Revista Colombiana de Estadística
Junio 2009, volumen 32, no. 1, pp. 123 a 143

Estimation of the Population Total using the Generalized Difference Estimator and Wilcoxon Ranks

Estimación del total poblacional usando el estimador de diferencia generalizada y los rangos de Wilcoxon

Hugo Andrés Gutiérrez[1,a], F. Jay Breidt[2,b]

[1]Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

[2]Department of Statistics, Colorado State University, Fort Collins, USA

Abstract

This paper presents a new regression estimator for the total of a population created by means of the minimization of a measure of dispersion and the use of the Wilcoxon scores. The use of a particular nonparametric model is considered in order to obtain a model-assisted estimator by means of the generalized difference estimator. First, an estimator of the vector of the regression coefficients for the finite population is presented and then, using the generalized difference principles, an estimator for the total a population is proposed. The study of the accuracy and efficiency measures, such as design bias and mean square error of the estimators, is carried out through simulation experiments.

***Key words*:** Finite population, Regression estimator, Wilcoxon score.

Resumen

Este artículo presenta un nuevo estimador de regresión para el total poblacional de una característica de interés, creado por la minimización de una medida de dispersión y el uso de los puntajes de Wilcoxon. Se considera el uso de un modelo no paramétrico con el fin de obtener un estimador asistido por modelos, que surge del estimador de diferencia gene ralizada. En primer lugar, se presenta un nuevo estimador del vector de coeficientes de regresión y luego, haciendo uso de los principios del estimador de diferencia generalizada, se propone un estimador para el total poblacional. El estudio de las medidas de precisión y eficiencias, como el sesgo y el error cuadrático medio, se lleva a cabo mediante experimentos de simulación.

***Palabras clave*:** estimador de regresión, población finita, puntaje de Wilcoxon.

[a]Director. E-mail: hugogutierrez@usantotomas.edu.co
[b]Professor and Chair. E-mail: jbreidt@stat.colostate.edu

1. Introduction

In survey sampling, some auxiliary variables are commonly incorporated in the estimation procedure by using a model, but the inferences are still design-based; this kind of approach is called *model-assisted*. In this approach, the model is used to increase the efficiency of the estimators, but even when the model is not correct, estimators typically remain design-consistent, as Breidt & Opsomer (2000, page 1026) claim. Auxiliary information on the finite population is often used to increase the precision of estimators of the population mean, total or the distribution function (Wu & Sitter 2001). As an example, the ratio estimator contains known information (population total) of some auxiliary variable. There are several methods that can be called model-assisted, but most of them have only been discussed in the context of linear parametric regression models. The main examples are the generalized regression estimators (GREG) (Cassel et al. 1976*a*, Särndal 1980), the calibration estimators (Deville & Särndal 1992), and empirical likelihood estimators (Chen & Qin 1993).

In this research, the use of some nonparametric models is considered in order to obtain a model-assisted estimator by means of the generalized difference estimator proposed by Cassel et al. (1976*b*). Specifically, we consider rank-based regression methods in order to describe the relationship between auxiliary variables and the study variable and also to improve the efficiency of the estimates. The results of several simulations done in this research show that the proposed estimator works very well under particular conditions found in the survey sampling context. In the following sections, the minimum dispersion criterion (Jaeckel 1972, Jurečková 1971) is used in order to build a rank-based sampling estimator of the regression coefficients. A comparison of the two approaches is achieved through Monte Carlo simulations where it could be observed that the proposed rank-based estimator gains in efficiency and its relative bias is negligible.

The outline of the paper is as follows: After a short introduction that describes briefly the model-assisted approach in survey sampling, Section 2 is focused in the construction of an estimator of regression coefficients obtained by the minimum dispersion approach. In Section 3, the generalized difference estimator is considered in order to build an estimator of the population total by means of results obtained in Section 2. Also in Section 3, some theoretical properties of the proposed estimator are reviewed. In Section 4, some empirical simulations show the good performance, in terms of low relative bias and high efficiency, of the proposed estimator -which is compared to traditional estimators under several scenarios- supported by favorable results in most cases.

1.1. Framework

Consider a finite population as a set of units $\{u_1, \ldots, u_k, \ldots, u_N\}$, where each one can be identified without ambiguity by a label. Let $U = \{1, \ldots, k, \ldots, N\}$ denote the set of labels. The size of the population N is not necessarily known.

The aim is to study a variable of interest y that takes the value y_k for unit k. Note that the y_k's are not random. The objective is to estimate a function of interest T of the y_k's:

$$T = f(y_1, \ldots, y_k, \ldots, y_N) \tag{1}$$

The most common functions are the population total, given by

$$t_y = \sum_{k \in U} y_k \tag{2}$$

and the population mean, given by

$$\overline{y}_U = \frac{\sum_{k \in U} y_k}{N} \tag{3}$$

Associated with the kth unit ($k = 1, \ldots, N$), there is a column vector of p auxiliary variables, $\mathbf{x}_k$. It is assumed that the population totals $\mathbf{t_x} = \sum_U \mathbf{x}_k$ are known.

A probability sample s is drawn from U, according to a sampling design $p(\cdot)$. Note that $p(s)$ is the probability of drawing the sample s. The sample size is $n(s)$, but, for a fixed size sampling design, the sample size is n. The sampling design determines the first order inclusion probability of the unit k, π_k, defined as

$$\pi_k = Pr(k \in s) = \sum_{s \ni k} p(s) \tag{4}$$

and the second order inclusion probability of the units k and l, defined as

$$\pi_{kl} = Pr(k, l \in s) = \sum_{s \ni k,l} p(s) \tag{5}$$

The study variable y is observed for the units in the sample.

The foundations of inference in survey sampling are based in pursuing a sampling strategy, that is the combination of a sampling design and an estimator. In this research it is assumed that the user knows the population behavior of the response variable, and chooses the appropriate sampling design. In this way, the pursuit is restricted to the estimator. Some sampling estimators for the total of a population are as follows.

