Survey Methodology, June 2004
Vol. 30, No. 1, pp. 67-78
Statistics Canada, Catalogue No. 12-001-XPB

Domain Estimation Using Linear Regression

Michael A. Hidiroglou and Zdenek Patak [1]

Abstract

One of the main objectives of a sample survey is the computation of estimates of means and totals for specific domains of interest. Domains are determined either before the survey is carried out (primary domains) or after it has been carried out (secondary domains). The reliability of the associated estimates depends on the variability of the sample size as well as on the *y*-variables of interest. This variability cannot be controlled in the absence of auxiliary information for subgroups of the population. However, if auxiliary information is available, the estimated reliability of the resulting estimates can be controlled to some extent. In this paper, we study the potential improvements in terms of the reliability of domain estimates that use auxiliary information. The properties (bias, coverage, efficiency) of various estimators that use auxiliary information are compared using a conditional approach.

Key Words: Domain estimation; Auxiliary data; Conditional properties.

1. Introduction

One of the main objectives of a sample survey is to compute estimates of means and totals of a number of characteristics associated with the units of a finite population *U*. The data are often used for analytic studies such as the comparison of means and totals for subgroups of the population. Such subgroups are referred to as *domains of study*. Hartley's (1959) paper is one of the first attempts to unify the theory of domain estimation. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimators that did not make use of auxiliary information. He did, however, consider the case of the ratio estimator where population totals were known for the domains. The use of auxiliary data in the context of domain estimation has been discussed in a number of articles. Särndal, Swensson and Wretman (1992) provided a unified treatment of domain estimation with auxiliary data. Estevao, Hidiroglou and Särndal (1995) were the first to recognize that the weights accounting for auxiliary data could be domain dependent or not domain dependent. Estevao and Särndal (1999) discussed desirable properties of regression estimators of domain totals using auxiliary data.

The existence of multivariate auxiliary data raises a number of questions in the context of domain estimation. Some of those questions are as follows. What is the effect of having auxiliary information that is not known on a population basis for the given domain of interest? How do we compute valid variance estimates in the context of domain estimators that use auxiliary data? If more than one estimator is possible for point estimation and/or variance estimation, what criteria should be used to choose the best estimator? Durbin (1969) supported the use of conditional inference to do such comparisons. He stated, "If the sample size is determined by a random mechanism and one happens to get a large sample, one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed". Holt and Smith (1979) favored conditional inference, and applied it to study the properties of the post-stratified estimator, given simple random sampling. Rao (1985) introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined *after* the sample has been drawn. In the case of domain estimation the number of units belonging to a particular domain is a random variable. Recognizable subsets in that context are those where the sample size is fixed within each domain. Comparison of the conditional statistical properties (*i.e.*, bias, mean squared error) of the different estimators can then be based on these subsets. The conditioning process assumes that population totals are known for each domain. In the case of simple random sampling, the number of units in the population domain is assumed known.

The main purpose of this paper is to study the unconditional and conditional properties of a number of domain estimators of totals in the presence of auxiliary data in the context of simple random sampling without replacement (SRSWOR). These conditional properties will be established by conditioning on fixed sample sizes within each domain.

1. Michael Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; Zdenek Patak, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

The paper is organized as follows. In section 2 we will introduce several estimators of domain totals. Their unconditional and conditional properties are provided in section 3. In section 4, we will present the results of a simulation study for the case of the ratio estimator of domain totals, and provide some concluding remarks in section 5.

2. Estimators of Domain Totals

We first introduce some notation to set up the framework, under which we will be assessing the performance of various estimators of domain totals. Let $U = \{1, ..., k, ..., N\}$ denote the finite population. A sample "*s*" is selected from this population using a sampling plan *P*(*s*). Let the first and second order inclusion probabilities be given by π_k and $\pi_{k\ell}$. The domain total $Y_d = \sum_{U_d} y_k$ is the parameter of interest for a variable "*y*". A domain U_d ($d = 1, ..., D$) is any subpopulation of *U*, for which a separate estimate may be required, before or after the planning stage. The number of population units in domain U_d is denoted N_d and $N = \sum_{d=1}^{D} N_d$ for *D* mutually exclusive and exhaustive domains spanning the entire population. The sample *s* is correspondingly divided into *D* domains $s_1, ..., s_d, ..., s_D$ where $s_d = U_d \cap s$. The realized sample size within s_d is a random variable that we denote n_d. Note that the sum of the n_d's over non-overlapping and exhaustive domains of the sample equals n. An estimator of the domain total $Y_d = \sum_{U_d} y_k$ that does not use auxiliary data is given by $\hat{Y}_{d,\mathrm{HT}} = \sum_{s_d} w_k\, y_k = \sum_s w_k\, y_{dk}$ where $w_k = \pi_k^{-1}$, and y_{dk} is equal to y_k if $k \in U_d$ and 0 otherwise.

Auxiliary information in the form of a *p*-dimensional vector **x** may be available at different levels of aggregation. It may be known for each unit in the population, or for subsets $U_g \subseteq U\,(g = 1, ..., G)$ of the population *U* that may coincide with the domains U_d. We denote such known totals $\mathbf{X}_g = \sum_{U_g} \mathbf{x}_k$; they are estimated by $\hat{\mathbf{X}}_{g,\mathrm{HT}} = \sum_{s_g} w_k\, \mathbf{x}_k$. A modified set of weights $\tilde{w}_k$ incorporating the auxiliary data can be computed using either calibration or linear regression procedures (LR). We chose the LR approach. In the case of *G* population groups, the LR estimator is given by

$$\hat{Y}_{\ell r} = \hat{Y}_{\mathrm{HT}} + \sum_{g=1}^{G} (\mathbf{X}_g - \hat{\mathbf{X}}_{g,\mathrm{HT}})' \hat{\mathbf{B}}_g \qquad (1.1)$$

where $\hat{\mathbf{B}}_g = (\sum_{s_g} w_k\, \mathbf{x}_k\, \mathbf{x}_k' / c_k)^{-1} \sum_{s_g} w_k\, \mathbf{x}_k\, y_k / c_k$, and c_k are suitable positive constants. The use of auxiliary data in the domain context offers a wide range of choices for various levels at which auxiliary totals are used and regression models are constructed. To simplify matters, we assume that $g = 1$ (*e.g.*: a single group *U*), yielding the simple regression estimator $\hat{Y}_{\ell r} = \hat{Y}_{\mathrm{HT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}})' \hat{\mathbf{B}}$, where $\hat{\mathbf{X}}_{\mathrm{HT}} = \sum_s w_k \mathbf{x}_k$.

We consider six estimators for estimating the domain population total Y_d. These estimators are based on whether we use the domain totals $\mathbf{X}_d$ or the population total $\mathbf{X}$, and whether we construct the regression estimator at the domain or at the population levels. The estimators are categorized into Horvitz-Thompson and "Hájek" types. We provide an example of the *ratio* estimator that is associated with each of these estimators.

2.1 Horvitz-Thompson Type Estimators

Case 1

We assume that the auxiliary information $\mathbf{x}_k$ is available at the population level *U*, $\mathbf{X} = \sum_U \mathbf{x}_k$ and that the domain specific y_{dk} variables are regressed on $\mathbf{x}_k$, $k \in U$. The resulting population regression parameter $\mathbf{B}_{1d} = (\sum_U \mathbf{x}_k\, \mathbf{x}_k' / c_k)^{-1} \sum_U \mathbf{x}_k\, y_{dk} / c_k$ is estimated by $\hat{\mathbf{B}}_{1d} = (\sum_s w_k \mathbf{x}_k\, \mathbf{x}_k' / c_k)^{-1} \sum_s w_k \mathbf{x}_k\, y_{dk} / c_k$ and the resulting estimator of the population total Y_d is

$$\hat{Y}_{d,\ell r_1} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}})' \hat{\mathbf{B}}_{1d}. \qquad (2.1)$$

Example: The domain ratio estimator given by $\hat{Y}_{d,\mathrm{RAT}} = X\hat{R}_{1d}$, where $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$. This estimator was first suggested by Hidiroglou (1991), and is discussed in more detail in Estevao *et al.* (1995).

If the auxiliary data totals are available at the domain level, $\mathbf{X}_d = \sum_{U_d} \mathbf{x}_k$, then two possible estimators of Y_d (cases 2 and 3) can be constructed, depending on how the population regression parameter is estimated.

