Survey Methodology, June 2016
Vol. 42, No. 1, pp. 121-135
Statistics Canada, Catalogue No. 12-001-X

A note on regression estimation with unknown population size

Michael A. Hidiroglou, Jae Kwang Kim and Christian Olivier Nambeu[1]

Abstract

The regression estimator is extensively used in practice because it can improve the reliability of the estimated parameters of interest such as means or totals. It uses control totals of variables known at the population level that are included in the regression set up. In this paper, we investigate the properties of the regression estimator that uses control totals estimated from the sample, as well as those known at the population level. This estimator is compared to the regression estimators that strictly use the known totals both theoretically and via a simulation study.

Key Words: Optimal estimator; Survey sampling; Weighting.

1 Introduction

Regression estimation has been increasingly used in large survey organizations as a means to improve the reliability of the estimators of parameters of interest (such as totals or means) when auxiliary variables are available in the population. A comprehensive overview of the regression estimator in survey sampling can be found in Cassel, Särndal and Wretman (1976) and Fuller (2009) among others. We next illustrate how the regression estimator can be used to estimate the total, $Y = \sum_{i \in U} y_i$ where $U = \{1, \ldots, N\}$ denotes the target population. A sample s of expected size n is selected according to a sampling plan $p(s)$ from U, where π_i is the resulting probability of inclusion of the first order. In the absence of auxiliary variables, we use the Horvitz-Thompson estimator given by $\hat{Y}_\pi = \sum_{i \in s} d_i y_i$ (Horvitz and Thompson 1952) where $d_i = 1/\pi_i$ is referred to as the weight survey associated with unit i. The regression estimator is given by

$$\hat{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}, \tag{1.1}$$

where $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$, and $\hat{\mathbf{B}}$ is a $p-$dimensional vector of estimated regression coefficients, which is computed as a function of the observed variables $\left(y_i, \mathbf{x}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ in the sample s.

Note that the components of the vector of population total $\mathbf{X}$ are known for each of the corresponding components variables in the vector $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$ used to compute $\hat{\mathbf{B}}$. However, there are instances when we have more observed auxiliary variables in the sample than in the population. Assume that the sample has q observed variables $(q > p)$, and that the p variables in the population are a subset of the q variables observed in the sample. Furthermore, suppose that some of the extra $q - p$ variables in the sample are well correlated with the variable of interest y. Can these extra variables be incorporated in the

1. Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, ON, Canada K1A 0T6. E-mail: hidirog@yahoo.ca; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: jkim@iastate.edu; Christian Olivier Nambeu, Business Survey Methods Division, Statistics Canada, ON, Canada K1A 0T6. E-mail: christianolivier.nambeu@canada.ca.

regression estimator so as to make it more efficient? Singh and Raghunath (2011) attempted to respond to that question for the case where $q = p + 1$. Their extra variable in the sample was the intercept. They used it to estimate the unknown population size N by $\hat{N} = \sum_{i \in s} d_i$.

In this article, we compare the estimator proposed by Singh and Raghunath (2011) to other regression estimators when N is known or unknown. In Section 2, we describe standard regression estimators for estimating totals when N is known as well as the regression proposed by Singh and Raghunath (2011) when N is unknown. In Section 3, an alternative estimator is proposed for the case where N is unknown. A simulation study is carried out in Section 4, to illustrate the performance of the various estimators studied in terms of bias and mean square error. Overall conclusions and recommendations are given in Section 5.

2 Regression estimators

Under general regularity conditions (Isaki and Fuller 1982; Montanari 1987), an approximation to the regression estimator (1.1) is

$$\tilde{Y}_{\text{REG}} = \hat{Y}_{\pi} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\pi}\right)^{\text{T}} \mathbf{B}, \tag{2.1}$$

where $\mathbf{B}$ is the limit in probability of $\hat{\mathbf{B}}$ when both the sample and the population sizes tend to infinity. For large samples, the variance of regression estimator (1.1) can be studied via (2.1). Note that $\tilde{Y}_{\text{REG}}$ is unbiased under the sampling plan $p(s)$ and can be re-expressed as:

$$\tilde{Y}_{\text{REG}} = \mathbf{X}^{\text{T}}\mathbf{B} + \sum_{i \in s} d_i E_i, \tag{2.2}$$

where $E_i = y_i - \mathbf{x}_i^{\text{T}}\mathbf{B}$.

The design variance for $\hat{Y}_{\text{REG}}$ can be approximated by

$$\text{AV}_p\left(\hat{Y}_{\text{REG}}\right) = \sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{2.3}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$ and π_{ij} is the second order inclusion probability for units i and j. Both the model-assisted (Särndal, Swensson and Wretman 1992) and the optimal-variance (Montanari 1987) approaches can be used to estimate $\mathbf{B}$. They both yield approximately unbiased estimators. In the case of the model-assisted approach, the basic properties (bias and variance terms) are valid even when the model is not correctly specified. Under the optimal-variance approach no assumption is made on the variable of interest.

