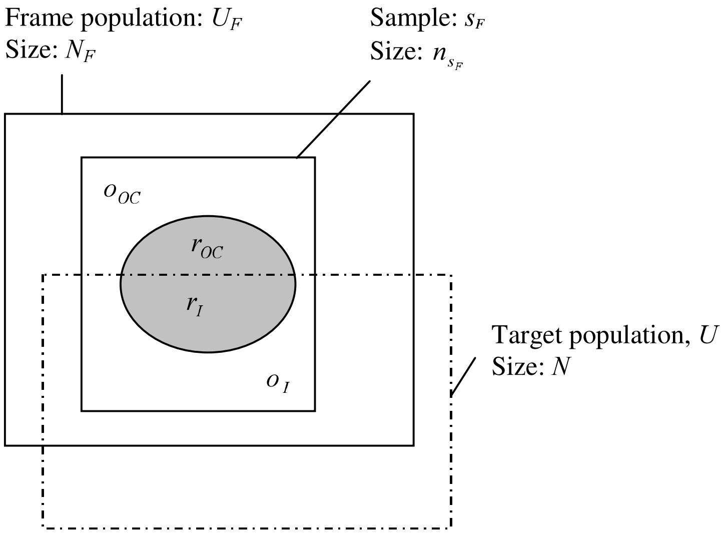
4.5 Conclusions

If no current register is at hand we will rely heavily on more or less speculative reasoning. However, when a perfect current register exists at the estimation stage of the survey, the estimation setup has improved considerably compared to the standard estimation setup. For every $k \in U_T$ we have exact knowledge of whether it belongs to U_{OC}, U_I, or U_{UC}, and hence N_{OC}, N_I, and N_{UC} are known. Furthermore, we have access to a current auxiliary vector $\mathbf{x}_k$ for every $k \in U$, and hence for every $k \in U_{UC}$. This means that better estimators can be used for t_{yU_I}, and that there are better prospects for estimation of the undercoverage total $t_{yU_{UC}}$, although it should be remembered that this latter estimation is not design unbiased; it relies on an assumption that it is possible to identify a reasonably strong linear relation between a study variable and the auxiliary vector, which is estimated using data outside U_{UC}. Also, it should be noted that this work may be laborious since it may have to be applied to many study variables separately.

5 On frame imperfection and nonresponse

In this paper we have discussed the problem of estimating a finite population total in the presence of frame imperfection, viz undercoverage and overcoverage. Typically, a survey will also suffer from nonresponse. The following figure illustrates the survey situation in the standard estimation setup.

Figure 2 *Target population, imperfect frame population and sample when nonresponse has occured. Standard estimation setup*



We now have four basic subsets of the selected sample s_F:

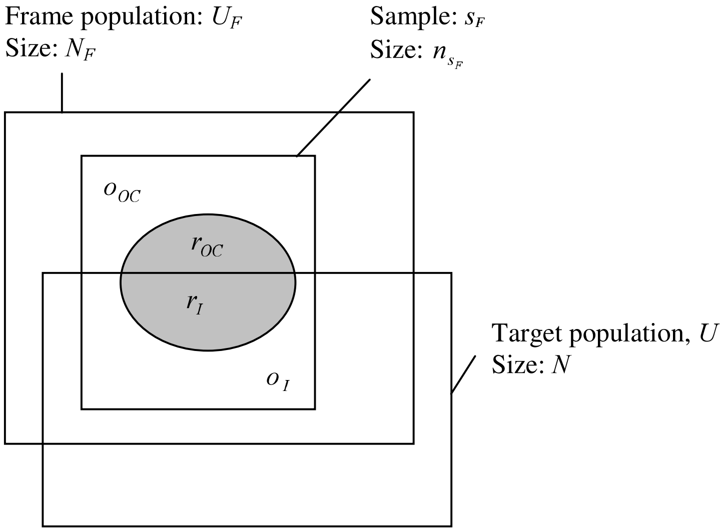
r_I	-	responding sample elements that belong to the target population - y values exist
r_{OC}	-	responding sample elements that belong to the overcoverage population - no y value exists
o_I	-	nonresponding sample elements that belong to the target population - y values exist but are missing.
o_{OC}	-	nonresponding sample elements that belong to the overcoverage population - no y value exists