Case 2

The population regression parameter

$$\mathbf{B}_{2d} = \left(\sum_{U_d} \mathbf{x}_k\, \mathbf{x}_k' / c_k \right)^{-1} \sum_{U_d} (\mathbf{x}_k\, y_k / c_k)$$

is estimated by regressing y_k on $\mathbf{x}_k$ for each domain U_d separately. Its estimator is given by

$$\hat{\mathbf{B}}_{2d} = \left(\sum_{s_d} w_k\, \mathbf{x}_k\, \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_d} \left(w_k\, \mathbf{x}_k\, y_k / c_k \right),$$

and the resulting regression estimator of a domain total is

$$\hat{Y}_{d,\ell r_2} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HT}})' \hat{\mathbf{B}}_{2d} \qquad (2.2)$$

where $\hat{\mathbf{X}}_d = \sum_s w_k\, \mathbf{x}_{dk}$ with $\mathbf{x}_{dk}$ defined similarly to y_{dk}.

Example: The Horvitz-Thompson post-stratified estimator given by $\hat{Y}_{d,\mathrm{POSTR}} = X_d \hat{R}_{2d}$, where $\hat{R}_{2d} = \hat{Y}_{d,\mathrm{HT}} / \hat{X}_{d,\mathrm{HT}}$.

Case 3

The population regression parameter

$$\mathbf{B}_3 = \left(\sum_U \left(\mathbf{x}_k \ \mathbf{x}_k'/c_k\right)\right)^{-1} \sum_U \left(\mathbf{x}_k \ y_k/c_k\right)$$

is estimated by regressing y_k on $\mathbf{x}_k$ using all units in U. The corresponding estimator is

$$\hat{\mathbf{B}}_3 = \left(\sum_s \ w_k \ \mathbf{x}_k \ \mathbf{x}_k'/c_k\right)^{-1} \sum_s \left(w_k \ \mathbf{x}_k \ y_k/c_k\right),$$

resulting in the regression estimator

$$\hat{Y}_{d,\ell r_3} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HT}})'\hat{\mathbf{B}}_3. \qquad (2.3)$$

Example: The alternate ratio estimator given by $\hat{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HT}} + (X_d - \hat{X}_{d,\mathrm{HT}})\hat{R}_3$, where $\hat{R}_3 = \hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$.

2.2 Hájek Type Estimators

Estimators (2.1)-(2.3) belong to the Horvitz-Thompson family. If the known population domain size N_d is also incorporated in the estimation, then we get the "Hájek" versions of the previously defined Horvitz-Thompson regression estimators. The Hájek regression estimators are obtained by replacing $\hat{Y}_{d,\mathrm{HT}}$, $\hat{\mathbf{X}}_{d,\mathrm{HT}}$, and $\hat{\mathbf{X}}_{\mathrm{HT}}$ by

$$\hat{Y}_{d,\mathrm{HA}} = \left(N_d / \hat{N}_d\right)\hat{Y}_{d,\mathrm{HT}}, \quad \hat{\mathbf{X}}_{d,\mathrm{HA}} = \left(N_d / \hat{N}_d\right)\hat{\mathbf{X}}_{d,}$$

and

$$\hat{\mathbf{X}}_{\mathrm{HA}} = \left(N/\hat{N}\right)\hat{\mathbf{X}}_{\mathrm{HT}},$$

where $\hat{N}_d = \sum_{s_d} w_k$ and $\hat{N} = \sum_s w_k$. The estimators are nearly conditionally unbiased for a given n_d, whereas their Horvitz-Thompson counterparts do not have this property. The "$\hat{\mathbf{B}}$"s contained within the Hájek regression estimators correspond exactly to their Horvitz-Thompson counterparts.

Case 4

$$\tilde{Y}_{d,\ell r_1} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HA}})'\hat{\mathbf{B}}_{1d}. \qquad (2.4)$$

Example: The Hájek ratio estimator given by $\tilde{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HA}} + (X - \hat{X}_{\mathrm{HA}})\hat{R}_{1d}$.

Case 5

$$\tilde{Y}_{d,\ell r_2} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HA}})'\hat{\mathbf{B}}_{2d}. \qquad (2.5)$$

Example: The Hájek post-stratified ratio estimator given by $\tilde{Y}_{d,\mathrm{POSTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\hat{R}_{2d}$. This estimator is identical to the Horvitz-Thompson post-stratified estimator.

Case 6

$$\tilde{Y}_{d,\ell r_3} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HA}})'\hat{\mathbf{B}}_3. \qquad (2.6)$$

Example: The Hájek alternate ratio estimator given by $\tilde{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\hat{R}_3$.

3. Properties of the Domain Estimators

Estimators (2.1)–(2.6) may be expressed as:

$$\hat{Y}_{d,\ell r} = \sum_s \ w_k \ a_{dk} \ y_{dk} = \sum_s \ \tilde{w}_{dk} \ y_{dk} \qquad (2.7)$$

where a_{dk} is an adjustment factor that may or may not be domain dependent. The product of the design weight w_k and the adjustment factor a_{dk} is known as the regression weight (or calibration weight) $\tilde{w}_{dk}$. Tables 1 and 2 provide a summary of these factors, as well as the residuals required for unconditional variance estimation. The population and sample residuals are denoted as E_{dk} and e_{dk}. The indicator variable δ_{dk} is equal to one if $k \in U_d$ and zero otherwise.

The approximate population and corresponding estimated variances of the Horvitz-Thompson estimators $\hat{Y}_{d,\ell r_j}$ $(j = 1, 2, 3)$ are:

$$V\left(\hat{Y}_{d,\ell r_j}\right) = \sum\sum_U \ \Delta_{k\ell} \left(\frac{E_{dk}}{\pi_k}\right)\left(\frac{E_{d\ell}}{\pi_\ell}\right) \qquad (2.8)$$

and

$$\mathrm{v}\left(\hat{Y}_{d,\ell r_j}\right) = \sum\sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \left(\frac{a_{dk}\, e_{dk}}{\pi_k}\right)\left(\frac{a_{d\ell}\, e_{d\ell}}{\pi_\ell}\right) \qquad (2.9)$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \ \pi_\ell$; $\pi_{k\ell} = \Pr\{k, \ell \in s\}$ with the appropriate E_{dk}'s, e_{dk}'s, a_{dk}'s defined in Table 1.

The approximate unconditional population and correspondding estimated variances of the Hájek-type estimators $\tilde{Y}_{d,\ell r_j}$ $(j = 1, 2, 3)$ are:

$$V\left(\tilde{Y}_{d,\ell r_j}\right) =$$

$$\begin{cases} \sum\sum_U \ \Delta_{k\ell} \left(\dfrac{E_{dk} - \left(\sum_U \ E_{dk} \ /N_d\right)\delta_{dk}}{\pi_k}\right) \times \\ \qquad \left(\dfrac{E_{d\ell} - \left(\sum_U \ E_{dk} \ /N_d\right)\delta_{d\ell}}{\pi_\ell}\right) \\ \qquad\qquad \text{for j} = 1 \\ \sum\sum_{U_d} \ \Delta_{k\ell} \left(\dfrac{E_{dk} - \tilde{E}_{U_d}}{\pi_k}\right)\left(\dfrac{E_{d\ell} - \tilde{E}_{U_d}}{\pi_\ell}\right) \\ \qquad\qquad \text{for j} = 2, 3 \end{cases} \qquad (2.10)$$

Table 1
Adjustment Factors and Residuals for Horvitz-Thompson Regression Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\hat{Y}_{d,\ell r_1}$	No	$1+(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HT}})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_k'\,\mathbf{B}_{1d}$ $e_{dk}=y_{dk}-\mathbf{x}_k'\,\hat{\mathbf{B}}_{1d}$
$\hat{Y}_{d,\ell r_2}$	Yes	$\delta_{dk}\left(1+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HT}})'\left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right)$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\mathbf{B}_{2d}$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\hat{\mathbf{B}}_{2d}$
$\hat{Y}_{d,\ell r_3}$	Yes	$\delta_{dk}+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HT}})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_3$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\hat{\mathbf{B}}_3$

Table 2
Adjustment Factors and Residuals for the Hájek-type Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\tilde{\tilde{Y}}_{d,\ell r_1}$	Yes	$\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HA}}\right)'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_k'\mathbf{B}_{1d}$ $e_{dk}=y_{dk}-\mathbf{x}_k'\,\hat{\mathbf{B}}_{1d}$
$\tilde{\tilde{Y}}_{d,\ell r_2}$	Yes	$\delta_{dk}\left(\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HA}}\right)'\left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right)$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\mathbf{B}_{2d}$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\hat{\mathbf{B}}_{2d}$
$\tilde{\tilde{Y}}_{d,\ell r_3}$	Yes	$\delta_{dk}\,\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HA}}\right)'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_3$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\,\hat{\mathbf{B}}_3$

and

$$\mathrm{v}\left(\tilde{\tilde{Y}}_{d,\ell r_j}\right)=\sum\sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}}\left(\frac{a_{dk}\,e_{dk}}{\pi_k}\right)\left(\frac{a_{d\ell}\,e_{d\ell}}{\pi_\ell}\right)$$

$$\text{for } j=1,2,3 \qquad (2.11)$$

where $\tilde{\tilde{E}}_{U_d}=\sum_{U_d} E_{dk}/N_d$. The appropriate E_{dk}'s, e_{dk}'s, and a_{dk}'s are defined in Table 2. Note that the form of the estimated unconditional variance is the same for both the Horvitz-Thompson and the Hájek-type estimators.