The model-assisted estimator of Särndal et al. (1992) assumes a working model between the variable of interest (y) and the auxiliary variables $(\mathbf{x})$. The working model is denoted by $m: y_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{\beta}$ is a vector of p unknown parameters, $E_m(\varepsilon_i \mid \mathbf{x}_i) = 0$, $V_m(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2$, and $\text{Cov}_m\left(\varepsilon_i, \varepsilon_j \mid \mathbf{x}_i, \mathbf{x}_j\right) = 0, i \neq j$. Under this approach, $\mathbf{B}$ in equation (2.1) is the ordinary least squares estimator of $\boldsymbol{\beta}$ in the population and it is given by

$$\mathbf{B}_{\text{GREG}} = \left(\sum_{i \in U} c_i \mathbf{x}_i \mathbf{x}_i^{\text{T}}\right)^{-1} \left(\sum_{i \in U} c_i \mathbf{x}_i y_i\right), \tag{2.4}$$

where $c_i = \sigma_i^{-2}$. This yields the following estimator for the total Y

$$\hat{Y}_{\text{GREG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{GREG}}, \tag{2.5}$$

where

$$\hat{\mathbf{B}}_{\text{GREG}} = \left(\sum_{i\in s} c_i d_i \mathbf{x}_i \mathbf{x}_i^{\text{T}}\right)^{-1} \left(\sum_{i\in s} c_i d_i \mathbf{x}_i y_i\right). \tag{2.6}$$

The optimal estimator of Montanari (1987), obtained by minimizing the design variance of

$$\tilde{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \mathbf{B},$$

is

$$\tilde{Y}_{\text{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \mathbf{B}_{\text{OPT}}, \tag{2.7}$$

where

$$\begin{aligned} \mathbf{B}_{\text{OPT}} &= \left\{V\left(\hat{\mathbf{X}}_\pi\right)\right\}^{-1} \text{Cov}\left(\hat{\mathbf{X}}_\pi, \hat{Y}_\pi\right) \\ &= \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\text{T}}}{\pi_j}\right)^{-1} \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \end{aligned} \tag{2.8}$$

The optimal estimator for the total Y is estimated by

$$\hat{Y}_{\text{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{OPT}}, \tag{2.9}$$

where

$$\hat{\mathbf{B}}_{\text{OPT}} = \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\text{T}}}{\pi_j}\right)^{-1} \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \tag{2.10}$$

Note that the computation of the regression vectors requires that the first component that defines them is invertible. We can ensure this by reducing the number of auxiliary variables that are input into the regression if not much loss in efficiency of the resulting regression estimator is incurred. If, on the other hand, there is a significant loss in efficiency, then we can invert these singular matrices using generalised inverses.

As mentioned in the introduction, not all population totals may be known for each component of the auxiliary vector $\mathbf{x}$. The regression normally uses the auxiliary variables for which a corresponding population total is known. Decomposing $\mathbf{x}_i$ as $\left(1, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ where $\mathbf{x}_i^* = \left(x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$, Singh and Raghunath (2011) proposed a GREG-like estimator that assumes that the regression is based on an intercept and the variable $\mathbf{x}^*$, even though only the population total of the $\mathbf{x}^*$ is known.

For the case that N is not known and that the population total of $\mathbf{x}^*$ is known, their estimator is

$$\hat{Y}_{\text{SREG}} = \hat{Y}_{\pi} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{2,\text{GREG}}, \tag{2.11}$$

where $\mathbf{X}^{*} = \sum_{i \in U} \mathbf{x}_i^{*}$ and $\hat{\mathbf{X}}_{\pi}^{*} = \sum_{i \in s} d_i \mathbf{x}_i^{*}$. The regression vector of estimated coefficients $\hat{\mathbf{B}}_{2,\text{GREG}}$ is obtained from $\hat{\mathbf{B}}_{\text{GREG}} = \left(\hat{B}_{1,\text{GREG}}, \hat{\mathbf{B}}_{2,\text{GREG}}^{\text{T}}\right)^{\text{T}}$ given by (2.6). The approximate design variance for $\hat{Y}_{\text{SREG}}$ takes the same form as equation (2.3), with $E_i = y_i - \mathbf{x}_i^{*\text{T}} \mathbf{B}_{2,\text{GREG}}$, where

$$\mathbf{B}_{2,\text{GREG}} = \left\{\sum_{i \in U} c_i \left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right)\left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right)^{\text{T}}\right\}^{-1} \sum_{i \in U} c_i \left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right) y_i$$

and $\bar{\mathbf{X}}_N^{*} = \sum_{i \in U} \mathbf{x}_i^{*} / N$.

The properties of (2.11) can be obtained by noting that

$$\begin{aligned}
\hat{Y}_{\text{SREG}} - Y &= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{2,\text{GREG}} \\
&= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \mathbf{B}_{2,\text{GREG}} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \left(\hat{\mathbf{B}}_{2,\text{GREG}} - \mathbf{B}_{2,\text{GREG}}\right).
\end{aligned}$$

Since $\hat{\mathbf{B}}_{2,\text{GREG}} - \mathbf{B}_{2,\text{GREG}} = O_p\left(n^{-1/2}\right)$ under some regularity conditions discussed in Fuller (2009, Chapter 2), the last term is of smaller order. Thus, ignoring the smaller order terms, we get the following approximation

$$\hat{Y}_{\text{SREG}} - Y \cong \sum_{i \in s} d_i E_i - \sum_{i \in U} E_i, \tag{2.12}$$

where $E_i = y_i - \mathbf{x}_i^{*\text{T}} \mathbf{B}_{2,\text{GREG}}$. Thus, $\hat{Y}_{\text{SREG}}$ is approximately design-unbiased. The asymptotic variance can be computed using

$$V\left\{\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right\} = E\left\{\left(\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right)^2\right\}.$$

As we can see, the asymptotic variance can be quite large unless $\sum_{i \in U} E_i = 0$.