As the figure reveals the situation has deteriorated compared to the situation described in section 1.4. For the responding sample elements we are able to tell whether element k belong to r_I or r_{OC}. However, for the nonresponding elements it will not be possible to determine whether element

k belongs to o_I or o_{OC}. Hence, if the overcoverage and the undercoverage is nonnegligible, we are in an akward situation, leading to highly unreliable estimates.

However, when one has access to a current register at the estimation stage of the survey and $U_R = U$, the situation improves since all subsets of s_F may be identified. The following figure illustrates this improved situation.

Figure 3. *Target population, imperfect frame population and sample when nonresponse has occured. Perfect current register at hand.*



In this situation we can use nonresponse techniques in order to estimate t_{yU_I}, where we also can take advantage of the fact that there will be no unknown overcoverage, and use techniques similar to those presented in the present paper in order to estimate $t_{yU_{UC}}$. Approaches along these lines will be presented in a forthcoming paper.

References

Deville, J. and Särndal, C. (1992). Calibration estimtors in survey sampling. **87**, 376–382.

Estevao, V., Hidiroglou, M., and Särndal, C. (1995). Methodological principles for a generalized estimation system at statistics canada. **11**, 181–204.

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

Appendix A: Derivations

A.1 Taylor linearization of $\tilde{t}_{yU_Ireg}$

We have

$$\tilde{t}_{yU_Ireg} = \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_FU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{B}}_{\mathbf{x}_F\mathbf{x}_Fs_I}$$

where

$$\begin{aligned}\widehat{\mathbf{B}}_{\mathbf{x}_F\mathbf{x}_Fs_I} &= \left(\widehat{B}_{1s_I}, \ldots, \widehat{B}_{qs_I}, \ldots, \widehat{B}_{Qs_I}\right)' \\ &= \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_Fs_I}\widehat{\mathbf{t}}_{\mathbf{x}_Fys_I} \\ &= \left(\sum\nolimits_{s_I} c_k\mathbf{x}_{Fk}\check{\mathbf{x}}'_{Fk}\right)^{-1} \sum\nolimits_{s_I} c_k\mathbf{x}_{Fk}\check{y}_k\end{aligned}$$

We can write this as

$$\begin{aligned}\tilde{t}_{yU_Ireg} &= \sum\nolimits_{s_I} \check{y}_k + \left(\tilde{\mathbf{t}}_{x_FU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F\mathbf{x}_Fs_I}\widehat{\mathbf{t}}_{\mathbf{x}_Fys_I} \\ &= f\left(\hat{t}_{yU_I\pi}, \hat{\mathbf{t}}_{x_FU_I\pi}, \widehat{\mathbf{T}}_{\mathbf{x}_F\mathbf{x}_Fs_I}, \widehat{\mathbf{t}}_{\mathbf{x}_Fys_I}\right)\end{aligned}$$

Thus, $\tilde{t}_{yU_Ireg}$ is a nonlinear function of the π estimators $\sum_{s_I} \check{y}_k$, $\sum_{s_I} \check{\mathbf{x}}_{Fk}$, $\widehat{\mathbf{T}}_{\mathbf{x}_F\mathbf{x}_Fs_I}$ and $\widehat{\mathbf{t}}_{\mathbf{x}_Fys_I}$, where $\widehat{\mathbf{T}}_{\mathbf{x}_F\mathbf{x}_Fs_I}$ and $\widehat{\mathbf{t}}_{\mathbf{x}_Fys_I}$ has the typical elements