Result 3.1: The Hájek-type regression estimator can be obtained as a by-product of the regression of y_k on

$$(\mathbf{x}_k^{\circ})'=\left(1,\left(\mathbf{x}_k-\bar{\mathbf{x}}_U\right)'\right),$$

where $\bar{\mathbf{x}}_U=N^{-1}\sum_U \mathbf{x}_k$. The resulting regression vector is

$$\hat{\mathbf{B}}^{\circ\,\prime}=\left(\hat{B}_1^{\circ},\hat{\mathbf{B}}_x'\right),$$

where

$$\hat{\mathbf{B}}_x=\left(\left(\sum_s w_k\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)'/c_k\right)\right)^{-1}\times \sum_s\left(w_k\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)y_k/c_k\right)$$

and $\hat{B}_1^{\circ}=\tilde{y}_s+\left(\bar{\mathbf{x}}_U-\tilde{\mathbf{x}}_s\right)'\hat{\mathbf{B}}_x$, with $\tilde{y}_s=\hat{Y}_{\mathrm{HT}}/\hat{N}$ and $\tilde{\mathbf{x}}_s=\hat{\mathbf{X}}_{\mathrm{HT}}/\hat{N}$.

The regression estimator of total $\hat{Y}_{\ell r}=N\hat{B}_1^{\circ}$ is equal to the Hájek form $\tilde{\tilde{Y}}_{\ell r}=\hat{Y}_{\mathrm{HA}}+(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HA}})'\hat{\mathbf{B}}_x$. The various Hájek-type domain regression estimators can be obtained using this approach. For instance, regressing y_{dk} on

$$(\mathbf{x}_k^{\circ})'=\left(1,\left(\mathbf{x}_{dk}-\bar{\mathbf{x}}_{U_d}\right)'\right)$$

yields $\tilde{\tilde{Y}}_{d,\ell r_1}$.

Proof: We first show how to arrive at the Hájek form of the regression estimator. Defining the auxiliary data vector $\mathbf{z}_k$ as $\mathbf{z}_k'=(x_{0k},\,\mathbf{x}_k')'$, the regression estimator is

$$\hat{\mathrm{Y}}_{\ell r}=\hat{Y}_{\mathrm{HT}}+\left(\mathbf{Z}-\hat{\mathbf{Z}}_{\mathrm{HT}}\right)'\hat{\mathbf{B}}_{\mathbf{z}}$$

where

$$\hat{\mathbf{B}}_{\mathbf{z}}=\left(\sum_s w_k\mathbf{z}_k\mathbf{z}_k'/c_k\right)^{-1}\left(\sum_s w_k\mathbf{z}_k y_k/c_k\right),$$

$\mathbf{Z}=\sum_U \mathbf{z}_k$ and $\hat{\mathbf{Z}}_{\mathrm{HT}}=\sum_s w_k\mathbf{z}_k$.

If $x_{0k}=1$, $\hat{\mathrm{Y}}_{\ell r}$ is exactly equivalent to $\hat{\mathrm{Y}}_{\ell r}=\mathbf{Z}'\,\hat{\mathbf{B}}_{\mathbf{z}}$. Decomposing $\hat{\mathbf{B}}_{\mathbf{z}}$ as

$$\hat{\mathbf{B}}_{\mathbf{z}}'=\left(\hat{B}_0,\hat{\mathbf{B}}_x'\right),$$

we have that $\hat{\mathrm{Y}}_{\ell r}=N\hat{B}_0+\sum_U \mathbf{x}_k'\hat{\mathbf{B}}_x$, where $\hat{B}_0=\tilde{y}_s-\tilde{\mathbf{x}}_s'\hat{\mathbf{B}}_x$ and

$$\hat{\mathbf{B}}_x=\left(\frac{\sum_s w_k\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)'}{c_k}\right)^{-1}\times \frac{\sum_s w_k\left(\mathbf{x}_k-\tilde{\mathbf{x}}_s\right)\left(\mathrm{y}_k-\tilde{\mathrm{y}}_s\right)}{c_k}.$$

Hence, the Hájek form of the regression estimator is

$$\tilde{\mathrm{Y}}_{\ell r} = \hat{\mathrm{Y}}_{\mathrm{HA}} + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\mathrm{HA}}\right)' \hat{\mathbf{B}}_x .$$

Regressing y_k on

$$(\mathbf{x}_k^{\circ})' = \left(1, (\mathbf{x}_k - \bar{\mathbf{x}}_U)'\right)$$

yields the estimated regression vector $\hat{\mathbf{B}}^{\circ\prime} = \left(\hat{B}_1^{\circ}, \hat{\mathbf{B}}_x'\right)$, where

$$\hat{\mathbf{B}}_x = \left(\frac{\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s)(\mathbf{x}_k - \tilde{\mathbf{x}}_s)}{c_k}\right)^{-1} \sum_s \frac{w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) y_k}{c_k}$$

and $\hat{B}_1^{\circ} = \tilde{y}_s + (\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_s)' \hat{\mathbf{B}}_x$. Substituting $\hat{B}_1^{\circ}$ into $\hat{Y}_{\ell r} = N\hat{B}_1^0$ yields the Hájek form $\tilde{\mathrm{Y}}_{\ell r}$.

Remark 3.1: (Additivity). Suppose that the domains U_d are mutually exclusive ($U_{d_1} \cap U_{d_2} = \emptyset$ for $d_1 \neq d_2$) and exhaustive ($\cup_{d=1}^{D} U_d = U$). Additivity over such domains means that $\sum_{d=1}^{D} \hat{Y}_{d,\ell r_1} = \sum_{d=1}^{D} \hat{Y}_{d,\ell r_3} = \hat{Y}_{\ell r}$ where

$$\hat{Y}_{\ell r} = \hat{Y}_{\mathrm{HT}} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}}\right)' \hat{\mathbf{B}}.$$

The additive property of $\hat{Y}_{d,\ell r_1}$ is desirable because a single set of calibration weights, $w_k\ a_{dk}$, can be used repeatedly to produce ad hoc domain estimates. Only two out of the six estimators, $\hat{Y}_{d,\ell r_1}$ and $\hat{Y}_{d,\ell r_3}$, are additive over all such domains.

Remark 3.2: (Calibrating on domain auxiliary data). Estevao *et al.* (1999) discussed some of the estimators provided in Tables 1 and 2 for the case of a single auxiliary variable x_k. They arrived at their estimators by controlling on domain information, either via auxiliary variables and /or control totals.

In what follows, we will assume that the sample *s* of size *n* has been selected using simple random sampling without replacement (SRSWOR) from a universe of size *N*. The estimated unconditional variance of the Horvitz-Thompson and Hájek-type estimators for this sampling plan is:

$$\mathrm{v}\left(\hat{Y}_{d,\ell r_j}\right) =$$

$$\mathrm{v}\left(\tilde{Y}_{d,\ell r_j}\right) = \sum_s \frac{N^2(1-f)}{n} \frac{\sum_s \left(a_{dk}\, e_{dk} - \overline{a_d\, e}\right)^2}{n-1} \quad (2.12)$$

where $\overline{a_d\, e} = \sum_s \left(a_{dk}\, e_{dk}/n\right)$ and $f = n/N$ is the sampling fraction.

3.1 Unconditional Properties

The choice between the various regression estimators should be based on the level at which the auxiliary totals are available, as well as bias and variance. All the above estimators are asymptotically unconditionally unbiased; however, their variances differ. We compare the unconditional population variances of the six domain regression estimators (2.1) – (2.6) by distinguishing two cases: (i) an intercept term is included in the regression; and (ii) no intercept term is included in the regression.

Result 3.2: Assume that an intercept is included in the regression, $c_k = c$ for all $k \in U$, and $N > p$, where *p* refers to the number of auxiliary variables. The following inequalities hold for the population variances of the domain regression estimators (2.1) – (2.6):

(i) $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$; $V(\hat{Y}_{d,\ell r_2}) \leq V(\hat{Y}_{d,\ell r_3})$; $V(\hat{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d,\ell r_1})$.