Remark 2.1 *If* $y_i = a + bx_i$, *we have* $\hat{Y}_{\text{SREG}} - Y = \left(\hat{N}_{\pi} - N\right)a$ *and this implies that* $V\left(\hat{Y}_{\text{SREG}}\right) = a^2 V\left(\hat{N}_{\pi}\right)$. *This means that if* $V\left(\hat{N}_{\pi}\right) > 0$, *we can artificially increases* $a^2 V\left(\hat{N}_{\pi}\right)$, *the variance of* $\hat{Y}_{\text{SREG}}$, *by choosing large values of* a.

Note that the optimal regression estimator using $\mathbf{x}^{*} = \left(x_2, \ldots, x_p\right)^{\text{T}}$ is also approximately design unbiased because

$$\begin{aligned}
\hat{Y}_{\text{OPT}}^{*} - Y &= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{OPT}}^{*} \\
&= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \mathbf{B}_{\text{OPT}}^{*} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \left(\hat{\mathbf{B}}_{\text{OPT}}^{*} - \mathbf{B}_{\text{OPT}}^{*}\right),
\end{aligned}$$

where $\mathbf{B}_{\text{OPT}}^{*}$ is obtained by replacing $\mathbf{x}_i$ by $\mathbf{x}_i^{*}$ in equation (2.8). Since $\hat{\mathbf{B}}_{\text{OPT}}^{*} - \mathbf{B}_{\text{OPT}}^{*} = O_p\left(n^{-1/2}\right)$ under some regularity conditions discussed in Fuller (2009, Chapter 2), ignoring the smaller order terms we get

$$\hat{Y}^{*}_{\text{OPT}} - Y \cong \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}^{*}_{\pi}\right)^{\text{T}} \mathbf{B}^{*}_{\text{OPT}}.$$

The asymptotic variance of $\hat{Y}^{*}_{\text{OPT}}$ is smaller than the one associated with $\hat{Y}_{\text{SREG}}$. The reason for this is that the optimal estimator minimizes the asymptotic variance among the class of estimators of the form

$$\hat{Y}_{B} = \hat{Y}_{\pi} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}^{*}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}} \tag{2.13}$$

indexed by $\hat{\mathbf{B}}$.

3 Alternative regression estimator

We now consider an alternative estimator that does not use the population size (N) information. Rather, it uses the known inclusion probabilities π_i provided that they are known for each unit in the population. Given that $\sum_{i\in U} \pi_i = n$, we can use $\mathbf{z}_i = \left(\pi_i, \mathbf{x}^{*\text{T}}_i\right)^{\text{T}}$ as auxiliary data in the model

$$y_i = \mathbf{z}^{\text{T}}_i \beta + e_i,$$

where $e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2 \pi_i\right)$. This means that the incorporation of the variance structure c_i of the error in the regression vector is given by $c_i = d_i / \sigma^2$. The resulting estimator is given by

$$\hat{Y}_{\text{KREG}} = \hat{Y}_{\pi} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}, \tag{3.1}$$

with $\mathbf{Z} = \sum_{i\in U} \mathbf{z}_i$, $\hat{\mathbf{Z}} = \sum_{i\in s} d_i \mathbf{z}_i$ and

$$\hat{\mathbf{B}}_{\text{KREG}} = \left(\sum_{i\in s} c_i d_i \mathbf{z}_i \mathbf{z}^{\text{T}}_i\right)^{-1} \sum_{i\in s} c_i d_i \mathbf{z}_i y_i. \tag{3.2}$$

This estimator corresponds exactly to the one given by Isaki and Fuller (1982).

Remark 3.1 *By construction,*

$$\sum_{i\in s} d^2_i \left(y_i - \mathbf{z}^{\text{T}}_i \hat{\mathbf{B}}_{\text{KREG}}\right) \mathbf{z}_i = \mathbf{0}.$$

Since π_i *is a component of* $\mathbf{z}_i$, *we have* $\sum_{i\in s} d_i \left(y_i - \mathbf{z}^{\text{T}}_i \hat{\mathbf{B}}_{\text{KREG}}\right) = 0$, *this leads to*

$$\hat{Y}_{\text{KREG}} = \mathbf{Z}^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}.$$

Thus, $\hat{Y}_{\text{KREG}}$ *is the best linear unbiased predictor of* $Y = \sum_{i=1}^{N} y_i$ *under the model*

$$y_i = \pi_i \beta_1 + \mathbf{x}^{*\text{T}}_i \boldsymbol{\beta}_2 + e_i,$$

where $e_i \sim \left(0, \sigma^2 \pi_i\right)$.

Note that $\hat{\mathbf{B}}_{\text{KREG}}$ can be expressed as $\hat{\mathbf{B}}_{\text{GREG}}$ by setting $c_i = d_i/\sigma^2$ and $\mathbf{x}_i = \mathbf{z}_i$. Thus, the proposed regression estimator can be viewed as a special case of GREG estimator. Using the argument similar to (2.12), we obtain

$$\hat{Y}_{\text{KREG}} - Y \cong \sum_{i\in s} d_i E_i^* - \sum_{i\in U} E_i^*, \tag{3.3}$$

where $E_i^* = y_i - \mathbf{z}_i^{\text{T}}\mathbf{B}_{\text{KREG}}$ and

$$\mathbf{B}_{\text{KREG}} = \left(\sum_{i\in U} c_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i\in U} c_i \mathbf{z}_i y_i.$$

The proposed estimator is approximately unbiased and its asymptotic variance

$$V\left\{\sum_{i\in s} d_i \left(y_i - \mathbf{z}_i^{\text{T}}\mathbf{B}_{\text{KREG}}\right)\right\} = \sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{E_i^*}{\pi_i}\frac{E_j^*}{\pi_j}$$

is often smaller than the asymptotic variance of Singh and Raghunath (2011)'s estimator.