$$\hat{t}_{qq',\pi} = \sum\nolimits_{s_I} c_kx_{Fqk}\check{x}_{Fq'k} = \hat{t}_{q'q,\pi}$$

$$\hat{t}_{q0,\pi} = \sum\nolimits_{s_I} c_kx_{Fqk}\check{y}_k$$

respectively. Using Taylor linearization technique we approximate the nonlinear $\tilde{t}_{yU_Ireg}$ by a linear pseudoestimator. In general, the nonlinear estimator $\hat{\theta}$ is approximated by the linear pseudo estimator $\hat{\theta}_0$ through

$$\hat{\theta} \doteq \hat{\theta}_0 = \theta + \sum_{h=1}^{a} a_h\left(\hat{t}_{h\pi} - t_h\right) \tag{92}$$

where

$$a_h = \left.\frac{\partial f}{\partial \hat{t}_{h\pi}}\right|_{\left(\hat{t}_{1\pi},\ldots,\hat{t}_{H\pi}\right)=(t_1,\ldots t_H)}$$

Letting $\mathbf{B}_{\mathbf{x}_F U_I} = (B_{1U_I}, \ldots, B_{qU_I}, \ldots, B_{QU_I})' = \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} \mathbf{t}_{\mathbf{x}_F y U_I}$, with $\mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} = \sum_{U_I} c_k \mathbf{x}_{Fk} \mathbf{x}'_{Fk}$ and $\mathbf{t}_{\mathbf{x}_F y U_I} = \sum_{U_I} c_k \mathbf{x}_{Fk} y_k$ we have $\theta = t_{yU_I} + \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum_{U_I} \mathbf{x}_{Fk}\right)' \mathbf{B}_{\mathbf{x}_F U_I}$. Furthermore we will need the following partial derivatives:

$$\frac{\partial f}{\partial \sum_{s_I} \check{y}_k} = 1$$

$$\frac{\partial f}{\partial \sum_{s_I} \check{x}_{Fqk}} = -\hat{B}_{qs_I}; \quad q = 1, \ldots, Q$$

$$\begin{aligned} \frac{\partial f}{\partial \hat{t}_{qq',\pi}} &= \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I} \Phi_{qq'} \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I}\right) \widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} \\ &= \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I} \Phi_{qq'} \widehat{\mathbf{B}}_{\mathbf{x}_F \mathbf{x}_F s_I}\right) \end{aligned}$$

$$\frac{\partial f}{\partial \hat{t}_{q0,\pi}} = \left(\tilde{\mathbf{t}}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right) \widehat{\mathbf{T}}^{-1}_{\mathbf{x}_F \mathbf{x}_F s_I} \boldsymbol{\phi}_q$$

where $\Phi_{qq'}$ is a $Q \times Q$ matrix with the value 1 in positions (q, q') and (q', q) and the value 0 everywhere else and $\boldsymbol{\phi}_q$ is a Q-vector with the qth component equal to one and zeros elsewhere.

Evaluating these partial derivatives at the expected value point

$\left(t_{yU_I}, \mathbf{t}_{x_F U_I}, \mathbf{T}_{\mathbf{x}_F\mathbf{x}_F U_I}, \mathbf{t}_{\mathbf{x}_F y U_I}\right)$ and inserting into (92) we obtain