1.1.1. The Horvitz-Thompson Estimator

The Horvitz-Thompson (HT) estimator (Horvitz & Thompson 1952) is defined by

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \tag{6}$$

This estimator is design-unbiased, that is $E_p\left(\widehat{t}_\pi\right) = t_y$ where $E_p(\cdot)$, denotes the expectation with respect to the sample design. Its variance is given by

$$Var_p\left(\widehat{t}_\pi\right) = \sum_{k,l \in U} \frac{y_k y_l \Delta_{kl}}{\pi_k \pi_l} \tag{7}$$

For more information about the properties of this estimator it is recommended to review Särndal et al. (1992, Ch. 2).

1.1.2. The Generalized Regression Estimator

The HT estimator does not use the auxiliary information in the estimation step[1]. However, it is of interest to improve its efficiency by using the auxiliary information. For this purpose, we suppose that the relationship between y_k and $\mathbf{x}_k$ could be described by a model (Cassel et al. 1976*b*) ξ, such that $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$ and

$$\begin{aligned} E_\xi(y_k) &= \mathbf{x}_k'\boldsymbol{\beta} \\ Var_\xi(y_k) &= \sigma_k^2 \end{aligned} \tag{8}$$

for $k = 1, \ldots, N$, where ε_k are independent random variables with mean zero and variance σ_k^2 and $\boldsymbol{\beta}$ is a vector of unknown constants. If (8) is adjusted with an intercept, then $x_{1k} \equiv 1 \; \forall k \in U$. Cassel et al. (1976*a*, p. 81) claim that the finite population is actually drawn from a larger universe and this is the model idea in "its most pure form". Särndal et al. (1992, pp. 225 - 226) explain that the hypothetical finite population fit of the model would result in estimating $\boldsymbol{\beta}$ by

$$\mathbf{B} = \left(\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \sum_U \frac{\mathbf{x}_k y_k}{\sigma_k^2} \tag{9}$$

When a sample s is drawn, $\mathbf{B}$ is estimated by

$$\widehat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \right)^{-1} \sum_s \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \tag{10}$$

Then, the Generalized Regression Estimator (GREG) (Cassel et al. 1976*b*) is given by

$$\widehat{t}_{GREG} = \widehat{t}_\pi + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} \right)' \widehat{\mathbf{B}} \tag{11}$$

where $\widehat{\mathbf{B}}$ is the vector of estimated regression coefficients.

[1]The HT estimator does not use auxiliary information explicitly. However, auxiliary information is often used implicitly in developing inclusion probabilities (as in a probability proportional to size sampling design) or developing stratification.

Särndal et al. (1989) give the approximate variance of the GREG as follows:

$$Var_p\left(\widehat{t}_{GREG}\right) \simeq \sum_{k,l\in U} \Delta_{kl} \frac{(y_k - \mathbf{x}_k'\mathbf{B})}{\pi_k} \frac{(y_l - \mathbf{x}_l'\mathbf{B})}{\pi_l} \tag{12}$$

which is small if y_k is well explained by the vector of auxiliary variables, $\mathbf{x}_k$. Isaki & Fuller (1982) and Deville & Särndal (1992) present the theoretical background of this estimator.

2. Estimating the Regression Coefficients

In this section, the traditional least squares estimation method of the vector of regression coefficients under the assumption of the model given in equation (8) is reviewed. After this, a new estimator of the vector of regression coefficients is obtained through the minimum dispersion approach.

2.1. Least Squares Estimation

When the least squares approach is used, $\boldsymbol{\beta}$ is estimated by (9). By using the principles of estimation proposed by (Horvitz & Thompson 1952), when a sample s is drawn, $\mathbf{B}$ is estimated by (10)[2] and its variance expression must be found. Särndal et al. (1992, section 5.10) show that when using the Taylor approach, an approximation of the variance of (10) is given by

$$AV\left(\widehat{\mathbf{B}}\right) = \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}_k'}{\sigma_k^2}\right)^{-1} \mathbf{V} \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}_k'}{\sigma_k^2}\right)^{-1} \tag{13}$$

where $\mathbf{V}$ is a symmetric matrix $p \times p$ with entries

$$v_{ij} = \sum\sum_U \Delta_{kl} \left(\frac{x_{ik}E_k}{\pi_k}\right)\left(\frac{x_{jl}E_l}{\pi_l}\right) \tag{14}$$

and $E_k = y_k - \mathbf{x}_k'\mathbf{B}$. The variance estimator is given by

$$\widehat{V}\left(\widehat{\mathbf{B}}\right) = \left(\sum_s \frac{\mathbf{x}_k\mathbf{x}_k'}{\sigma_k^2\pi_k}\right)^{-1} \widehat{\mathbf{V}} \left(\sum_s \frac{\mathbf{x}_k\mathbf{x}_k'}{\sigma_k^2\pi_k}\right)^{-1} \tag{15}$$

where $\mathbf{V}$ is a symmetric matrix $p \times p$ with entries

$$\widehat{v}_{ij} = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{x_{ik}e_k}{\pi_k}\right)\left(\frac{x_{jl}e_l}{\pi_l}\right) \tag{16}$$

and $e_k = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}$. Note that $i, j = 1, \ldots, p$. In this research, we assume a model like 8 supposing that ε_k deviates from the Gaussian distribution.

[2]Note that (10) is biased but asymptotically design unbiased and consistent under mild assumptions.

Besides this particular case, if the scatterplot shows some points of influence or some outliers, as in Figure 1, the use of the least squares approach is not suitable in order to estimate $\widehat{\mathbf{B}}$. Jaeckel (1972) proposes some alternatives to find a nonparametric estimate of the vector of coefficients. As usual, in a linear model, the problem is to find those values of the coefficients which make the residuals as small as possible.

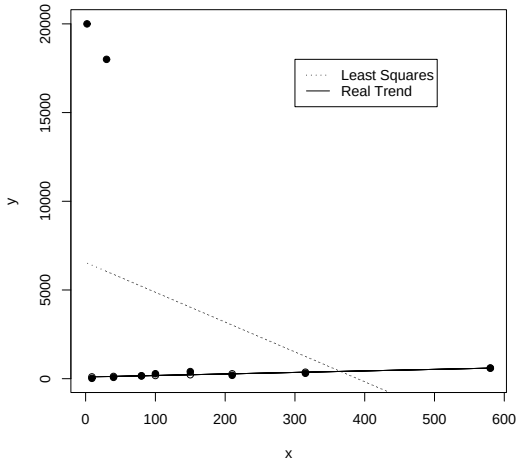

FIGURE 1: Two regression lines in the finite population.