The optimal version of $\hat{Y}_{\text{KREG}}$ uses $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ as auxiliary data. It is given by

$$\hat{Y}_{\text{KOPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{KOPT}}, \tag{3.4}$$

where $\hat{\mathbf{B}}_{\text{KOPT}}$ is obtained by substituting $\mathbf{x}_i$ by $\mathbf{z}_i$ in equation (2.10).

Remark 3.2 *For fixed-size sampling designs, we have* $V_p\left(\sum_{i\in s} d_i\pi_i\right) = 0$. *In this case, the optimal regression coefficient vector* $\mathbf{B}_{\text{KOPT}} = V_p\left(\hat{\mathbf{Z}}_\pi\right)^{-1} \text{Cov}_p\left(\hat{\mathbf{Z}}_\pi, \hat{Y}_\pi\right)$ *cannot be computed because the variance-covariance matrix* $V_p\left(\hat{\mathbf{Z}}_\pi\right)$ *is not invertible. Thus, the optimal estimator with* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ *reduces to the optimal estimator* (2.9) *only using* $\mathbf{x}_i^*$.

Remark 3.3 *For random-size sampling designs,* $V_p\left(\sum_{i\in s} d_i\pi_i\right) \geq 0$. *In this case, all of the components of* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ *can be used in the design-optimal regression estimator* (2.9).

A difficulty with using the optimal estimator $\hat{Y}_{\text{KOPT}}$ is that it requires the computation of the joint inclusion probabilities π_{ij}: these may be difficult to compute for certain sampling designs. An estimator that does not require the computation of the joint inclusion probabilities is obtained by assuming that $\pi_{ij} = \pi_i\pi_j$. We refer to this estimator as the pseudo-optimal estimator, $\hat{Y}_{\text{POPT}}$. It is given by

$$\hat{Y}_{\text{POPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}, \tag{3.5}$$

where

$$\hat{\mathbf{B}}_{\text{POPT}} = \left(\sum_{i\in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i\in s} c_i d_i \mathbf{z}_i y_i$$

and

$$c_i = d_i - 1.$$

In general, the pseudo-optimal estimator $\hat{Y}_{\text{POPT}}$ should yield estimates that are quite close to those produced by $\hat{Y}_{\text{KREG}}$ when the sampling fraction is small. Note that $\hat{Y}_{\text{POPT}}$ is exactly equal to the optimal estimator $\hat{Y}_{\text{KOPT}}$ in the case of Poisson sampling. In this sampling design the inclusion probabilities of units in the sample are independent. The approximate design variance for $\hat{Y}_{\text{KREG}}$, $\hat{Y}_{\text{KOPT}}$ and $\hat{Y}_{\text{POPT}}$ have the same form as the one given in equation (2.3) with the E_i's respectively given by $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{KREG}}$, $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{KOPT}}$ and $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{POPT}}$.

4 Simulations

We carried out two simulation studies. The first one used a dataset provided in the textbook of Rosner (2006) and the second one was based on an artificial population created according to a simple linear regression model. The first simulation assessed the performance of all of the estimators with respect to different sample schemes while the second simulation study focused on the impact of changing the intercept value in the model.

The parameter of interest for these two simulations is the total of the variable of interest y: $Y = \sum_{i \in U} y_i$. All estimators were used $\left(\hat{Y}_{\text{GREG}}, \hat{Y}_{\text{OPT}}, \hat{Y}_{\text{POPT}}, \hat{Y}_{\text{SREG}}, \hat{Y}_{\text{KREG}} \text{ and } \hat{Y}_{\text{KOPT}}\right)$ with the available auxiliary data. Table 4.1 summarizes the auxiliary data and the variance structure of the errors (when applicable) associated with the estimators used in the two studies.

Table 4.1
Estimators used in simulation

***N* known**	***N* unknown**
$\hat{Y}_{\text{GREG2}}$ as defined by (2.5) with $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ and $c_i = c$	$\hat{Y}_{\text{SREG1}}$ as defined as special case of (2.11) with $\mathbf{x}_i^* = (x_{2i})$
$\hat{Y}_{\text{OPT2}}$ as defined by (2.9) with $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{OPT1}}$ as defined by (2.9) with $\mathbf{x}_i = (x_{2i})$
$\hat{Y}_{\text{OPT3}}$ as defined by (2.9) with $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{KREG2}}$ as defined by (3.1) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i / \sigma^2$
$\hat{Y}_{\text{POPT3}}$ as defined by (3.5) with $\mathbf{z}_i = (1, \pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i - 1$	$\hat{Y}_{\text{KOPT2}}$ as defined as (3.4) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$
	$\hat{Y}_{\text{POPT2}}$ as defined as (3.5) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i - 1$

The performance of all estimators was evaluated based on the relative bias, the Monte Carlo relative efficiency and the approximate relative efficiency. Expressions of these quantities as shown below.

1. *Relative bias:*

$$\mathrm{RB}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{100}{R}\sum_{i=1}^{R}\frac{\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)}{Y}, \tag{4.1}$$

where $\hat{Y}_{\mathrm{EST}(r)}$ represents one of the estimators presented in Table 4.1 as computed in the r^{th} Monte Carlo sample.