$$\begin{aligned}
\tilde{t}_{yU_I reg} &\doteq t_{yU_I} + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + 1\left(\sum\nolimits_{s_I} \check{y}_k - t_{yU_I}\right) - \\
&\quad - \sum_{q=1}^{Q} B_{qU_i}\left(\sum\nolimits_{s_I} \check{x}_{Fqk} - \sum\nolimits_{U_I} x_{Fqk}\right) - \\
&\quad - \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \sum_{q=1}^{Q}\sum_{q' \leq q} \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I} \Phi_{qq'} \mathbf{B}_{\mathbf{x}_F U_I}\left(\hat{t}_{qq',\pi} - t_{qq'}\right) + \\
&\quad + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \sum_{q=1}^{Q} \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I} \boldsymbol{\phi}_q \left(\hat{t}_{q0,\pi} - t_{q0}\right) \\
&= \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + \\
&\quad - \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I}\left(\widehat{\mathbf{T}}_{\mathbf{x}_F\mathbf{x}_F s_I} - \mathbf{T}_{\mathbf{x}_F\mathbf{x}_F U_I}\right)\mathbf{B}_{\mathbf{x}_F U_I} + \\
&\quad + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I}\left(\widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} - \mathbf{t}_{\mathbf{x}_F y U_I}\right) \\
&= \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_{Fk}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + \\
&\quad + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I}\left(\widehat{\mathbf{t}}_{\mathbf{x}_F y s_I} - \widehat{\mathbf{T}}_{\mathbf{x}_F\mathbf{x}_F s_I}\mathbf{B}_{\mathbf{x}_F U_I}\right) \\
&= \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{s_I} \check{E}_k + \mathbf{t}'_{x_F U_I}\mathbf{B}_{\mathbf{x}_F U_I} + \\
&\quad + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I}\sum\nolimits_{s_I} c_k \mathbf{x}_{Fk}\check{E}_k \\
&= \tilde{\mathbf{t}}'_{x_F U_I}\mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{s_I}\left(1 + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I} c_k \mathbf{x}_{Fk}\right)\check{E}_k \\
&= \sum\nolimits_{s_I} \check{E}_k^{\alpha_I}
\end{aligned}$$

where $E_k^{\alpha_I} = \left(1 + \alpha_I c_k \mathbf{x}_{Fk}\right) E_k$ with $\alpha_I = \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F\mathbf{x}_F U_I}$ and $E_k = y_k - \mathbf{x}'_{Fk}\mathbf{B}_{\mathbf{x}_F U_I}$.
Thus,

$$\tilde{t}_{yU_I reg} \doteq \tilde{\mathbf{t}}'_{x_F U_I}\mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{s_I} \check{E}_k^{\alpha_I}$$

Expected value of $\tilde{t}_{yU_I reg}$

$$E\left(\tilde{t}_{yU_I reg}\right) \doteq \tilde{\mathbf{t}}'_{x_F U_I}\mathbf{B}_{\mathbf{x}_F U_I} + E\left(\sum\nolimits_{s_I} \check{E}_k^{\alpha_I}\right) = \tilde{\mathbf{t}}'_{x_F U_I}\mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{U_I} E_k^{\alpha_I}$$

where

$$\begin{aligned}
\sum\nolimits_{U_I} E_k^{\alpha_I} &= \sum\nolimits_{U_I} \left(1 + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} c_k \mathbf{x}_{Fk}\right) E_k \\
&= \sum\nolimits_{U_I} E_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} \sum\nolimits_{U_I} c_k \mathbf{x}_{Fk} E_k \\
&= \sum\nolimits_{U_I} E_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} \sum\nolimits_{U_I} c_k \mathbf{x}_{Fk} y_k - \\
&\quad - \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{T}^{-1}_{\mathbf{x}_F \mathbf{x}_F U_I} \sum\nolimits_{U_I} c_k \mathbf{x}_{Fk} \mathbf{x}'_{Fk} \mathbf{B}_{\mathbf{x}_F U_I} \\
&= \sum\nolimits_{U_I} E_k
\end{aligned}$$

and we have

$$E\left(\tilde{t}_{yU_I reg}\right) \doteq \tilde{\mathbf{t}}'_{x_F U_I} \mathbf{B}_{\mathbf{x}_F U_I} + \sum\nolimits_{U_I} E_k = \sum\nolimits_{U_I} y_k + \left(\tilde{\mathbf{t}}_{x_F U_I} - \mathbf{t}_{x_F U_I}\right)' \mathbf{B}_{\mathbf{x}_F U_I}$$