(ii) $V(\tilde{Y}_{d,\ell r_2}) < V(\tilde{Y}_{d,\ell r_1})$ and $V(\tilde{Y}_{d,\ell r_2}) < V(\tilde{Y}_{d,\ell r_3})$; $V(\tilde{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\tilde{Y}_{d,\ell r_1})$.

Proof: In the case of simple random sampling without replacement, $V(\hat{Y}_{d,\ell r_j}) = A\sum_U (E_{dk} - \bar{E}_{U_d})^2$ for $\ell = 1, 2, 3$, where $A = N^2(1-f)/(n(N-1))$ and $\bar{E}_{U_d} = \sum_{U_d} E_{dk}/N$. Given that the regression contains an intercept, it follows that $\sum_U E_{dk} = 0$ or that $\sum_{U_d} E_{dk} = 0$, depending on which regression estimator we use. We only show that (i) holds: the proof for (ii) is similar. The population variances for $\hat{Y}_{d,\ell r_1}$ and $\hat{Y}_{d,\ell r_2}$ are respectively

$$V\left(\hat{Y}_{d,\ell r_1}\right) = A\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2$$

and

$$V\left(\hat{Y}_{d,\ell r_2}\right) = A\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d})^2 .$$

The population variance of $\hat{Y}_{d,\ell r_3}$ is

$$V\left(\hat{Y}_{d,\ell r_3}\right) = A\sum_U (E_{dk} - \tilde{E}_{U_d})^2 ,$$

where

$$\tilde{E}_{U_d} = N^{-1}\sum_U E_{dk} = \left(\frac{N_d}{N}\right)\left(\bar{y}_{U_d} - \bar{\mathbf{x}}_{U_d}' \mathbf{B}_3\right),$$

with $\bar{y}_{U_d} = N_d^{-1}\sum_{U_d} y_{dk}$ and $\bar{\mathbf{x}}_{U_d}'$ similarly defined.

We first show that $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$. To this end, we decompose $\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2$ into its within domain U_d and outside domain $U_{\bar{d}}$ components, yielding

$$\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 = \sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 + \sum_{U_{\bar{d}}} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 .$$

Since

$$\sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 = \sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d})^2 + \sum_{U_d} \left(\mathbf{x}_k' (\mathbf{B}_{2d} - \mathbf{B}_{1d})\right)^2 ,$$

it follows that $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$.

Next, we show that

$$V\left(\hat{Y}_{d,\ell r_2}\right) \leq V\left(\hat{Y}_{d,\ell r_3}\right).$$

The variance $V(\hat{Y}_{d,\ell r_3})$ can be re-expressed as

$$V\left(\hat{Y}_{d,\ell r_3}\right)=\sum_{U_d}\left(\begin{array}{l}(y_k-\mathbf{x}_k'\mathbf{B}_3)^2\\ \quad-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}-\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)^2\end{array}\right).$$

The difference between $V(\hat{Y}_{d,\ell r_3})$ and $V(\hat{Y}_{d,\ell r_2})$ is:

$$\begin{aligned}&V\left(\hat{Y}_{d,\ell r_3}\right)-V\left(\hat{Y}_{d,\ell r_2}\right)\\ &=A\left\{\begin{array}{l}\sum_{U_d}\left((y_k-\mathbf{x}_k'\mathbf{B}_3)^2\right)-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}-\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)^2\\ \qquad\qquad-\sum_{U_d}(y_k-\mathbf{x}_k'\mathbf{B}_{2d})^2\end{array}\right\}\\ &=A\left\{\begin{array}{l}(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\\ \quad-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)\end{array}\right\}\\ &\geq A\left\{\begin{array}{l}(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\\ \quad-N_d\left(\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)\end{array}\right\}.\end{aligned}$$

Noting that $\bar{y}_{U_d}=\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}$ it follows that:

$$\begin{aligned}&\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}'\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\\ &=\mathbf{B}_{2d}'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}\\ &\qquad+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\\ &=(\mathbf{B}_3-\mathbf{B}_{2d})'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'(\mathbf{B}_3-\mathbf{B}_{2d})\end{aligned}$$

Since

$$\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'-N_d\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'=\sum_{U_d}\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)',$$

the difference $V(\hat{Y}_{d,\ell r_3})-V(\hat{Y}_{d,\ell r_2})$ can be expressed as:

$$\begin{aligned}&V\left(\hat{Y}_{d,\ell r_3}\right)-V\left(\hat{Y}_{d,\ell r_2}\right)\\ &=A\left\{(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'-N_d\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\right\}\\ &=A\left\{(\mathbf{B}_3-\mathbf{B}_{2d})'\sum_{U_d}\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)'(\mathbf{B}_3-\mathbf{B}_{2d})\right\}\\ &\geq 0.\end{aligned}$$

Finally, we show that $V(\hat{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d,\ell r_1})$ by constructing examples:

(i) $V(\hat{Y}_{d,\ell r_3}) < V(\hat{Y}_{d,\ell r_1})$, if $\mathbf{B}_3=\mathbf{B}_{2d}$;

(ii) $V(\hat{Y}_{d,\ell r_3}) = V(\hat{Y}_{d,\ell r_1})$, if $\mathbf{B}_3=\mathbf{B}_{1d}$;

(iii) $V(\hat{Y}_{d,\ell r_3}) > V(\hat{Y}_{d,\ell r_1})$, if the fit of y_k on $\mathbf{x}_k$ is much poorer than the fit y_{dk} on $\mathbf{x}_k$ for $k\in U$.

It can also be shown that $V(\tilde{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_2})$; $V(\tilde{Y}_{d,\ell r_3}) < V(\hat{Y}_{d,\ell r_3})$; and $V(\tilde{Y}_{d,\ell r_1}) < V(\hat{Y}_{d,\ell r_1})$. The estimator with the smallest variance is $\tilde{Y}_{d,\ell r_2}$. However, if it is assumed that the $\hat{\mathbf{B}}_{2d}$'s are similar across all domains, and that there are very few observations in s_d, it may be preferable to use $\tilde{Y}_{d,\ell r_3}$. The choice between $\tilde{Y}_{d,\ell r_2}$ and $\tilde{Y}_{d,\ell r_3}$ should not always be based on the asymptotic variance. If there are very few observations in s_d, this can cause significant bias in $\tilde{Y}_{d,\ell r_2}$ and also cause the exact variance of $\tilde{Y}_{d,\ell r_2}$ to be larger than that of $\tilde{Y}_{d,\ell r_3}$, so that the latter may be preferred.

Remark 3.3: If there is no intercept in the regression, then it does not necessarily follow that Result 3.2 holds.

Proof: We illustrate this statement using the elementary ratio versions of cases 1 and 2. They are respectively the *Horvitz-Thompson ratio* estimator $\hat{Y}_{d,\mathrm{RAT}}=\hat{Y}_{d,\mathrm{HT}}(X/\hat{X}_{\mathrm{HT}})$ and the *Horvitz-Thompson post-stratified ratio* estimator $\hat{Y}_{d,\mathrm{POSTR}}=\hat{Y}_{d,\mathrm{HT}}(X_d/\hat{X}_{d,\mathrm{HT}})$. Also, suppose that the elements of the data vector (y_k,x_k) are positive for all $k\in U$. The population variances for $\hat{Y}_{d,\mathrm{RAT}}$ and $\hat{Y}_{d,\mathrm{POSTR}}$ are $V(\hat{Y}_{\mathrm{d,POSTR}})=A\sum_{U_d}(y_k-B_{2d}x_k)^2$ and $V(\hat{Y}_{d,\mathrm{RAT}})=A\sum_U(y_{dk}-B_{1d}x_k)^2$, where $B_{2d}=Y_d/X_d$, and $B_{1d}=Y_d/X$.

The difference $V(\hat{Y}_{d,\mathrm{RAT}})-V(\hat{Y}_{d,\mathrm{POSTR}})$ can be re-expressed as:

$$\begin{aligned}&A\sum_{U_d}(B_{1d}-B_{2d})^2x_k^2\\ &\quad+2A(B_{1d}-B_{2d})\sum_{U_d}(y_k-B_{2d}x_k)x_k\\ &\quad+A\sum_{U_{\bar{d}}}(y_{dk}-B_{1d}x_k)^2.\end{aligned}$$

Since the second term of this expression can be positive, negative or zero, the difference $V(\hat{Y}_{d,\mathrm{RAT}})-V(\hat{Y}_{d,\mathrm{POSTR}})$ can be negative.

3.2 Conditional Properties

For a given sample *s*, let n_d be the realized sample size of s_d. The following result can be used to evaluate the conditional bias of estimators (2.1) to (2.6).