2. *Monte Carlo Relative efficiency*

$$\mathrm{RE}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.2}$$

where

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)^2.$$

The RE measures the relative efficiency of the estimator $\hat{Y}_{\mathrm{EST}}$ with respect to $\hat{Y}_{\mathrm{GREG2}}$.

3. *Approximate Relative efficiency*

$$\mathrm{AR}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.3}$$

where

$$\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i}{\pi_i}\frac{E_j}{\pi_i},$$

is the approximate variance of $\hat{Y}_{\mathrm{EST}}$ with $E_i = y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{EST}}$. The approximate relative efficiency (AR) measures the relative gain in efficiency of $\hat{Y}_{\mathrm{EST}}$ with respect to $\hat{Y}_{\mathrm{GREG2}}$ using the population residual obtained by Taylor linearisation. It is expected that RE and AR give comparable results. However, as we will see, this may not be the case.

4.1 Simulation 1

The population was the dataset (FEV.DAT) available on the CD that accompanies the textbook by Rosner (2006). The data file contains 654 records from a study on Childhood Respiratory Disease carried out in Boston. The variables in the file were: age, height, sex (male female), smoking (indicates whether the individual smokes or not) and Forced expiratory volume (FEV). Singh and Raghunath (2011) used the same data set. The parameter of interest is the total height (y) of the population. The variable age (x_1) was used as auxiliary variable in the regression. The variable FEV (x_2) was chosen as the size variable to compute probabilities of selection for the sampling schemes that are considered in this simulation. The two variables sex and smoking were discarded from the simulation. Table 4.2 summarizes the central tendency measures of the three variables in the population. For each variable, the mean and median were similar. This indicates that the three variables have a symmetrical distribution.

Table 4.2
Descriptive statistics of y, x_1 and x_2

	Min	Q1	Median	Mean	Q3	Max
y	46	57	61.5	61.14	65.5	74
x_1	3	8	10	9.931	12	19
x_2	0.79	1.98	2.55	2.64	3.12	5.79

Figure 4.1 displays the relationship between the variable of interest y and the auxiliary variable x_1. The relationship between Height (y) and the age (x_1) appears to be linear but does not go through the origin. The Pearson correlation coefficient between y and x_1 was 0.79.

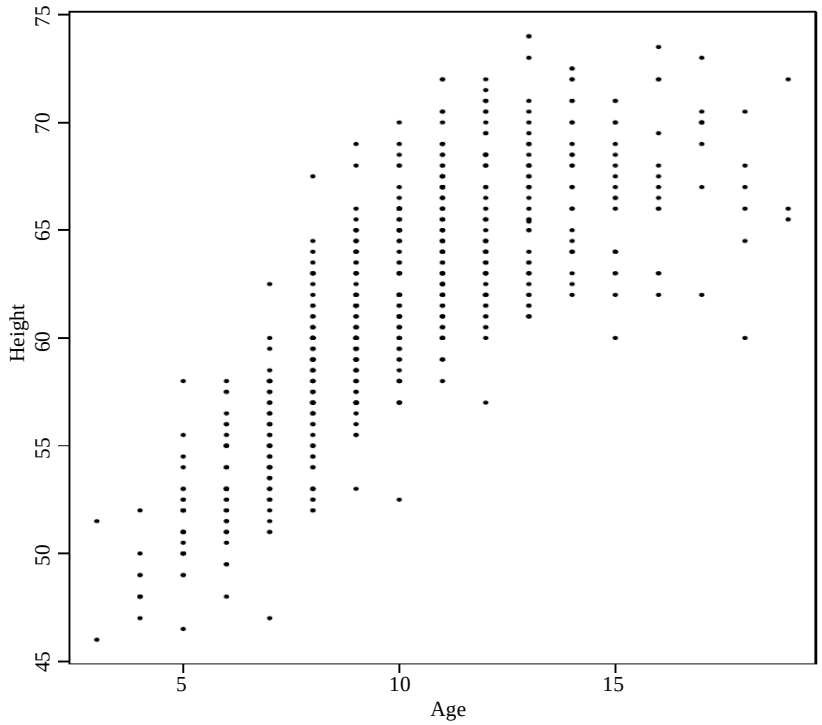

Figure 4.1 Relationship between the variable of interest *Height* and the auxiliary variable *Age*.

The objective of this simulation study was to evaluate the performance of the estimators presented in Table 4.1 using different sampling designs. We considered the Midzuno, the Sampford and the Poisson sampling designs. The variable x_2 were used as a size measure for the three sampling schemes to compute the inclusion probabilities. These sampling designs are as follows:

1. *Midzuno sampling* (see Midzuno 1952): The first unit is sampled with probability p_i and the remaining $n-1$ units are selected as a simple random sampling without replacement from the remaining $N-1$ remaining units in the population. The probabilities of selection p_i for unit i

is given by $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. The first order inclusion probability for unit i is given by $\pi_i = (N-1)^{-1}\left[(N-n)\, p_i + (n-1)\right]$.

2. *Sampford sampling* (see Sampford 1967): The algorithm for selecting the sample is carried out as follows. The first unit is selected with probability $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$ and the remaining $n-1$ units are selected with replacement with probability $\lambda_i = (1 - np_i)^{-1}\, p_i$. If any of the units are selected more than once, the procedure is repeated until all elements of the sample are different. The probability of inclusion of the first order is given by $\pi_i = np_i$.
3. *Poisson sampling*: Each unit is selected independently, resulting in a random sample size. The probability of selecting unit i is $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. The inclusion probability associated with unit i is $\pi_i = np_i$. A good description of this procedure can be found in Särndal et al. (1992).