A.2 Taylor linearization of $\hat{t}^*_{yUreg}$

We can write

$$\begin{aligned}
\hat{t}^*_{yUreg} &= \sum\nolimits_{s_I} \check{y}_k + \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I} \widehat{\mathbf{t}}_{\mathbf{x}ys_I} + \mathbf{t}'_{uU_C} \widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I} \widehat{\mathbf{t}}_{\mathbf{u}ys_I} = \\
&= f\left(\hat{t}_{yU_I\pi}, \hat{\mathbf{t}}_{xU_I\pi}, \widehat{\mathbf{T}}_{\mathbf{xx}s_I}, \widehat{\mathbf{t}}_{\mathbf{x}ys_I}, \widehat{\mathbf{T}}_{\mathbf{uu}s_I}, \widehat{\mathbf{t}}_{\mathbf{u}ys_I}\right)
\end{aligned}$$

Thus, $\hat{t}^*_{yUreg}$ is a nonlinear function of the π estimators $\sum_{s_I} \check{y}_k$, $\sum_{s_I} \check{\mathbf{x}}_k$, $\widehat{\mathbf{T}}_{\mathbf{xx}s_I}$, $\widehat{\mathbf{t}}_{\mathbf{x}ys_I}$, $\widehat{\mathbf{T}}_{\mathbf{uu}s_I}$ and $\widehat{\mathbf{t}}_{\mathbf{u}ys_I}$ where $\widehat{\mathbf{T}}_{\mathbf{xx}s_I} = \sum_{s_I} c_k \mathbf{x}_k \check{\mathbf{x}}'_k$ and $\widehat{\mathbf{t}}_{\mathbf{x}ys_I} = \sum_{s_I} c_k \mathbf{x}_k \check{y}_k$ has the typical elements

$$\hat{t}_{jj',\pi} = \sum\nolimits_{s_I} c_k x_{jk} \check{x}_{j'k} = \hat{t}_{j'j,\pi}$$

$$\hat{t}_{j0,\pi} = \sum\nolimits_{s_I} c_k x_{jk} \check{y}_k$$

respectively and $\widehat{\mathbf{T}}_{\mathbf{uu}s_I}$ and $\widehat{\mathbf{t}}_{\mathbf{u}ys_I}$ follow analogously by replacing $\mathbf{x}_k$ by $\mathbf{u}_k$ and c_k by h_k with typical elements

$$\hat{t}_{pp',\pi} = \sum\nolimits_{s_I} h_k u_{pk} \check{u}_{p'k} = \hat{t}_{p'p,\pi}$$

$$\hat{t}_{p0,\pi} = \sum\nolimits_{s_I} h_k u_{pk} \check{y}_k$$

We will need the following partial derivatives:

$$\frac{\partial f}{\partial \sum_{s_I} \check{y}_k} = 1$$

$$\frac{\partial f}{\partial \sum_{s_I} \check{x}_{jk}} = -\hat{B}_{js_I}; \quad j = 1, \ldots, J$$

$$\begin{aligned} \frac{\partial f}{\partial \hat{t}_{jj',\pi}} &= \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I} \Phi_{jj'} \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I}\right) \widehat{\mathbf{t}}_{\mathbf{x}ys_I} \\ &= \left(\mathbf{t}_{x_F U_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right)' \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I} \Phi_{jj'} \widehat{\mathbf{B}}_{\mathbf{xx}s_I}\right) \end{aligned} \tag{93}$$

$$\frac{\partial f}{\partial \hat{t}_{j0,\pi}} = \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I} \check{\mathbf{x}}_k\right) \widehat{\mathbf{T}}^{-1}_{\mathbf{xx}s_I} \boldsymbol{\phi}_j \tag{94}$$

$$\begin{aligned} \frac{\partial f}{\partial \hat{t}_{pp',\pi}} &= \mathbf{t}'_{uU_{UC}} \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I} \Psi_{pp'} \widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I}\right) \widehat{\mathbf{t}}_{\mathbf{u}ys_I} \\ &= \mathbf{t}'_{uU_{UC}} \left(-\widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I} \Psi_{pp'} \widehat{\mathbf{B}}_{\mathbf{uu}s_I}\right) \end{aligned}$$

$$\frac{\partial f}{\partial \hat{t}_{p0,\pi}} = \mathbf{t}'_{uU_I} \widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I} \boldsymbol{\psi}_p$$