Result 3.3: Let $\mathbf{z}_k$ be an arbitrary *p*-dimensional vector, that is $\mathbf{z}_k=(z_{k1},...,z_{kp})'$, and suppose that $n_d\geq 1$. The conditional expectation of $\bar{\mathbf{z}}_s=n^{-1}\sum_s\mathbf{z}_k$ given n_d can be written as:

$$\begin{aligned}E(\bar{\mathbf{z}}_s\mid n_d)&=\frac{1}{n}\left[f_d\sum_{U_d}\mathbf{z}_k+f_{\bar{d}}\left(\sum_U\mathbf{z}_k-\sum_{U_d}\mathbf{z}_k\right)\right]\\ &=\bar{\mathbf{z}}_U+\frac{w_d-W_d}{1-W_d}\left(\tilde{\mathbf{z}}_{U_d}-\bar{\mathbf{z}}_U\right)\end{aligned}\qquad(3.1)$$

where $\bar{\mathbf{z}}_U=N^{-1}\sum_U\mathbf{z}_k$, $\tilde{\mathbf{z}}_{U_d}=N_d^{-1}\sum_{U_d}\mathbf{z}_k$, $w_d=n_d/n$, $W_d=N_d/N$, $f_d=n_d/N_d$, $f_{\bar{d}}=n_{\bar{d}}/N_{\bar{d}}$ with $n_{\bar{d}}=n-n_d$, and $N_{\bar{d}}=N-N_d$.

Proof: Rewriting $\bar{\mathbf{z}}_s$ as

$$\frac{1}{n}\left(\sum_{s_d}\mathbf{z}_k+\sum_{s_{\bar{d}}}\mathbf{z}_k\right),$$

we have that

$$E(\bar{\mathbf{z}}_s\mid n_d)=\frac{1}{n}\left[\frac{n_d}{N_d}\sum_{U_d}\mathbf{z}_k+\frac{n-n_d}{N-N_d}\sum_{U_{\bar{d}}}\mathbf{z}_k\right]$$

where $s_{\bar{d}} = \{ k \in s \text{ and } k \notin s_d \}$ and

$$U_{\bar{d}} = \{ k \in U \text{ and } k \notin U_d \}.$$

Since $\sum_{U_{\bar{d}}} \mathbf{z}_k = \sum_U \mathbf{z}_k - \sum_{U_d} \mathbf{z}_k$, we obtain the required result, that is

$$E(\bar{\mathbf{z}}_s \mid n_d) = \bar{\mathbf{z}}_U + \frac{w_d - W_d}{1 - W_d}(\tilde{\mathbf{z}}_{U_d} - \bar{\mathbf{z}}_U).$$

Result 3.4: The conditional population variance of $\bar{\mathbf{z}}_s$ given n_d, can be written as

$$V(\bar{\mathbf{z}}_s \mid n_d) = \frac{w_d^2}{n_d}(1 - f_d)\mathbf{V}_{z_{U_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1 - f_{\bar{d}})\mathbf{V}_{z_{U_{\bar{d}}}},$$

where

$$\mathbf{V}_{z_{U_d}} = \frac{1}{N_d - 1}\sum_{U_d}(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_d})(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_d})',$$

$$\mathbf{V}_{z_{U_{\bar{d}}}} = \frac{1}{N_{\bar{d}} - 1}\sum_{U_{\bar{d}}}(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_{\bar{d}}})(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_{\bar{d}}})',$$

with $\tilde{\mathbf{z}}_{U_{\bar{d}}} = N_{\bar{d}}^{-1}\sum_{U_{\bar{d}}} \mathbf{z}_k$, and $w_{\bar{d}} = 1 - w_d$.

The estimator of the conditional population variance $\mathbf{V}(\bar{\mathbf{z}}_s \mid n_d)$ is given by

$$\mathbf{v}(\bar{\mathbf{z}}_s \mid n_d) = \frac{w_d^2}{n_d} f_d \, \mathbf{v}_{z_{s_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1 - f_{\bar{d}})\mathbf{v}_{z_{s_{\bar{d}}}},$$

where

$$\mathbf{v}_{z_{s_d}} = \frac{1}{n_d - 1}\sum_{s_d}(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_d})(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_d})'$$

and

$$\mathbf{v}_{z_{s_{\bar{d}}}} = \frac{1}{n_{\bar{d}} - 1}\sum_{s_{\bar{d}}}(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_{\bar{d}}})(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_{\bar{d}}})',$$

with $\tilde{\mathbf{z}}_{s_d} = n_d^{-1}\sum_{s_d} \mathbf{z}_k$, $\tilde{\mathbf{z}}_{s_{\bar{d}}} = n_{\bar{d}}^{-1}\sum_{s_{\bar{d}}} \mathbf{z}_k$.

Proof: It follows using arguments similar to those used in Result 3.3. We first illustrate how Result 3.3 can be used to obtain the conditional bias for the simpler estimators of domain totals. This includes the Horvitz-Thompson estimator $\hat{Y}_{d,\mathrm{HT}}$, as well as post-stratified ratio estimator $\hat{Y}_{d,\mathrm{POSTR}} = (X_d / \hat{X}_{d,\mathrm{HT}})\hat{Y}_{d,\mathrm{HT}}$. Let $\mathbf{z}_k$ be the domain variable y_{dk}. Using Result 3.3, we have that $E(\hat{Y}_{d,\mathrm{HT}} \mid n_d) = N w_d \tilde{y}_{U_d}$, where $\tilde{y}_{U_d} = Y_d / N_d$. The conditional bias of $\hat{Y}_{d,\mathrm{HT}}$ given n_d is therefore $\mathrm{Bias}(\hat{Y}_{d,\mathrm{HT}} \mid n_d) = N(w_d - W_d)\tilde{y}_{U_d}$. For the post-stratified ratio estimator, note that $\hat{Y}_{d,\mathrm{POSTR}} - Y_d \doteq \hat{Y}_{d,\mathrm{HT}} - (Y_d / X_d)\hat{X}_{d,\mathrm{HT}}$. Defining $\mathbf{z}_k$ as $y_{dk} - (Y_d / X_d)x_{dk}$, we obtain that $\mathrm{Bias}(\hat{Y}_{d,\mathrm{POSTR}} \mid n_d) \doteq 0$.

We next proceed to evaluate the conditional bias and variance of estimators (2.1) – (2.6). We only illustrate the procedure for the regression estimator $\hat{Y}_{d,\ell r_1}$, as the steps are similar for the other estimators. Conditional on n_d, the distribution of s_d is that of an SRSWOR. This means that, for each sample s_d, n_d can be considered as having been selected from N_d. We express $\hat{Y}_{d,\ell r_1}$ as $\hat{Y}_{d,\ell r_1} = \sum_U \hat{y}_k + N/n\sum_s e_{dk}$, where $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$ and $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$.

Following Särndal and Hidiroglou (1989), we define the conditional regression vector $\mathbf{B}_{1d}^*$ as

$$\mathbf{B}_{1d}^* = \left[E\left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)\Big| n_d\right]^{-1} \times E\left[\left(\sum_s \frac{\mathbf{x}_k y_k}{c_k}\right)\Big| n_d\right]. \quad (3.2)$$

The estimated regression vector $\hat{\mathbf{B}}_{1d}$ will converge to $\mathbf{B}_{1d}^*$ (under appropriate conditions) in conditional design probability as n_d and N_d increase.

We have that

$$E\left[\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{n c_k}\Big| n_d\right] = \sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{N c_k} + \mathbf{R}_c$$

and

$$E\left[\sum_s \frac{\mathbf{x}_k y_k}{n c_k}\Big| n_d\right] = \sum_U \frac{\mathbf{x}_k y_k}{N c_k} + \mathbf{r}_c,$$

where

$$\mathbf{R}_c = \frac{w_d - W_d}{1 - W_d}\left(\frac{1}{N_d}\sum_{U_d} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} - \frac{1}{N}\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right) \doteq \mathbf{0}$$

and

$$\mathbf{r}_c = \frac{w_d - W_d}{1 - W_d}\left(\frac{1}{N_d}\sum_{U_d} \frac{\mathbf{x}_k y_k}{c_k} - \frac{1}{N}\sum_U \frac{\mathbf{x}_k y_k}{c_k}\right) \doteq \mathbf{0}.$$

Consequently, using Result 3.3 and assuming that $(w_d - W_d)/(1 - W_d) \doteq 0$, we have that $\hat{\mathbf{B}}_{1d} \doteq \mathbf{B}_{1d}^*$.