2.2. Estimation of B through the Minimum Dispersion approach

Without loss of generality, it is supposed that the k-th unit has only one auxiliary variable associated. The reason for this is the convenience for the theoretical development, but the reader must note that the estimation of the regression coefficients can be extended to the multiparameter case. Then $y_1, \ldots, y_k, \ldots, y_N$ is a realization of a linear working model, $\xi : y_k = \beta_0 + \beta x_k + \varepsilon_k$, where we denote by $F_\varepsilon(\cdot)$ the continuous distribution function of the residuals of this model and $f_\varepsilon(\cdot)$ their corresponding probability density function.

The following definitions (Hettmansperger 1984, section 3.4) are required in order to develop an estimator that could be considered as suitable under the former assumptions.

Definition 1. Let $D(\cdot)$ be a measure of variability in the finite population that satisfies the following properties:

1. $D(\mathbf{E} + \mathbf{1}_N c) = D(\mathbf{E})$

2. $D(-\mathbf{E}) = D(\mathbf{E})$

for any $N \times 1$ vector $\mathbf{E}$ and any scalar c. Note that $\mathbf{1}_N$ is a vector of ones of size N. Then $D(\cdot)$ is called a translation-invariant measure of dispersion.

Let $\mathbf{x}$ be a vector of size N of known auxiliary information and $\mathbf{y} = (y_1, \ldots, y_k, \ldots, y_N)'$. By minimizing $D(\mathbf{y} - \beta\mathbf{x})$, an estimate of β, generated by $D(\cdot)$, is obtained. Jaeckel (1972) defined the following measure of dispersion for any vector $\mathbf{E} = \mathbf{y} - \beta\mathbf{x}$

$$D(\mathbf{E}) = \sum_{k=1}^{N} a(R_k)E_k \tag{17}$$

where $R_1, \ldots, R_N$ are the ranks of $E_1, \ldots, E_N$, and $a(k)$ are a non-decreasing set of scores. Using the former measure we have the following definition.

Definition 2. A rank estimate of β is the value b that minimizes

$$D(\mathbf{E}) = \sum_{k=1}^{N} a\big[R(E_k)\big](E_k) \tag{18}$$

where $E_k = y_k - \beta_0 - \beta x_k$ and $\mathbf{E} = (E_1, \ldots, E_k, \ldots, E_N)$

Note that we shall not estimate β_0, through the dispersion measure approach, because of the first condition of the Definition 1, which implies that the estimate of β does not depend on β_0. We expect that the estimates generated by minimizing 18 will be more robust than least-square estimates because the influence of outliers enters in a linear rather than quadratic fashion.

Result 1. *Without loss of generality, the estimate b that minimizes 18 is the same as the one that minimizes the measure of dispersion in terms of centered data.*

Proof. Using the properties of a measure of variability, we have that:

$$\begin{aligned} D(\mathbf{E}) &= \sum_{k=1}^{N} a\big[R(E_k)\big](E_k) \\ &= \sum_{U} a(R(y_k - b_0 - bx_k))(y_k - b_0 - bx_k) \\ &= \sum_{U} a(R(y_k - \overline{y}_U - bx_k + b\overline{x}_U))(y_k - \overline{y}_U - bx_k + b\overline{x}_U) \\ &= \sum_{U} a\Big(R\big(y_k^c - bx_k^c\big)\Big)\big(y_k^c - bx_k^c\big) \end{aligned} \tag{19}$$

where $y_k^c = y_k - \overline{y}_U$ and $x_k^c = x_k - \overline{x}_U$. □

Result 2. *Jaeckel (1972, theorem 4) states that when using the Wilcoxon scores, defined by*[3,4]

$$a(k) = \frac{k}{N+1} - \frac{1}{2} \tag{20}$$

then (18) is minimized when b, the estimator of β, is the weighted median of the set of pairwise slopes given by

$$b_{kl} = \frac{y_k - y_l}{x_k - x_l} \qquad k, l = 1, \ldots, N \tag{21}$$

for $x_k > x_l$, where each slope is weighted proportional to $x_k - x_l$, and b_{kl} are assumed all distinct.

Note that 18 is translation-invariant, so we can obtain the estimate b_0 as the median of $y_k - bx_k$. Draper (1988) explains that "to calculate a weighted median, sort the observations from smallest to largest, carrying their weights along with them, find the overall sum S of the weights, and begin adding the weights from the top or bottom of the sorted list until $S/2$ is reached. The corresponding observation is the weighted median".

The estimator of β_0 in the finite population is given by the following result.

Result 3. *Let b be the estimator of β which minimizes 18. Then the estimator of β_0 which satisfies the condition that the median point* $(\text{med}(x), \text{med}(y))$ *must lie in the regression line is given by*

$$b_0 = \text{med}(y - bx) \tag{22}$$

2.2.1. Slope Estimation

In practice, we just have a sample of the finite population, so that both b_0 and b remain unknown, but can be estimated by a sample estimator involving the inclusion probability of each element in the selected sample.

Result 4. *A rank-based sampling estimator of the slope regression coefficient is given by $\widetilde{b}$, which is a weighted median of*

$$\widetilde{b}_{kl} = \frac{\check{y}_k^c - \check{y}_l^c}{\check{x}_k^c - \check{x}_l^c} \tag{23}$$

for $\check{x}_k^c > \check{x}_l^c$, where each term is weighted proportional to $\check{x}_k^c - \check{x}_l^c$. With

$$\check{y}_k^c = \frac{y_k - \overline{y}_s}{\pi_k}$$

[3]The Wilcoxon procedures are robust and highly efficient in the sense that the effect of outliers (in the variable of interest) is smaller than the least squares procedures; *i.e.*, Wilcoxon procedures provide protection against outlying responses, see (Terpstra & McKean 2005).