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, $\boldsymbol{\phi}_j$ is a J-vector with the jth component equal to one and zeros elsewhere and where $\Psi_{pp'}$ is a $P \times P$ matrix with the value 1 in positions (p, p') and (p', p) and the value 0 everywhere else and $\boldsymbol{\psi}_p$ is a P-vector with the pth component equal to one and zeros elsewhere.

Letting $\mathbf{B}_{\mathbf{x}U_I} = \mathbf{T}^{-1}_{\mathbf{xx}U_I} \mathbf{t}_{\mathbf{x}yU_I} = \left(\sum_{U_I} c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_{U_I} c_k \mathbf{x}_k y_k$ and $\mathbf{B}_{\mathbf{u}U_I} = \mathbf{T}^{-1}_{\mathbf{uu}U_I} \mathbf{t}_{\mathbf{u}yU_I} = \left(\sum_{U_I} h_k \mathbf{u}_k \mathbf{u}'_k\right)^{-1} = \sum_{U_I} h_k \mathbf{u}_k y_k$ we have $\theta = t_{yU_I} + \mathbf{t}'_{uU_{UC}} \mathbf{B}_{\mathbf{u}U_I}$. Next, we evaluate the partial derivatives at the expected value point

$\left(t_{yU_I}, \mathbf{t}_{xU_I}, \mathbf{T}_{\mathbf{xx}U_I}, \mathbf{t}_{\mathbf{x}yU_I}, \mathbf{T}_{\mathbf{uu}U_I}, \mathbf{t}_{\mathbf{u}yU_I}\right)$. The partial derivatives given by (93) and (94) conveniently vanish at this point and we obtain

$$\begin{aligned}
\hat{t}^*_{yUreg} &\doteq t_{yU_I} + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + 1\left(\sum\nolimits_{s_I}\check{y}_k - t_{yU_I}\right) - \\
&\quad -\sum_{j=1}^{J} B_{jU_i}\left(\sum\nolimits_{s_I}\check{x}_{jk} - \sum\nolimits_{U_I} x_{jk}\right) - \\
&\quad -\mathbf{t}'_{uU_{UC}}\sum_{p=1}^{P}\sum_{p'\leq p}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\Psi_{pp'}\mathbf{B}_{\mathbf{u}U_I}\left(\hat{t}_{pp',\pi} - t_{pp'}\right) \\
&\quad +\mathbf{t}'_{uU_{UC}}\sum_{p=1}^{P}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\boldsymbol{\psi}_p\left(\hat{t}_{p0,\pi} - t_{p0}\right) \\
&= \sum\nolimits_{s_I}\check{y}_k + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I}\check{\mathbf{x}}_k\right)'\mathbf{B}_{\mathbf{x}U_I} - \\
&\quad -\mathbf{t}'_{uU_{UC}}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\left(\widehat{\mathbf{T}}_{\mathbf{uu}s_I} - \mathbf{T}_{\mathbf{uu}U_I}\right)\mathbf{B}_{\mathbf{u}U_I} \\
&\quad +\mathbf{t}'_{uU_{UC}}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\left(\widehat{\mathbf{t}}_{\mathbf{u}ys_I} - \mathbf{t}_{\mathbf{u}yU_I}\right) \\
&= \sum\nolimits_{s_I}\check{y}_k + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + \left(\mathbf{t}_{xU_I} - \sum\nolimits_{s_I}\check{\mathbf{x}}_k\right)'\mathbf{B}_{\mathbf{x}U_I} + \\
&\quad +\mathbf{t}'_{uU_{UC}}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\left(\widehat{\mathbf{t}}_{\mathbf{u}ys_I} - \widehat{\mathbf{T}}_{\mathbf{uu}s_I}\mathbf{B}_{\mathbf{u}U_I}\right) \\
&= \sum\nolimits_{s_I}\check{E}_{xk} + \mathbf{t}'_{xU_I}\mathbf{B}_{\mathbf{x}U_I} + \mathbf{t}'_{uU_{UC}}\mathbf{T}^{-1}_{\mathbf{uu}U_I}\sum\nolimits_{s_I}h_k\mathbf{u}_k\check{E}_{uk} \\
&\quad +\mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} \\
&= \mathbf{t}'_{xU_I}\mathbf{B}_{\mathbf{x}U_I} + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + \sum\nolimits_{s_I}\check{F}_k
\end{aligned}$$