The total of $Y = \sum_{i \in U} y_i$ was the parameter of interest. Based on each of these sampling schemes, we selected $R = 2{,}000$ Monte Carlo samples of size $n = 50$. Estimators in Table 4.1 were then computed for each sample. The performance of the estimators was then assessed using the Relative Bias, the Monte Carlo Relative Efficiency and the Approximate Relative Efficiency as described by the equations (4.1), (4.2) and (4.3) respectively.

4.2 Simulation 1 results

Simulation results are presented in Table 4.3. All estimators studied are approximately unbiased, and their relative bias is smaller than 1%. We discuss separately the approximate relative efficiency (AR) and the relative efficiency (RE) of the estimators when the population size N is known and unknown.

Case 1: *Population size N is known*

We compare the AR and the RE for the following estimators in Table 4.3: $\hat{Y}_{\text{GREG2}}$, $\hat{Y}_{\text{OPT2}}$, $\hat{Y}_{\text{OPT3}}$ and $\hat{Y}_{\text{POPT3}}$ for each of the three sampling designs. We can do so for almost all these estimators except for $\hat{Y}_{\text{OPT3}}$ for the Midzuno and the Sampford sampling schemes. In this case, we cannot compute $\mathbf{B}_{\text{OPT3}}$ for a similar reason as the one described in Remark 3.2.

On the basis of both AR and RE, the pseudo-optimal estimator $\hat{Y}_{\text{OPT3}}$ is the most reliable estimator regardless of the sampling scheme. It is close to the optimal estimator $\hat{Y}_{\text{OPT2}}$ only in terms of AR. Both the RE and the AR of the optimal estimator $\hat{Y}_{\text{OPT2}}$ were not as close as expected under the Midzuno sampling design. The poor behaviour of the RE of the optimal estimator $\hat{Y}_{\text{OPT2}}$ has also been observed by Montanari (1998). Figure 4.2 explains what is happening. We observe that most estimates obtained for the optimal estimator $\hat{Y}_{\text{OPT2}}$ for the 2,000 Monte Carlo samples are close to the mean. However, in some samples, the estimates are quite far from it. This is in contrast to $\hat{Y}_{\text{POPT3}}$ where the values are tightly centered around the mean: note that the associated RE and AR are quite close to one another.

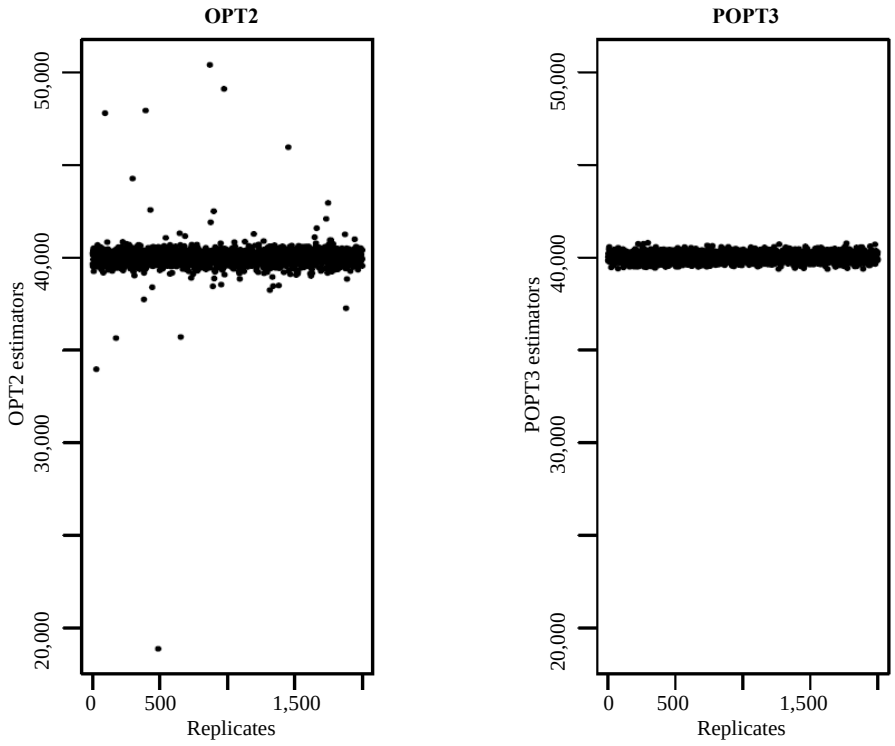

Figure 4.2 Scatter plots of Monte Carlo estimators under the Midzuno Sampling Design.

The optimal estimator $\hat{Y}_{\text{OPT3}}$ is equivalent to the pseudo-optimal estimator $\hat{Y}_{\text{POPT3}}$ in the case of Poisson sampling scheme. Recall that the optimal estimator $\hat{Y}_{\text{OPT2}}$ used $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ as auxiliary data. The optimal estimator $\hat{Y}_{\text{OPT3}}$ used $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$ as auxiliary data. The addition of the π_i has significantly improved the efficiency of the optimal estimator for the Poisson sampling scheme.