[4]This paper considers only the case where the set of scores corresponds to the Wilcoxon scores. The reason is that Wilcoxon procedures are more efficient than least squares procedures when the data are non-normal and feature 95.5% efficiency when the data are normally distributed (Hettmansperger & McKean 1998, pp.163-164). There are many other choices for the set of scores and could be considered for future research.

$$\breve{x}_k^c = \frac{x_k - \overline{x}_s}{\pi_k}$$

$\overline{y}_s$ *and* $\overline{x}_s$ *are the sample mean of the response variable and the sample mean of the auxiliary variable, respectively. Note that* $\widetilde{b_{kl}}$ *are assumed to be all distinct with* $k, l = 1, \ldots, n$.

Proof. The former estimator is quite intuitive: from the Result 1, we obtained that the measure of dispersion to minimize is $D = \sum_U a\Big(R\big(y_k^c - bx_k^c\big)\Big)\Big(y_k^c - bx_k^c\Big)$. As it was mentioned, in practice the y_k's are not available in the whole population, so a natural estimation of D is given by including the first order inclusion probabilities in the measure, as follows:

$$\begin{aligned}
\widetilde{D} &= \sum_s \frac{a\left(R\left(\frac{y_k^c - bx_k^c}{\pi_k}\right)\right)\left(y_k^c - bx_k^c\right)}{\pi_k} \\
&= \sum_s a\left(R\left(\frac{y_k^c - bx_k^c}{\pi_k}\right)\right)\left(\frac{y_k^c - bx_k^c}{\pi_k}\right) \\
&= \sum_s a\Big(R\big(\breve{y}_k^c - b\breve{x}_k^c\big)\Big)\Big(\breve{y}_k^c - b\breve{x}_k^c\Big)
\end{aligned} \tag{24}$$

Then, the proof is complete when using Result 1. □

There are many choices in the estimation of the population dispersion, the reason that we use π-expansion in the denominator of expression (24) is that D could be seen as a population total and its corresponding HT estimator must be a sample total expanded by the inclusion probability of each unit in the selected sample, s. The π-expansion is included in the rank function $R(\cdot)$ because it must maintain the original weights given by the inclusion probability of each element. Note that (24) takes a form similar to (19), and applying the Result 2 an estimator of b is obtained.

2.2.2. Intercept

The estimation of the intercept b_0 can be found by estimating the median (with respect to the pseudo-residuals $e_k^* = y_k - \widetilde{b}x_k$, $k = 1, \ldots, n$) of the finite population.

Result 5. *A sampling estimator of the intercept regression coefficient is given by* $\widetilde{b}_0$

$$\widetilde{b}_0 = \widetilde{F}^{-1}(0.5) \tag{25}$$

$\widetilde{F}^{-1}$ *is the inverse function of* $\widetilde{F}(0.5)$ *given by*

$$\widetilde{F}(0.5) = \sum_s \frac{z_{k,0.5}}{\pi_k}\left(\sum_s \frac{1}{\pi_k}\right)^{-1} \tag{26}$$

and

$$z_{k,0.5} = \begin{cases} 1 & if\ e_k^* \leq 0.5, \\ 0 & if\ e_k^* > 0.5 \end{cases} \quad where \quad e_k^* = y_k - \widetilde{b}x_k \tag{27}$$

The general procedure suggested for the estimation of a median has the following steps (Särndal et al. 1992, p. 197):

1. First, obtain the estimated distribution function, $\widetilde{F}$
2. Estimate the median by $\widetilde{F}^{-1}(0.5)$.

2.3. Properties of the Rank-Based Estimator of Regression Coefficients

In this section, the results of a Monte Carlo simulation are used in order to show that the rank-based estimator of the regression coefficients has a good performance (lower relative bias and mean square error than the least squares approach) under two specific scenarios.

A size $N = 1000$ finite population is simulated from a superpopulation model, ξ. To do this, it is supposed that the relationship between y_k and x_k can be described through a model ξ, such that $y_k = \beta_0 + \beta x_k + \varepsilon_k$ and

$$\begin{aligned} E_\xi(y_k) &= \beta_0 + \beta x_k \\ Var_\xi(y_k) &= \sigma_k^2 \end{aligned} \tag{28}$$

The first simulation scenario is when the values of x come from a gamma distribution with scale and shape parameter equal to one[5]. The second scenario is similar to the first, but 5% of the data in the response variable is contaminated. This process was done by contaminating the errors through a mixture of normal densities. The R code of the contamination step is available by requesting to the first author. Figure 2 shows the corresponding scatterplot for the second scenario.

The value of the parameter β was set to two and the value of the parameter β_0 was set to ten such that $y_k > 0 \forall k \in U$. For the non-contaminated units, it is assumed that ε_k are independent and identically distributed as $N(0, \sigma^2)$.

In each run of the simulation, random samples were drawn, according to a simple random sampling design without replacement (SI). Each sample was of size $n = 100$. The parameters were estimated using least squares and the minimum dispersion approach. This process was repeated $M = 1000$ times. The simulation was written in the statistical software R 2.6.0. (Team 2007). In the simulation, the performance of an estimator $\widehat{b}$ was evaluated with its relative bias, (RB) and its mean square error, (MSE), defined as follows:

[5]The values of this distribution are non negative and its shape is right-skewed. This is common in practice (Wu 2003, p. 946).

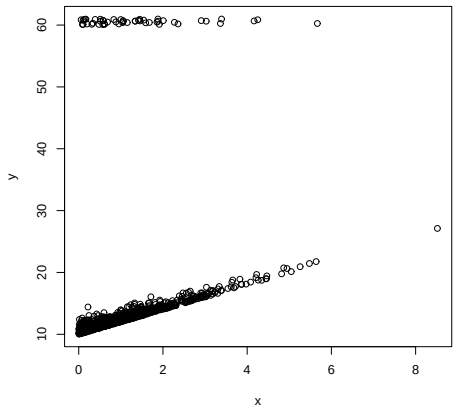

FIGURE 2: Scatter plot of the contaminated response variable against the simulated auxiliary variable.

$$RB = 100\% M^{-1} \sum_{m=1}^{M} \frac{\widehat{b}_m - \beta}{\beta} \tag{29}$$

$$MSE(\widehat{b}) = M^{-1} \sum_{m=1}^{M} (\widehat{b}_m - \beta)^2 \tag{30}$$

respectively, and $\widehat{b}_m$ was computed in the m-th simulated sample.