Define the "conditional residual" for the k^{th} unit as

$$E_{dk}^* = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}^*. \quad (3.3)$$

The deviation of $\hat{Y}_{d,\ell r_1}$ from the true value Y_d can be written as

$$\hat{Y}_{d,\ell r_1} - Y_d = -\sum_U E_{dk}^* + \frac{N}{n}\sum_s E_{dk}^* - \Delta_{1d}^* \quad (3.4)$$

where

$$\Delta_{1d}^* = \left(\frac{N}{n}\sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k\right)'(\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*).$$

In equation (3.4), Δ_{1d}^* is of lower order than $N/n\sum_s E_{dk}^*$. To see this, note that

$$E\left[\left(\frac{N}{n}\sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k\right)\Big| n_d\right] = N\frac{w_d - W_d}{1 - W_d}(\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U),$$

where $(w_d - W_d)/(1 - W_d)$ should be close to zero.

Also, as noted earlier, $\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*$ is near the vector $\mathbf{0}$ in conditional design probability. Hence $E_{dk}^* = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}^* \doteq y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d} = E_{dk}$. This implies that we can write (3.4) as

$$\hat{Y}_{d,\ell r_1} - Y_d \doteq -\sum_U E_{dk} + \frac{N}{n}\sum_s E_{dk}. \quad (3.5)$$

The conditional expectation of $\hat{Y}_{d,\ell r_1} - Y_d$ is approximately:

$$E\left[(\hat{Y}_{d,\ell r_1} - Y_d)\,|\,n_d\right] = N\frac{w_d - W_d}{1-W_d}\left(\tilde{\bar{E}}_{U_d} - \bar{E}_{U_d}\right) \quad (3.6)$$

where $\tilde{\bar{E}}_{U_d} = \sum_{U_d} E_{dk}/N_d$ and $\tilde{\bar{E}}_{U_d} = \sum_{U} E_{dk}/N$. Since $\bar{y}_{U_d} = W_d\,\tilde{\bar{y}}_{U_d}$, the conditional expectation (3.6) can be re-expressed as:

$$E[(\hat{Y}_{d,\ell r_1} - Y_d)\,|\,n_d] = N\frac{w_d - W_d}{1-W_d}\,[\tilde{y}_{U_d}(1-W_d) - (\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U)'\mathbf{B}_{1d}]. \quad (3.7)$$

The term $\sum_U E_{dk}$ is constant in (3.5). Using Result 3.4, the conditional population variance of $\hat{Y}_{d,\ell r_1}$ and its estimated value are respectively

$$V\left(\hat{Y}_{d,\ell r_1}\,|n_d\right) = N^2\left[\frac{w_d^2}{n_d}(1-f_d)V_{E_{U_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1-f_{\bar{d}})V_{E_{U_{\bar{d}}}}\right]$$

and

$$v\left(\hat{Y}_{d,\ell r_1}\,|n_d\right) = N^2\left[\frac{w_d^2}{n_d}(1-f_d)v_{e_{s_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1-f_{\bar{d}})v_{e_{s_{\bar{d}}}}\right]$$

where $V_{E_{U_d}} = (N_d - 1)^{-1}\sum_{U_d}\left(E_{dk} - \tilde{\bar{E}}_{U_d}\right)^2$,

$$V_{E_{U_{\bar{d}}}} = \left(N_{\bar{d}} - 1\right)^{-1}\sum_{U_{\bar{d}}}\left(E_{dk} - \tilde{\bar{E}}_{U_{\bar{d}}}\right)^2,$$

$$v_{e_{s_d}} = (n_d - 1)^{-1}\sum_{s_d}\left(a_{dk}e_k - \sum_{s_d}\frac{a_{dk}\,e_k}{n_d}\right)^2,$$

and

$$v_{e_{s_{\bar{d}}}} = \left(n_{\bar{d}} - 1\right)^{-1}\sum_{s_{\bar{d}}}\left(a_{dk}e_{dk} - \sum_{s_{\bar{d}}}\frac{a_{dk}\,e_{dk}}{n_{\bar{d}}}\right)^2.$$

The conditional bias and variances of the remaining five estimators can be derived similarly. Table 3 presents a summary of these properties. The required adjustment factors a_{dk} and residual terms e_{dk} are given in Tables 1 and 2.

4. Simulation Study

A simulation study was carried out to illustrate the conditional and unconditional properties of the ratio version of estimators (2.1) – (2.6). We studied these properties using a population of 1,000 bivariate observations (y, x). This population resulted from the concatenation of two generated population domains: a large domain of size 900 and a small domain of size 100. The (y, x) observations were generated within each domain assuming a ratio model $y_k = \beta x_k + \varepsilon_k$ where $E(\varepsilon_k) = 0$ and $V(\varepsilon_k) = \sigma^2 x_k$. The β coefficients were 1.0 and 3.0 in the large and small domains. The auxiliary variable x was generated using a gamma distribution $\Gamma(a,b)$, where $a = 3$ and $b = 16$. The dependent variable y was also generated by a gamma distribution, $\Gamma(A,B)$ such that the parameters A and B satisfied $E(y_k) = \beta x_k = AB$ and $V(y_k) = \sigma^2 x_k = AB^2$. After solving for A and B, we obtained $A = \beta^2/\sigma^2$ and $B = \sigma^2/\beta$. The term σ^2 was chosen to satisfy a set correlation between x and y defined by

$$\rho_{X,Y} = \frac{\beta\, b}{\sqrt{\sigma^2 b + \beta^2 b^2}}.$$

The preceding equation yields the constant term

$$\sigma^2 = \beta^2 b\left(\frac{1}{\rho_{X,Y}^2} - 1\right)$$

of the error variance. Common correlation values $\rho_{x,y}$ were used for both domains, ranging from 0.1 to of 0.9 in steps of 0.1, resulting in nine different populations. Random samples (M = 10,000) of size 250 were then repeatedly selected from the populations. For each sample, estimates of domain totals were computed using the estimators given in Table 4. We do not include the Hájek post-stratified estimator, $\tilde{Y}_{d,\ell r_2}$, as it corresponds exactly to its Horvitz-Thompson analogue, $\hat{Y}_{d,\ell r_2}$.

Table 3
Conditional Bias and Variance of Estimators (2.1) – (2.6)

Estimator	Conditional Bias	Estimated Conditional Variance
$\hat{Y}_{d,\ell r_1}$	$N\left((w_d - W_d)/(1-W_d)\right)\left(\tilde{y}_{U_d}(1-W_d) - (\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U)'\mathbf{B}_{1d}\right)$	$N^2\left[\left(w_d^2/n_d\right)(1-f_d)v_{e_{s_d}} + \left(w_{\bar{d}}^2/n_{\bar{d}}\right)(1-f_{\bar{d}})v_{e_{s_{\bar{d}}}}\right]$
$\hat{Y}_{d,\ell r_2}$	Almost 0	$\left(N_d^2(1-f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\hat{Y}_{d,\ell r_3}$	$N\left(w_d - W_d\right)\left(\tilde{y}_{U_d} - \tilde{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)$	$\left((N_d w_d)^2(1-f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\tilde{Y}_{d,\ell r_1}$	$N\left(w_d - W_d\right)\left(\left(\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_{U_d}\right)'\mathbf{B}_{1d}/(1-W_d)\right)$	$(N/w_d)^2\left[\left(w_d^2/n_d\right)(1-f_d)v_{e_{s_d}} + \left(w_{\bar{d}}^2/n_{\bar{d}}\right)(1-f_{\bar{d}})v_{e_{s_{\bar{d}}}}\right]$
$\tilde{Y}_{d,\ell r_2}$	Almost 0	$\left(N_d^2(1-f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\tilde{Y}_{d,\ell r_3}$	Almost 0	$\left(N_d^2(1-f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$

Table 4
Estimators and Associated Error Terms

Estimator	Ratio Version	Error Term
HT ratio: $\hat{Y}_{d,\ell r_1}$	$\hat{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HT}}\left(X / \hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_{1d}\ x_k$, $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$
HT post-stratified ratio: $\hat{Y}_{d,\ell r_2}$	$\hat{Y}_{d,\mathrm{POSTR}} = \hat{Y}_{d,\mathrm{HT}}\left(X_d / \hat{X}_{d,\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_{2d}\ x_{dk}$, $\hat{R}_{2d} = \hat{Y}_{d,\mathrm{HT}} / \hat{X}_{\mathrm{d,HT}}$
HT alternate ratio: $\hat{Y}_{d,\ell r_3}$	$\hat{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HT}} + (X_d - \hat{X}_{d,\mathrm{HT}})\left(\hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_3 x_k$, $\hat{R}_3 = \hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$
Hájek ratio: $\tilde{Y}_{d,\ell r_1}$	$\tilde{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HA}} + (X - \hat{X}_{\mathrm{HA}})\left(\hat{Y}_{d,\mathrm{HT}} / \hat{X}_{\mathrm{HT}}\right).$	$e_{dk} = y_{dk} - \hat{R}_{1d} x_k$, $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$
Hájek alternate ratio: $\tilde{Y}_{d,\ell r_3}$	$\tilde{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\left(\hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_3 x_k$, $\hat{R}_3 = \hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$

4.1 Unconditional Results

The unconditional properties of the estimators were assessed using two performance measures: (i) root mean squared error (RMSE) and (ii) coverage rate (CR). They are:

i. The *RMSE* is defined as

$$\sqrt{\sum_{m=1}^{M} (\hat{Y}_d^{(m)} - Y_d)^2 / M}\ ,$$

where $\hat{Y}_d^{(m)}$ is the estimated total (either Horvitz-Thompson or Hájek type) based on sample *m*, and *M* is the total number of samples drawn for the simulation.

ii. The coverage rate *CR* for a given estimator $\hat{Y}$ is defined as the ratio of the number of times that the 95% confidence interval

$$\hat{Y}_d^{(m)} \pm 1.96\sqrt{v(\hat{Y}_d^{(m)})}$$

contains the true population total to the number of replicates. We used the unconditional variances given by (2.12), and the error terms in Table 4 to estimate the required variances.