Singh and Raghunath (2011) used $\hat{Y}_{\text{SREG1}}$ when N was known, but did not include it as a control count. Nonetheless, they observed that $\hat{Y}_{\text{SREG1}}$ was quite comparable to $\hat{Y}_{\text{GREG2}}$ in terms of AR and RB for the Midzuno sampling design. The reason for this is that this sampling scheme is quite close to simple random sampling without replacement. However, using these two measures, $\hat{Y}_{\text{SREG1}}$ is by far the worst estimator for the other two sampling schemes.

Case 2: *Population size N is unknown*

Five estimators are reported in Table 4.3 for this case. However, as $\hat{Y}_{\text{KREG2}}$ is quite close to $\hat{Y}_{\text{KOPT2}}$ and $\hat{Y}_{\text{POPT2}}$, we comment on the results obtained for $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$. Estimators $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$ were very similar in terms of relative efficiency and approximate relative efficiency for the Midzuno sampling design. For the Sampford sampling scheme, $\hat{Y}_{\text{OPT1}}$, $\hat{Y}_{\text{KREG2}}$ and $\hat{Y}_{\text{POPT2}}$ were comparable and slightly better than $\hat{Y}_{\text{SREG1}}$. Under the Poisson sampling scheme, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$ outperformed $\hat{Y}_{\text{SREG1}}$. We can also see that $\hat{Y}_{\text{SREG1}}$ was very inefficient with an RE at least 10 times larger than those associated with $\hat{Y}_{\text{KREG2}}$ or $\hat{Y}_{\text{POPT2}}$. Note that $\hat{Y}_{\text{KREG2}}$ was better than $\hat{Y}_{\text{OPT1}}$: this is reasonable as $\hat{Y}_{\text{KREG2}}$ uses two auxiliary variables whereas $\hat{Y}_{\text{OPT1}}$ uses the single auxiliary variable x_{2i}.

Table 4.3
Comparison of estimators in terms of relative bias and relative efficiencies

		Population size known				Population size unknown				
		$\hat{Y}_{GREG2}$	$\hat{Y}_{OPT2}$	$\hat{Y}_{OPT3}$	$\hat{Y}_{POPT3}$	$\hat{Y}_{SREG1}$	$\hat{Y}_{OPT1}$	$\hat{Y}_{KREG2}$	$\hat{Y}_{KOPT2}$	$\hat{Y}_{POPT2}$
Midzuno	RB (in %)	0.08	0.04		0.07	0.07	0.07	0.07		0.07
	RE	1.00	5.84		0.54	0.94	0.93	0.93		0.93
	AR	1.00	0.55		0.55	0.94	0.93	0.93		0.93
Sampford	RB (in %)	0.11	0.11		0.07	-0.01	0.07	0.02		0.02
	RE	1.00	0.59		0.58	14.72	13.69	13.55		13.56
	AR	1.00	0.55		0.56	15.77	14.39	14.39		14.40
Poisson	RB (in %)	0.11	0.11	0.08	0.08	0.09	0.14	0.16	0.16	0.16
	RE	1.00	0.96	0.57	0.57	160.47	15.49	13.85	13.85	13.85
	AR	1.00	0.96	0.55	0.56	180.36	16.73	14.40	14.39	15.73

Note: We do not provide results for $\hat{Y}_{OPT3}$ and $\hat{Y}_{KOPT2}$ for the Midzuno and Sampford designs because the variance-covariance matrix is not invertible.

4.3 Simulation 2

The performance of the estimators was assessed for different values of the intercept in the model. We restricted ourselves to the Poisson sampling design to illustrate Remark 2.1 in Section 2: that is the efficiency of $\hat{Y}_{SREG}$ deteriorates as the intercept gets bigger. The population was generated according to the following model

$$y_i = a + x_i + e_i. \tag{4.4}$$

The e_i values were generated from a normal distribution with mean 0 and variance $\sigma_i^2 = 1$. The x values were generated according to a chi-square distribution with one degree of freedom. Three populations of size $N = 5{,}000$ were generated using (4.4) with different values of the intercept a. Note that $x-$values were re-generated for each population. The three populations were labelled as A, B and C depending on the intercept used. The intercept values were set to 3, 5 and 10 respectively for populations A, B and C. From each of these populations we drew $R = 2{,}000$ Monte Carlo samples with expected sample size $n = 50$ using the Poisson sampling design. The first inclusion probability was set equal to $\pi_i = nz_i \Big/ \sum_{i \in U} z_i$ for each unit i. The z values were generated according to the following model

$$z_i = 0.5y_i + u_i,$$

where u_i was a random error generated according to an exponential distribution with mean k equals to 0.5 or 1.

4.4 Simulation 2 results

Numerical results are given in Table 4.4 for $k = 1$ and Table 4.5 for $k = 0.5$. All estimators are approximately unbiased with relative biases smaller than 1%.

Case 1: *Population size N is known*

As expected, both optimal estimators $\hat{Y}_{\text{OPT2}}$ and $\hat{Y}_{\text{OPT3}}$ are more efficient than $\hat{Y}_{\text{GREG2}}$. The optimal estimator $\hat{Y}_{\text{OPT2}}$ based on $(1, x_{2i})^{\text{T}}$ is slightly better than $\hat{Y}_{\text{GREG2}}$. The inclusion of the additional variable π_i resulting in $\hat{Y}_{\text{OPT3}}$ yields significant gains in terms of RE and AR: these gains decrease as the intercept gets larger. Once more, $\hat{Y}_{\text{SREG1}}$ is quite inefficient, and as noted in Remark 2.1, this inefficiency increases as the intercept gets larger. The previous observations are valid regardless of k. The efficiency of both optimal estimators $\hat{Y}_{\text{OPT2}}$ and $\hat{Y}_{\text{OPT3}}$ decreases as k gets smaller.