Table 1 shows the relative bias of the estimators of β_0 and β. The sampling estimators based in the minimization of the sampling dispersion through Wilcoxon ranks have a smaller bias than the least squares estimators under a normal model with no contaminated data. The difference is huge under a normal model with contaminated data in the response variable demonstrating the robustness of the proposed estimator.

TABLE 1: Relative bias of the estimators.

	Minimum dispersion		Least Squares	
	β_0	β	β_0	β
Not contaminated	-0.37%	-0.33%	-0.51%	-0.62%
Contaminated	-3.98%	-0.19%	-33.94%	23.09%

Regardless to the efficiency of the proposed estimators, Table 2 shows that under a model with contaminated data, the estimator performs well and it could be stated that the fit is good in comparison with the least squares estimator. The proposed estimator gains in efficiency under the model with contaminated data; this gain is very high in the slope estimation of the regression line.

Table 2: Mean square error of the estimators.

	Minimum dispersion		Least Squares	
	β_0	β	β_0	β
Not contaminated	0.13	0.0004	0.25	0.0001
Contaminated	0.16	0.0005	1.15	0.21

3. Estimating the Population total Through Minimum Dispersion

If b_0 and b were known, then a design-unbiased estimator of the population total could be constructed using the generalized difference estimator (Cassel et al. 1976*b*) given by the following expression:

$$\widehat{t}_y = \sum_{k\in s} \frac{y_k - b_0 - bx_k}{\pi_k} + \sum_{k\in U}(b_0 + bx_k) \tag{31}$$

The design variance of this estimator is given by

$$Var\left(\widehat{t}_y\right) = \sum\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k}\frac{E_l}{\pi_l} \tag{32}$$

where $E_k = y_k - b_0 - bx_k$. It is expected that this variance would be smaller than the variance of the HT estimator.

In practice, we have only a sample of the finite population, so that both b_0 and b remain unknown, but can be estimated by a sample estimator involving the inclusion probability of each element in the selected sample.

Result 6. *A rank-based survey regression estimator for the population total is given by the following expression*

$$\widetilde{t}_y = \sum_{k\in s} \frac{y_k - \widetilde{b}_0 - \widetilde{b}x_k}{\pi_k} + \sum_{k\in U}\left(\widetilde{b}_0 + \widetilde{b}x_k\right) \tag{33}$$

where $\widetilde{b}$ *is given by* (23) *and* $\widetilde{b}_0$ *is given by* (25).

3.1. Properties of the Rank-Based Estimator of the Population Total

It is straightforward to show that (33) can be written as

$$\widetilde{t}_y = \widehat{t}_{y\pi} + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)'\widetilde{\mathbf{B}} \tag{34}$$

where $\widehat{t}_{y\pi}$ is the HT estimator for the variable of interest, $\mathbf{t_x} = \left(N, \sum_U x_k\right)'$, $\widehat{\mathbf{t}}_{\mathbf{x}\pi} = \left(\sum_s \frac{1}{\pi_k}\sum_s \frac{x_k}{\pi_k}\right)'$ and $\widetilde{\mathbf{B}} = \left(\widetilde{b}_0, \widetilde{b}\right)'$.

3.1.1. Simple Random Sampling

If simple random sampling without replacement is considered, then $\pi_k = \dfrac{n}{N}$ and $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$ for $l \neq l$. For this sampling design, the rank-based regression estimator is defined as

$$\widetilde{t}_y = \frac{N}{n}\sum_s y_k + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)'\widetilde{\mathbf{B}} \tag{35}$$

where $\widehat{\mathbf{t}}_{\mathbf{x}\pi} = \left(N, \frac{N}{n}\sum_s x_k\right)'$ and $\widetilde{\mathbf{B}} = \left(\widetilde{b}_0, \widetilde{b}\right)'$. Under SI, $\widetilde{b}$ is the median of the set of pairwise slopes given by:

$$b_{kl} = \frac{y_k - y_l}{x_k - x_l} \tag{36}$$

and $\widetilde{b}_0$ is the median of $y_k - \widetilde{b}x_k$. Note that the second term of (35) can be considered as a rank-based correction for the estimated population total.

3.1.2. Variance Estimation Through the Difference Estimator

If it is suspected that the variability in $\widetilde{\mathbf{B}}$ is dominated by the variability in $\widehat{t}_{y\pi}$ and $\widehat{\mathbf{t}}_{\mathbf{x}\pi}$, then

$$\widetilde{t}_y - t_y = \left(\widehat{t}_{y\pi} - t_y\right) + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)'\mathbf{B} + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)'\left(\widetilde{\mathbf{B}} - \mathbf{B}\right) \tag{37}$$

where $\mathbf{B} = (b_0, b_1)'$ is the vector of finite population regression coefficients that would be obtained from the rank-based procedure if the entire finite population were observed. The last term above is the product of two terms, each converging to zero, and is supposed of smaller order than either of the first two terms (*small* $\times$ *small* = *negligible*). This means that the proposed rank-based model-assisted estimator is well approximated by a generalized difference estimator, from which a variance estimator could be found straightforwardly.

Under the previous scenario, it follows that the proposed estimator behaves in large samples the generalized difference estimator (Särndal et al. 1992, p. 221) and then:

$$\widetilde{t}_y \approx \widehat{t}_{y\pi} + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)'\mathbf{B}, \tag{38}$$

The design variance of this estimator is given by (32). An estimator of this variance is given by:

$$\widehat{Var}\left(\widehat{t}_y\right) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l} \tag{39}$$

where $e_k = y_k - \widetilde{b}_0 - \widetilde{b}_1 x_k$.