The four graphs provided in Figures 1 and 2, summarize the unconditional analysis for small and large domains. Also shown is the impact of increasing $\rho_{x,y}$. The square root of the average mean squared error and coverage rates are used to compare the estimators.

In Figure 1, we note that the RMSE decreases substantially with increasing $\rho_{x,y}$. This can be attributed to the decreasing dispersion of the dependent variable conditional on the independent variable as the correlation between the two increases. We also note that the spread of the RMSE is narrower for the large domain than for the small domain. The ranking of the estimators in terms of RMSE from worst to best is as follows: (i) HT ratio (HT RAT), (ii) Hájek ratio (HA RAT), (iii) HT alternate ratio (HT ALTR), (iv) Hájek alternate ratio (HA ALTR), and (v) HT post-stratified ratio (HT POSTR). This ranking is in agreement with Result 3.2.

In Figure 2, we note that the unconditional coverage rates are similar across all the estimators regardless of the correlation $\rho_{x,y}$. For small domains the Horvitz-Thompson estimators exhibit a slight degradation in the coverage rate when $\rho_{x,y}$ is weak. But as the correlation increases, their coverage rate becomes comparable to the Hájek type estimators. The Hájek estimators have a better overall coverage rate than their Horvitz-Thompson counterparts.

4.2 Conditional Results

The conditional properties of the estimators were studied using: (i) average relative conditional bias and (ii) conditional coverage rates. They are defined as:

i. $\mathrm{ARB}_d = (100 / M_d)\sum_{m=1}^{M_d} (\hat{Y}_d^{(m)} - Y_d) / Y_d$, where M_d is the number of samples of size n_d.

ii. The conditional coverage rate has the same definition as its unconditional counterpart. The associated variance is

$$v_d^2 = \frac{1}{M_d - 1}\sum_{m=1}^{M_d} (\hat{Y}_d^{(m)} - \bar{Y}_d)^2$$

where

$$\bar{Y}_d = \frac{1}{M_d}\sum_{m=1}^{M_d} \hat{Y}_d^{(m)}\ .$$

Table 5 summarizes the conditional biases of the ratio versions of estimators (2.1) – (2.4) and (2.6). They were obtained from Table 3 using a single auxiliary variable.

The relative conditional bias and coverage rates of the estimators are summarized in Figures 3, 4a, and 4b with respect to the realized sample size n_d for large and small domains, and for two correlations ($\rho_{X,Y} = 0.90$ and $\rho_{X,Y} = 0.60$).

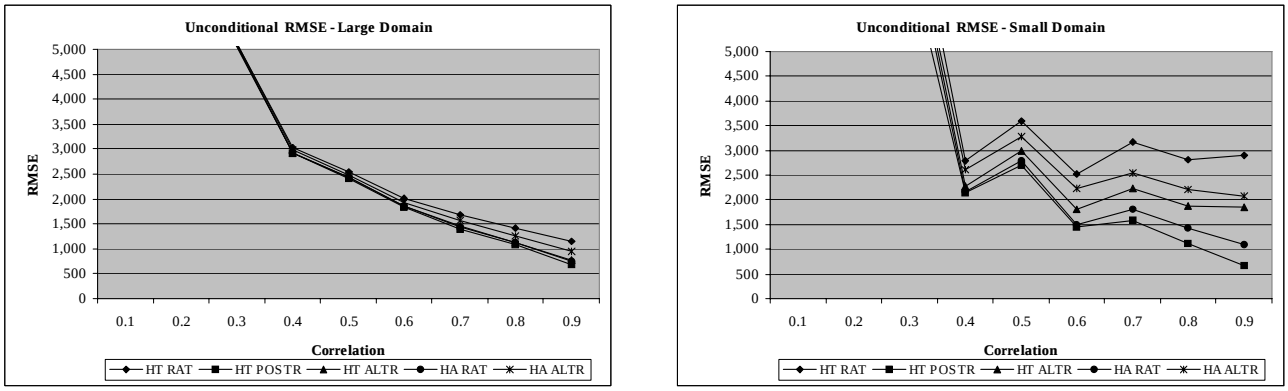

Figure 1. Unconditional RMSE.

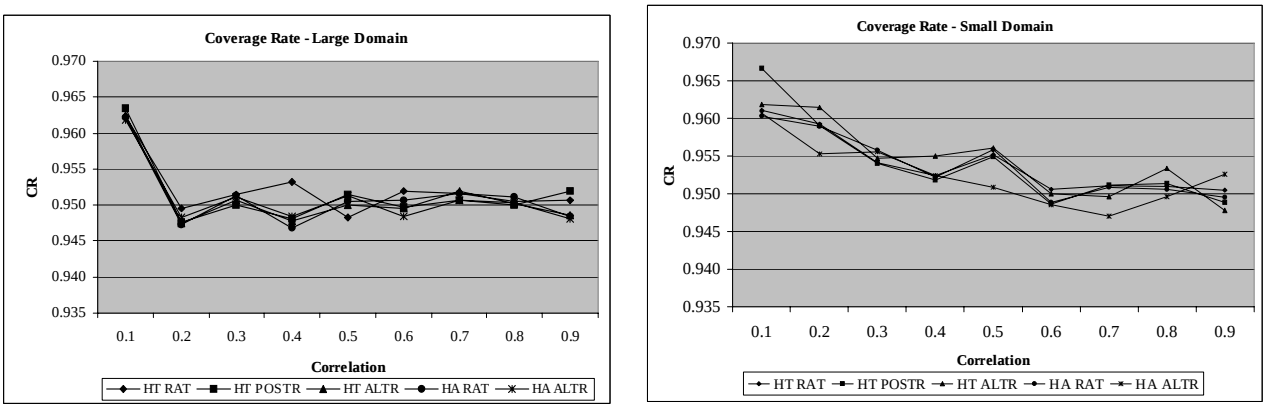

Figure 2. Unconditional Coverage Rates.

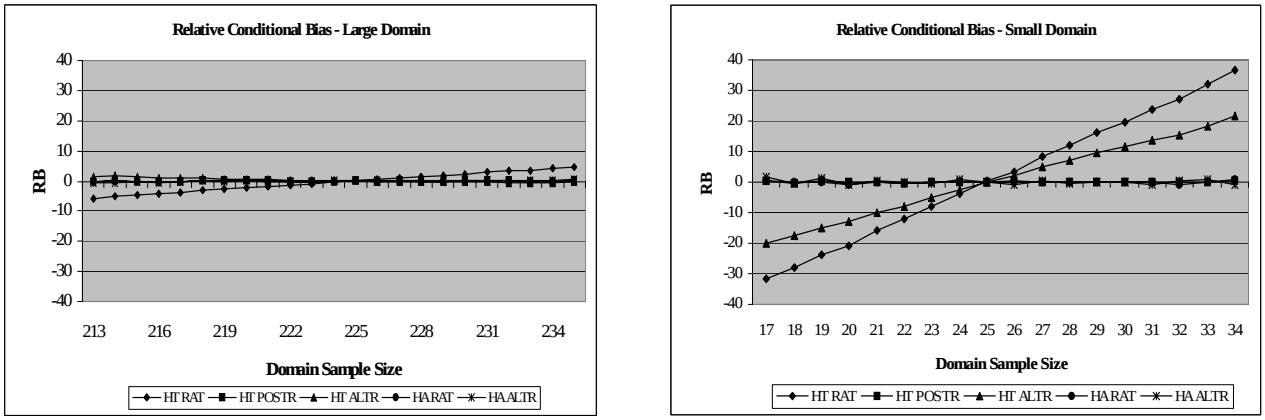

Figure 3. Average Relative Conditional Bias for $\rho_{X,Y} = 0.90$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