Case 2: *Population size N unknown*

The most efficient estimator is $\hat{Y}_{\text{KREG2}}$. It outperforms $\hat{Y}_{\text{OPT1}}$ as it uses more auxiliary variables. Estimator $\hat{Y}_{\text{SREG1}}$ is by far the most inefficient one. As the intercept in the population model increases, the relative efficiency (both in terms of RE and AR) is fairly stable for $\hat{Y}_{\text{KREG2}}$. On the other hand, the relative efficiencies associated with $\hat{Y}_{\text{SREG1}}$ and $\hat{Y}_{\text{OPT1}}$ deteriorate rapidly, as the intercept in the population model increases. The effect of k on the efficiencies of the estimators is as described when the population size is known.

Table 4.4
Relative bias and relative efficiencies of the estimators for $k = 1$ under Poisson sampling design

Intercept		Population size known				Population size unknown				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (in %)	0.23	0.38	0.56	0.56	0.18	0.77	0.22	0.22	0.22
	RE	1.00	0.95	0.67	0.67	7.72	5.42	0.94	0.94	0.94
	AR	1.00	0.94	0.60	0.98	7.08	5.01	0.85	0.85	0.91
5	RB (in %)	0.04	0.07	0.18	0.18	-0.01	0.67	-0.07	-0.07	-0.07
	RE	1.00	0.99	0.76	0.76	23.91	16.63	1.50	1.50	1.50
	AR	1.00	0.98	0.70	0.73	23.48	16.20	1.45	1.45	1.52
10	RB (in %)	-0.01	-0.02	0.06	0.06	-0.57	0.79	-0.02	-0.02	-0.02
	RE	1.00	1.00	0.80	0.80	88.30	67.47	2.20	2.20	2.20
	AR	1.00	0.99	0.73	0.74	97.92	66.13	2.15	2.15	2.20

Table 4.5
Relative bias and relative efficiencies of the estimators for $k = 0.5$ under Poisson sampling design

Intercept		Population size known				Population size unknown				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (in %)	0.13	0.25	0.42	0.42	-0.18	0.54	-0.02	-0.02	-0.02
	RE	1.00	0.99	0.89	0.89	8.42	5.93	1.78	1.78	1.78
	AR	1.00	0.96	0.83	0.95	8.30	5.83	1.79	1.79	2.10
5	RB (in %)	0.03	0.09	0.22	0.22	0.72	1.49	0.18	0.18	0.18
	RE	1.00	1.00	0.91	0.91	24.35	17.39	3.26	3.26	3.26
	AR	1.00	0.98	0.88	0.94	23.83	16.41	3.15	3.15	3.54
10	RB (in %)	0.06	0.07	0.12	0.12	0.33	1.42	0.13	0.13	0.13
	RE	1.00	1.00	0.96	0.96	98.69	73.93	6.26	6.26	6.26
	AR	1.00	0.99	0.91	0.92	98.65	66.20	5.89	5.89	6.24

5 Conclusions

The regression estimator can be quite efficient if the auxiliary data that it uses are well correlated with the variable of interest. Furthermore, it requires that population totals corresponding to the auxiliary variables are available. In this article, we investigated the behavior of the regression estimator $\left(\hat{Y}_{\text{SREG}}\right)$ proposed by Singh and Raghunath (2011). This estimator uses estimated population count as a control total and the known population totals for the auxiliary variables. We compared it to the Generalized Regression estimator $\left(\hat{Y}_{\text{GREG}}\right)$, its optimal analogue $\left(\hat{Y}_{\text{OPT}}\right)$, and to an alternative estimator $\left(\hat{Y}_{\text{KREG}}\right)$ that uses the first-order inclusion probabilities and auxiliary data for which the population totals are known. As the optimal regression estimator requires the computation of second-order inclusion probabilities, we also included a pseudo-optimal estimator $\left(\hat{Y}_{\text{POPT}}\right)$ that does not require them. We investigated the properties of these estimators in terms of bias and efficiency via a simulation that included various sampling designs, and different values of the intercept in the model for a generated artificial population. We compared the results when the population size was known and unknown.

When the population size is known, the most efficient estimator is the optimal estimator $\hat{Y}_{\text{OPT}}$. However, since this estimator can be unstable, the pseudo-optimal estimator $\hat{Y}_{\text{POPT}}$ is a good alternative to it. This is in line with Rao (1994) who favoured the optimal estimator $\hat{Y}_{\text{POPT}}$ over the Generalized Regression estimator $\hat{Y}_{\text{GREG}}$. The Singh and Raghunath (2011) proposition to use $\hat{Y}_{\text{SREG}}$ is not viable, as it can be quite inefficient. When the population size is not known, the alternative regression estimator $\hat{Y}_{\text{KREG}}$ is the best one to use.

Acknowledgements

The authors kindly acknowledge suggestions for improved readability provided by the Associate Editor and the referees.

References

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimators and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 1, 69-77.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary data information at the estimation stage. *Journal of Official Statistics*, 10(2), 153-165.

Rosner, B. (2006). *Fundamentals of Biostatistics*. Sixth edition, Duxbury Press.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of section. *Biometrika*, 54, 499-513.

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

Singh, S., and Raghunath, A. (2011). On calibration of design weights. *METRON International Journal of Statistics*, vol. LXIX, 2, 185-205.