The rigorous study of the properties of the estimator for the variance in (39) requires theoretical sampling-based arguments that are beyond of the scope of

this research. However, in this section we will proceed through simulations to show that the difference estimator approach is reasonable. For this purpose, the performance of the variance estimator in (39) is evaluated. A finite population of size $N = 1000$ was simulated from a superpopulation model, ξ. It is supposed that the relationship between y_k and $\mathbf{x}_k$ could be described by means of the very first model ξ (non contaminated data) in the section 2.3, such that $y_k = \beta_0 + \beta x_k + \varepsilon_k$. The auxiliary information is generated in the same way as in the previous section.

In particular, the model $y_k = 10 + 2x_k + \varepsilon_k$ is considered such that $y_k > 0$ $\forall k \in U$. It is assumed that ε_k are independent and distributed as $N(0, \sigma_k^2)$.

In each run of the simulation, simple random samples were drawn; each sample was of size $n = 100$. The parameters (β_0, β_1) were estimated using the least squares approach and the minimum dispersion approach. This process was repeated $M = 1000$ times. The simulation was written in the statistical software R 2.6.0. (Team 2007). In the simulation, the performance of the proposed variance estimator using the principles of the generalized difference estimator, (44), was evaluated using the percent relative bias, $(RB\%)$ that was 0.963%. The value of the relative bias is very close to zero and even though it is an empirical exercise, the use of the estimator appears reasonable under the standard model ξ.

3.1.3. Jackknife Variance Estimation

The exact design-based variance of the proposed estimator does not have a closed form because the estimator $\widetilde{\mathbf{B}}$ is a nonlinear one. On this subject Lohr (1999, p. 293) claims that Jackknife methods are convenient for multiparameter and non-parametric problems and provides an attractive alternative in this cases.

Let $\widetilde{t}_{y_{(j)}}$ denote the estimator of the population total omitting the j-th unit. Then, for a simple random sample we define the delete 1-Jackknife variance estimator of $\widetilde{t}_y$ as

$$\widehat{V}_{JK}\left(\widetilde{t}_y\right) = \frac{n-1}{n}\sum_{j=1}^{n}\left(\widetilde{t}_{y_{(j)}} - \widetilde{t}_y\right)^2 \tag{40}$$

This method provides a consistent estimation of the variance.

3.1.4. Representative Strategies

Definition 3. Given $\mathbf{x}$ an auxiliary information vector, a sampling strategy $\left(p, \widehat{t}\right)$ is called representative with respect to $\mathbf{x}$, if and only if[6]

$$\widehat{t}(\mathbf{x}) = t_{\mathbf{x}} \tag{41}$$

for every s with $p(s) > 0$; that is, the estimator applied to the auxiliary variables reproduces exactly the population total of each auxiliary variable.

[6]The combination $\left(p, \widehat{t}\right)$ denoting an estimator $\widehat{t}$ based on a sample drawn accordingly to a design p is called a strategy.

Result 7. *Under any sampling design $p(s)$, the proposed population total estimator induces a representative strategy because the pair $\left(p(s), \widetilde{t}\right)$ estimates the population total of the auxiliary variables with null variance.*

Proof. It is straightforward to show that $\widetilde{b}_1 = 1$ because it is the weighted median of $\widetilde{b}_{kl} = \frac{\breve{x}_k^c - \breve{x}_l^c}{\breve{x}_k^c - \breve{x}_l^c}$, and $\widetilde{b}_0 = 0$ because it is the sampling estimation of $\text{med}\left(x_k - \widetilde{b}_1 x_k\right)$. Therefore,

$$\begin{aligned} \widetilde{t}_x &= \sum_s \left(\frac{x_k - \widetilde{b}_0 - \widetilde{b}_1 x_k}{\pi_k} \right) + \sum_U \left(\widetilde{b}_0 + \widetilde{b}_1 x_k \right) \\ &= \sum_s \left(\frac{x_k - 0 - x_k}{\pi_k} \right) + \sum_U (0 + x_k) \\ &= \sum_U x_k = t_x \end{aligned} \tag{42}$$

Note that $Var\left(\widetilde{t}_x\right) = Var(t_x) = 0$. □

3.1.5. Cochran-Consistency

The definition of Cochran-Consistency (Särndal et al. 1992, p. 168) claims that an estimator is consistent for a parameter in a finite population if $s = U$ implies that the estimator is equal to the parameter.

Result 8. *Under SI designs, the proposed estimator is Cochran-consistent.*

Proof. It is straightforward to show that if $s = U$ under the family of SI designs, then $\mathbf{t_x} = \widehat{\mathbf{t}}_{\mathbf{x}\pi}$, so

$$\begin{aligned} \widetilde{t}_y &= t_{y\pi} + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi} \right)' \widetilde{\mathbf{B}} \\ &= t_y + \left(\mathbf{t_x} - \mathbf{t}_x \right)' \widetilde{\mathbf{B}} = t_y \end{aligned} \tag{43}$$

□

4. Empirical Simulation

In this section, some simulation experiments are carried out in order to compare the performance of the proposed estimator given by (33) and referred to as JAC, with the Horvitz-Thompson (HT) estimator and the regression estimator (REG).

A size $N = 1000$ finite population is simulated from a superpopulation model ξ. It is supposed that the relationship between y_k and x_k can be described through a model ξ, such that $y_k = \beta_0 + \beta x_k + \varepsilon_k$ and

$$\begin{aligned} E_\xi(y_k) &= 10 + 2x_k \qquad\qquad y_k > 0 \\ Var_\xi(y_k) &= \sigma_k^2 \qquad\qquad k = 1, \ldots, n \end{aligned} \tag{44}$$

The values of the vector of auxiliary information are generated from a gamma distribution with scale and shape parameter equal to 1. It is assumed that the values of ε_k are independent and distributed as $N\left(0, \sigma_k^2\right)$. This is the real model used in the construction of the proposed estimator. Note that even though the model includes a term for the variance, the resulting rank-based estimator (JAC) does not contain this variance term nor does the HT estimator. The REG estimator takes into account the variance term, and this is an interesting feature in the simulation.