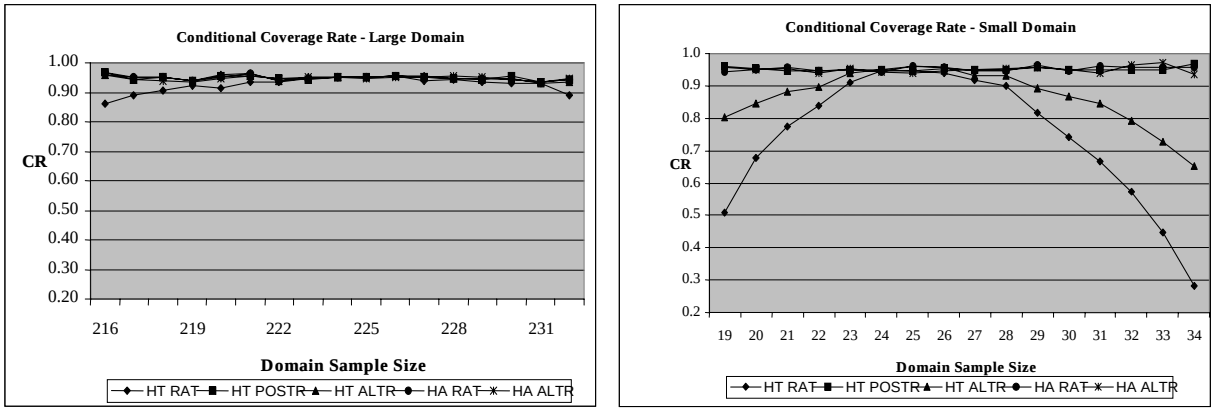

Figure 4a. Conditional Coverage Rates for $\rho_{X,Y} = 0.90$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

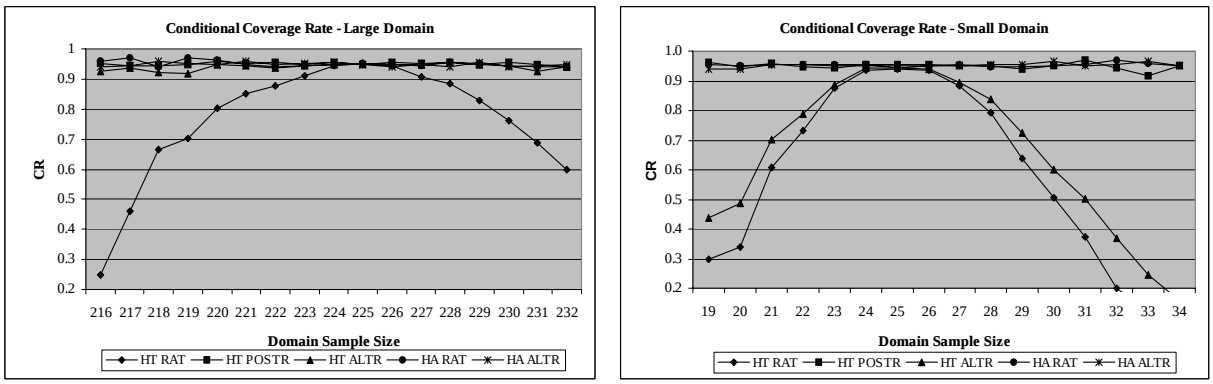

Figure 4b. Conditional Coverage Rates for $\rho_{X,Y} = 0.60$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

Table 5
Conditional Biases of Ratio Versions of Estimators (2.1) – (2.4) and (2.6)

Estimator	Conditional Bias
HT Ratio: $\hat{Y}_{d,\ell r_1}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{(\bar{x}_U - W_d\tilde{x}_{U_d})}{\bar{x}_U(1-W_d)}$
HT post-stratified ratio: $\hat{Y}_{d,\ell r_2}$	Almost 0
HT alternate ratio: $\hat{Y}_{d,\ell r_3}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{(\bar{x}_U - \tilde{x}_{U_d}\bar{y}_U / \tilde{y}_{U_d})}{\bar{x}_U}$
Hájek ratio: $\tilde{\tilde{Y}}_{d,\ell r_1}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{W_d(\bar{x}_U - \tilde{x}_{U_d})}{(1-W_d)\bar{x}_U}$
Hájek alternate ratio: $\tilde{\tilde{Y}}_{d,\ell r_3}$	Almost 0

The conditional bias presented in Figure 3 supports the theoretical results presented in Table 5. The three Hájek estimators are nearly conditionally unbiased. The magnitude of the conditional bias of both the HT ratio estimator and the HT alternate ratio estimator is in agreement with the theoretical conditional bias. But it should be noted that the conditional bias associated with the HT alternate ratio estimator is smaller than the one of the HT ratio estimator. Also, in larger domains, this conditional bias is less pronounced for the HT alternate ratio estimator.

The conditional coverage rates are given in Figures 4a and 4b. We note that the three Hájek estimators follow closely the nominal 95% coverage probability. The coverage rate of the HT alternate ratio estimator is reasonable in larger domains despite its being conditionally biased. But its coverage deteriorates substantially in smaller domains. The coverage rate of the HT ratio estimator is not acceptable. But it should be noted that the coverage rates of the conditionally biased estimators improve as the realized sample size n_d approaches the expected domain sample size $E(n_d)$.

In summary, the simulation study identified the three Hájek estimators, *Hájek post-stratified ratio*, *Hájek alternate ratio*, and *Hájek ratio* as the best estimators in terms of their conditional and unconditional properties. Note that even though the *Hájek ratio* estimator uses the least domain auxiliary data (it uses domain population counts N_d), its mean squared error is still reasonable. The *Hájek post-stratified ratio* is the best estimator in terms of its conditional and unconditional properties.

5. Concluding Remarks

We have studied six possible regression estimators of domain totals, each using various levels of auxiliary information at the domain and/or population level. The only estimator that has regression weights that are not domain dependent and that also have the additive property is Horvitz-Thompson estimator $\hat{Y}_{d,\ell r_1}$. This estimator is constructed using auxiliary information at the population level: the domain dependent independent variable y_{dk} is regressed on the auxiliary vector $\mathbf{x}_k$. However, it can be seriously conditionally biased and the associated confidence intervals can be understated.

The Hájek-type estimators have two the disadvantages: (i) they do not have the additive property; and (ii) their associated regression weights are domain dependent. However, they have the best conditional properties. They are nearly conditionally unbiased, and the conditional confidence intervals associated with the estimators follow closely the nominal coverage rate. They also have the smaller unconditional MSE's. The Hájek estimator that uses the least auxiliary data at the domain level is $\tilde{\tilde{Y}}_{d,\ell r_1}$. It requires domain population counts N_d ($d = 1, \ldots, D$), and the population totals $\mathbf{X}$. Its conditional and unconditional properties are reasonable.

The best Hájek estimator, $\tilde{\tilde{Y}}_{d,\ell r_2}$, uses auxiliary information at the domain level. The Hájek regression type estimator $\tilde{\tilde{Y}}_{d,\ell r_1}$ can be made domain independent using a single set of regression weights as follows. Suppose that the most important domains are $U_g \subseteq U$ $(g = 1, ..., G)$, and that these domains are mutually exclusive and exhaustive. The resulting Hájek estimator is

$$\tilde{\tilde{Y}}_{d,\ell r_1} = \sum_{g=1}^{G} \left[\hat{Y}_{g,\mathrm{HA}} + (\mathbf{X}_g - \hat{\mathbf{X}}_{g,\mathrm{HA}})' \hat{\mathbf{B}}_{1g} \right]$$

where

$$\hat{Y}_{g,\mathrm{HA}} = \left(N_g / \hat{N}_g \right) \hat{Y}_{g,\mathrm{HT}}, \quad \hat{Y}_{g,\mathrm{HT}} = \sum_{s_g} w_k \, y_{dk}$$

and

$$\hat{\mathbf{B}}_{1g} = \left(\sum_{s_g} w_k \, \mathbf{x}_k \, \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_g} w_k \, \mathbf{x}_k \, y_k / c_k .$$

References

Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: Wiley, Interscience.

Estevao, V.M., Hidiroglou, M.A. and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 2, 181-204.

Estevao, V.M., and Särndal, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 213-231.

Hartley, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.

Hidiroglou, M.A. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada report, September, 1991.

Holt, D., and Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

Rao, J.N.K. (1985). Conditional Inferences in Survey Sampling. *Survey Methodology*, 15-32.

Särndal, C.-E., and Hidiroglou, M.A. (1989). Small Domain Estimation: A conditional analysis. *Journal of American Statistical Association*, 84, 405, 266-275.

Särndal, C.-E., Swensson B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.