where $E_{xk} = y_k - \mathbf{x}'_k\mathbf{B}_{\mathbf{x}U_I}, E_{uk} = y_k - \mathbf{u}'_k\mathbf{B}_{\mathbf{u}U_I}$ and $F_k = E_{xk} + \boldsymbol{\alpha}'_u h_k\mathbf{u}_k E_{uk}$ with $\boldsymbol{\alpha}'_u = \mathbf{t}'_{uU_{UC}}\mathbf{T}^{-1}_{\mathbf{uu}U_I}$.

The expected value of $\hat{t}^*_{yUreg}$ is given by

$$\begin{aligned}
E\left(\hat{t}^*_{yUreg}\right) &\doteq \mathbf{t}'_{xU_I}\mathbf{B}_{\mathbf{x}U_I} + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + E\left(\sum\nolimits_{s_I}\check{F}_k\right) \\
&= \mathbf{t}'_{xU_I}\mathbf{B}_{\mathbf{x}U_I} + \mathbf{t}'_{uU_{UC}}\mathbf{B}_{\mathbf{u}U_I} + \sum\nolimits_{U_I}F_k
\end{aligned}$$

where

$$\begin{aligned}
\sum\nolimits_{U_I} F_k &= \sum\nolimits_{U_I} \left(E_{xk} + \mathbf{a}' h_k \mathbf{u}_k E_{uk}\right) \\
&= \sum\nolimits_{U_I} E_{xk} + \mathbf{a}' \sum\nolimits_{U_I} h_k \mathbf{u}_k E_{uk} \\
&= \sum\nolimits_{U_I} E_{xk} + \mathbf{t}'_{uU_C} \mathbf{T}^{-1}_{\mathbf{uu}U_I} \sum\nolimits_{U_I} h_k \mathbf{u}_k y_k - \\
&\quad -\mathbf{t}'_{uU_C} \mathbf{T}^{-1}_{\mathbf{uu}U_I} \sum\nolimits_{U_I} h_k \mathbf{u}_k \mathbf{u}'_k \mathbf{B}_{\mathbf{u}U_I} \\
&= \sum\nolimits_{U_I} E_{xk}
\end{aligned}$$

and we have

$$\begin{aligned}
E\left(\hat{t}^*_{yUreg}\right) &\doteq \mathbf{t}'_{xU_I} \mathbf{B}_{\mathbf{x}U_I} + \mathbf{t}'_{uU_C} \mathbf{B}_{\mathbf{u}U_I} + \sum\nolimits_{U_I} E_{xk} \\
&= \sum\nolimits_{U_I} y_k + +\mathbf{t}'_{uU_C} \mathbf{B}_{\mathbf{u}U_I}
\end{aligned}$$