In each run, random samples according to a SI design were drawn. Each sample was of size $n = 100$. The parameters (β_0, β_1) were estimated using least squares and the minimum dispersion approach. This process was repeated $M = 1000$ times. The simulation was written in the statistical software R 2.6.0. (Team 2007). In the simulation, the performance of an estimator $\widehat{t}_y$ of t_y is tracked by the Percent Relative Bias:

$$RB = 100\% M^{-1} \sum_{m=1}^{M} \frac{\widehat{t}_{y,m} - t_y}{t_y} \tag{45}$$

and the relative efficiency

$$RE\left(\widehat{t}_y\right) = \frac{MSE\left(\widehat{t}\right)_y}{MSE\left(\widehat{t}_{y\pi}\right)} \tag{46}$$

where $\widehat{t}_{y,m}$ is computed in the mth simulated sample, $m = 1, \ldots, 1000$. The Mean Square Error (MSE) is defined by

$$MSE\left(\widehat{t}_y\right) = M^{-1} \sum_{m=1}^{M} \left(\widehat{t}_{y,m} - t_y\right)^2 \tag{47}$$

Note that the HT estimator is the baseline estimator for efficiency comparison.

Specifically, we consider the robustness and the absence of normality of the residuals as the main issues that moves us to consider a model-assisted survey rank-based regression estimator. We used the minimum dispersion procedure (Jurečková 1971) and (Jaeckel 1972) to build the proposed estimator. It is known that rank-based procedures outperform, in the sense of efficiency, traditional (least squares) procedures when the distribution function of the residuals in the model is deviated from the normal distribution (Hettmansperger 1984, Hettmansperger & McKean 1998).

The estimators are considered under a wide range of model specifications. The simplest of these ones is simple linear regression with normal, uncorrelated, homoscedastic errors. Departures from this simple model (mean function is not

linear, errors are not normal, errors are heteroscedastic) would all be of interest. It is expected that the rank-based procedure would continue to work well across a whole range of simulated models. The model specifications are as follows:

1. M1: normal linear model with correctly specified variance structure and uncorrelated, homoscedastic errors;
2. M2: normal linear model with correctly specified variance structure and uncorrelated, heteroscedastic errors[7];
3. M3: linear model with correctly specified variance structure and non-normal errors, uncorrelated, homoscedastic errors[8];
4. M4: normal linear model with incorrectly specified variance structure and uncorrelated, heteroscedastic errors[9];
5. M5: nonlinear model[10];
6. M6: normal linear model with five percent of contaminated data[11].

These cases represent a range of correct and incorrect model specifications for the estimators that are considered. Table 3 reports the simulated relative bias for the estimators. In all of these cases, the relative bias is negligible.

Table 3: Relative Bias of the estimators.

Model	HT	REG	JAC
M1	−0.024%	0.008%	0.007%
M2	0.034%	0.013%	0.014%
M3	−0.020%	0.002%	0.003%
M4	0.034%	0.014%	0.014%
M5	0.002%	0.002%	0.002%
M6	−0.009%	−0.021%	−0.023%

The motivation of this research was the construction of an estimator able to gain in efficiency compared with the traditional estimators in the survey sampling context. From results shown in Table 4, it can be seen that the proposed estimator in the M1 model gains against the others, it can be seen that the MSE of the estimators that uses the auxiliary information is less than the HT estimator in all cases, specifically, under a regular model like M1 the estimator features very well. Under the M2 model, the REG estimator loses efficiency compared with the

[7]In this step, the population values of the y_k are generated by adding $N(0, \sigma_k^2)$ errors. Notice that σ_k^2 was set in order that the model had a strong heteroscedastic structure.

[8]This model assumes errors ε_k following an exponential distribution with parameter equals to one.

[9]In this step the true model has a different variance for each sample point, but it is wrongly assumed that the model has a constant variance for all sample points.

[10]The model ξ is such that $y_k = \dfrac{1}{(10 + 2x_k)^3} + \varepsilon_k$.

[11]In this step the contaminated data follow the same specifications as in the simulation of section 2.3

proposed estimator. This is an important issue due that the proposed estimator does not take into account any term of variance. A similar situation occurs in the M3 model where the errors are non-normal; in this scenario both of the estimators features well. The results of the M4 model are the same than the M2 model for the proposed estimator because it does not take into account any variance term. When the model is not linear, all of the estimators features well. In the M6 model, when dealing with contaminated data in the response variable, the proposed estimator work very well, and in this point the gain in efficiency is higher. In general conditions, the proposed estimators plays a good role and it is supported by this empirical experiment.

Table 4: Mean Square Error of the estimators.

Model	HT	REG	JAC
M1	311.11	33.83	33.82
M2	349.38	44.23	43.12
M3	326.88	9.86	9.83
M4	349.38	43.10	43.12
M5	2.11	2.12	2.11
M6	509	244.62	243.55

Table 5 shows the relative efficiency of the HT and REG estimators in comparison with the proposed estimator, JAC[12]. In all of the models, the proposed estimator performs better than the HT estimator, except the nonlinear model. The efficiency of the proposed estimator in comparison with REG estimator is almost bigger than 1. Notice that the use of auxiliary information is very relevant in the M3 model and does not affect in the M5 model. The efficiency of the proposed estimator is very close to one, in most cases, in comparison with the REG estimator.

Table 5: Relative efficiency of the proposed estimator.

Model	HT	REG
M1	9.19	1.00
M2	8.10	1.02
M3	33.2	1.00
M4	8.11	0.99
M5	1.00	1.00
M6	2.08	1.00

4.1. Small Sample Sizes

So far, the proposed estimator features very well in comparison with the HT estimator and performs at least as well as REG estimator. There is a particular

[12]The Relative Efficiency, **RE**, of an estimator $\widehat{t}_y$ is given by the ratio $\mathbf{RE}\left(\widehat{t}_y\right) = \frac{MSE\left(\widehat{t}_y\right)}{MSE\left(\widehat{t}_{JAC}\right)}$. Ratios bigger than one favor the proposed estimator.

case when the proposed estimator gains more than 40% in comparison with REG estimator: when the sample size is small and the percentage of the contaminated data is between 1 and 10%, the proposed estimator is clearly better than REG estimator, as is shown in Table 6.