The approximate variance is given by

$$AV\left(\hat{t}^*_{yUreg}\right) = \sum\sum\nolimits_{U_I} \Delta_{Fkl} \check{F}_k \check{F}_l$$

and a variance estimator is given by

$$\hat{V}\left(\hat{t}^*_{yUreg}\right) = \sum\sum\nolimits_{s_I} \check{\Delta}_{Fkl} \check{f}_k \check{f}_l$$

where $f_k = e_{xk} + \hat{\boldsymbol{\alpha}}'_u h_k \mathbf{u}_k e_{uk}$ with $e_{xk} = y_k - \mathbf{x}'_k \widehat{\mathbf{B}}_{\mathbf{x}s_I}$, $\hat{\boldsymbol{\alpha}}'_u = \mathbf{t}'_{uU_C} \widehat{\mathbf{T}}^{-1}_{\mathbf{uu}s_I}$ and $e_{uk} = y_k - \mathbf{u}'_k \widehat{\mathbf{B}}_{\mathbf{u}s_I}$.

Appendix B: Tables

B.1 Rates of change in the BR

Table B.1.A *Rates of change in the BR: Number of enterprises per size group in November 2000 versus March 2001*

		March 2001							
		0	**1-4**	**5-9**	**10-19**	**20-49**	**50-**	**Deaths**	**Total**
November 2000	**0**	578 297	47	28	17	27	28	22 893	601 337
	1-4	993	143 205	32	13	2	-	2 219	146 464
	5-9	142	85	32 334	93	11	2	381	33 048
	10-19	43	21	96	16 952	131	3	180	17 426
	20-49	14	1	13	82	9 542	120	138	9 910
	50-	9	-	-	2	61	5 865	124	6 061
	Births	20 626	2 493	395	116	72	82	-	23 784
	Total	600 124	145 852	32 898	17 275	9 846	6 100	25 935	

Table B.1.B *Rates of change in the BR: Number of enterprises per size group in March 2001 versus May 2001*

		May 2001							
		0	**1-4**	**5-9**	**10-19**	**20-49**	**50-**	**Deaths**	**Total**
March 2001	**0**	567 183	19 917	1 087	291	89	27	11 530	600 124
	1-4	15 405	121 106	7 176	457	88	4	1 616	145 852
	5-9	629	5 369	22 958	3 456	163	7	316	32 898
	10-19	147	272	2 208	12 886	1 588	25	149	17 275
	20-49	54	66	54	817	8 334	447	74	9 846
	50-	10	-	1	11	210	5 823	45	6 100
	Births	14 640	1 787	230	75	51	29	-	16 812
	Total	598 068	148 517	33 714	17 993	10 523	6 362	13 730	

Table B.1.C *Rates of change in the BR: Number of enterprises per size group in May 2001 versus August 2001*

		August 2001							
		0	**1-4**	**5-9**	**10-19**	**20-49**	**50-**	**Deaths**	**Total**
May 2001	**0**	591 419	27	6	12	6	7	6 591	598 068
	1-4	466	146 923	20	2	3	1	1 102	148 517
	5-9	59	21	33 408	18	4	1	203	33 714
	10-19	31	10	17	17 784	36	4	111	17 993
	20-49	13	4	-	36	10 354	33	83	10 523
	50-	3	1	-	3	21	6 294	40	6 362
	Births	11 421	1 239	157	64	33	17	-	12 931
	Total	603 412	148 225	33 608	17 919	10 457	6 357	8 130	

Table B.1.D *Rates of change in the BR: Number of enterprises per size group in August 2001 versus November 2001*

		November 2001							
		0	**1-4**	**5-9**	**10-19**	**20-49**	**50-**	**Deaths**	**Total**
August 2001	**0**	592 180	1 534	50	18	17	4	9 609	603 412
	1-4	1 893	144 259	293	25	12	2	1 741	148 225
	5-9	195	217	32 675	123	3	-	395	33 608
	10-19	50	16	108	17 463	67	2	213	17 919
	20-49	12	6	5	64	10 202	38	130	10 457
	50-	3	1	3	1	28	6 227	94	6 357
	Births	18 811	1 722	281	69	52	56	-	20 991
	Total	613 144	147 755	33 415	17 763	10 381	6 329	12 182	