TABLE 6: Relative efficiency of the proposed estimator in comparison with REG estimator.

% contaminated	n=100	n=50	n=20	n=10	n=5	n=3
0.1%	0.99	1.08	1.01	1.21	0.87	1.16
1%	1.00	1.02	1.07	1.16	1.20	1.04
5%	1.00	1.02	1.04	1.14	1.04	1.46
10%	0.99	1.03	1.06	1.12	1.16	1.17
20%	1.00	1.03	1.04	1.09	0.98	1.06
40%	1.00	1.03	1.05	1.08	1.07	0.99
50%	1.01	1.04	1.06	1.09	1.08	0.96

The simulation was done following the same specifications as in Section 2.3 The value of the parameter β_1 was set to two and the value of the parameter β_0 was set to ten such that $y_k > 0 \forall k \in U$. It is assumed that ε_k follows a mixture of two normal densities with means 0 and 10 and identical variance equal to 2[13].

Regarding the percentage of the contaminated data, the proposed estimator (JAC) does not perform well when there is too much contamination. When a small probability of a large contamination is used, *i.e.* with a few outliers, the rank based method performs better. The simulation was done using different sample sizes and different percent of contaminated data in the response variable.

Table 6 reports the relative efficiency of the proposed estimator in comparison with REG estimator and it can be noted that, when the sample size decreases, the good performance of the REG estimator decreases too, in comparison with the JAC estimator[14]. When the sample size is equal to 100, the proposed estimator performs as well as the REG estimator and the percent of contaminated data does not influence the performance of the estimators.

When the percentage of contaminated data is higher than 10%, the efficiency of the proposed estimator tends to decrease. Note that when the percentage of contaminated data is high, REG estimator has a very good behavior, even when the sample size is small. Specifically, it is recommended to use the method proposed in this research when the percentage of contaminated data is less than 20% because when the sample size decreases, the efficiency of the estimator increases substantially and it indicates that the JAC estimator performs better than traditional estimators in the survey sampling literature and still maintains a very small bias.

[13]The contamination of the response variable is done with the creation of an indicator that converts the error term to a mixture of normal densities with different means.

[14]The Relative Bias of both estimators, REG and JAC, is always less than 3% and is not reported.

5. Conclusions and Further Research

This research was motivated by the construction of an estimator able to gain in efficiency under some particular conditions. The estimator was built under a model-assisted approach using the minimum dispersion criterion and the generalized difference estimator as baseline. In order to construct a population total estimator that involves a regression model it was necessary to build the estimators of such regression coefficients. These estimator were motivated by some particular cases where the traditional least squares approach did not fit well (such as the contaminated response variable scenario). In this pursuit of the rank-based estimators for the slope and intercept, the minimum dispersion criterion was used and the behavior of such estimators was completely satisfactory, in the sense of high efficiency, compared with the least squares approach.

When the good performance of these regression estimators was observed, the next step was the construction of a population total estimator. The form of the generalized difference estimator was used and the construction of the variance estimator of the population total estimator was proposed. The results of several simulations done in this research show that the proposed estimator works very well under particular conditions consistent with the survey sampling context. The proposed estimator and its implementation in the R software is open and available in case needed.

Of course, there are many open questions in this research. There are many other choices for the set of scores and it will be interesting to show how the choice of the scores affects the estimation of the parameters. Note that the poststratification estimator was not considered in the simulation study. If the auxiliary information is not continuous but discrete, as in the poststratification estimator, robust poststratification through the minimum dispersion criterion be an interesting alternative.

[Recibido: julio de 2008 — Aceptado: marzo de 2009]

References

Breidt, F. J. & Opsomer, J. D. (2000), 'Local Polynomial Regression Estimators in Survey Sampling', *The Annals of Statistics* **28**, 1026–1053.

Cassel, C. M., Särndal, C. E. & Wretman, J. (1976*a*), *Foundations of Inference in Survey Sampling*, Wiley, New York, United States.

Cassel, C. M., Särndal, C. E. & Wretman, J. (1976*b*), 'Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations', *Biometrika* **63**, 615–620.

Chen, J. & Qin, J. (1993), 'Empirical Likelihood Estimation for Finite Populations and the Efectivene Usage of Auxiliary Information', *Biometrika* **80**, 107–116.

Deville, J. C. & Särndal, C. E. (1992), 'Calibration Estimators in Survey Sampling', *Journal of the American Statistical Association* **87**, 376–382.

Draper, D. (1988), 'Rank-Based Robust Analysis of Linear Models I. Exposition and Review', *Statistical Science* **3**, 239–257.

Hettmansperger, T. P. (1984), *Statistical Inference Based on Ranks*, Wiley, New York, United States.

Hettmansperger, T. P. & McKean, J. W. (1998), *Robust Nonparametric Statistical Methods*, Arnold.

Horvitz, D. G. & 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. & Fuller, W. A. (1982), 'Survey Design under the Regression Superpopulation Model', *Journal of the American Statistical Association* **767**, 89–96.

Jaeckel, L. (1972), 'Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals', *The Annals of Mathematical Statistics* **43**, 1449–1458.

Jurečková, J. (1971), 'Nonparametric Estimate of Regression Coefficients', *The Annals of Mathematical Statistics* **42**, 1328–1338.

Lohr, S. (1999), *Sampling: Design and Analysis*, Duxbury Press, California, United States.

Särndal, C. E. (1980), 'On π-inverse Weighting Versus best Linear Unbiased Weighting in Probability Sampling', *Biometrika* **67**, 639–650.

Särndal, C. E., Swensson, B. & Wretman, J. (1989), 'The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Popoulation Total', *Biometrika* **76**, 527–537.

Särndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York, United States.

Team, R. D. C. (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN.

Terpstra, J. F. & McKean, J. W. (2005), 'Rank-based analyses of linear models using R', *Journal of Statistical Software* **14**, 1–26.

Wu, C. (2003), 'Optimal Calibration Estimators in Survey Sampling', *Biometrika* **90**, 937–951.

Wu, C. & Sitter, R. R. (2001), 'A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data', *Journal of the American Statistical Association* **96**, 185–193.