DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 101–119, http://dx.doi.org/10.1515/JOS-2015-0005

# On Estimating Quantiles Using Auxiliary Information

*Yves G. Berger*[1] *and Juan F. Muñoz*[2]

We propose a transformation-based approach for estimating quantiles using auxiliary information. The proposed estimators can be easily implemented using a regression estimator. We show that the proposed estimators are consistent and asymptotically unbiased. The main advantage of the proposed estimators is their simplicity. Despite the fact the proposed estimators are not necessarily more efficient than their competitors, they offer a good compromise between accuracy and simplicity. They can be used under single and multistage sampling designs with unequal selection probabilities. A simulation study supports our finding and shows that the proposed estimators are robust and of an acceptable accuracy compared to alternative estimators, which can be more computationally intensive.

*Key words:* Distribution function; inclusion probabilities; regression estimator; sample survey.

## 1. Introduction

Estimation of quantiles is of considerable interest when measuring income distribution and poverty lines (e.g. Osier 2009; Verma and Betti 2011; Eurostat 2003; Berger and Skinner 2003). For instance, the median is regarded as a more appropriate measure of location than the mean when variables of interest, such as income, expenditure, and so on, have highly skewed distributions, because the median is less sensitive to outliers than the mean. For this reason, the median is also used by most household wealth surveys, such as the Household Finance and Consumption Survey (HFCS) carried out by the European Central Bank among the Eurozone countries. In addition, quantile estimation has many practical applications, for example, when measuring poverty (e.g. Osier 2009; Eurostat 2012; Eurostat 2003).

In sample surveys, auxiliary information is often used at the estimation stage to improve the estimation of target parameters. The use of auxiliary information has been studied extensively for estimation of means and totals. However, it has no obvious extensions to the estimation of quantiles. In this article, we propose a transformation-based approach for estimating quantiles, which takes into account of the auxiliary information.

We consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ containing $N$ units. Let $y_1, \ldots, y_N$ denote the values of a variable of interest, $y$, and $x_1, \ldots, x_N$ denote the values of an auxiliary variable, $x$. Our proposed approach can be easily extended to several auxiliary variables. A sample $s$ of size $n$ is selected randomly from $U$ according to

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK. Email: y.g.berger@soton.ac.uk
[2] Department of Quantitative Methods in Economics and Business, University of Granada, Granada, 18071, Spain. Email: jfmunoz@ugr.es

a sampling design. We consider a design-based approach where the $y_i$ and $x_i$ are fixed (nonrandom) quantities and the sampling distribution is specified by the sampling design. The aim is to estimate the population quantile

$$Y_\alpha = F^{-1}(\alpha), \tag{1}$$

where $F^{-1}(\cdot)$ is the inverse of the population distribution function

$$F(t) = \frac{1}{N}\sum_{i\in U}\delta(y_i \le t)$$

and $0 < \alpha < 1$. The function $\delta(\cdot)$ takes the value 1 if its argument is true and 0 otherwise. Throughout this article, we define the inverse of any function $G(\cdot)$ by $G^{-1}(\alpha) = inf\{t : G(t) \ge \alpha\}$.

A customary estimator for $Y_\alpha$ is obtained by substituting $F(t)$ by its estimator into (1). For example, the 'Hájek type' estimator of $Y_\alpha$ is defined by

$$\widehat{Y}_{\pi,\alpha} = \widehat{F}_\pi^{-1}(\alpha), \tag{2}$$

where $\widehat{F}_\pi(t)$ is the Hájek (1971) estimator defined by

$$\widehat{F}_\pi(t) = \frac{1}{\widehat{N}}\sum_{i\in s}\frac{1}{\pi_i}\delta(y_i \le t) \tag{3}$$

with $\widehat{N} = \sum_{i\in s}\pi_i^{-1}$, where $\pi_i$ denotes the first-order inclusion probability of unit $i$. A wide range of estimators exists for the distribution function $F(\cdot)$, some of which use auxiliary information (see Section 2).

The proposed approach consists in inverting the distribution function at $\widehat{\alpha}_{reg}$ rather than at $\alpha$. The quantity $\widehat{\alpha}_{reg}$, defined in (19), takes the auxiliary information into account. The proposed estimators can be justified by using a transformation of the variable of interest. The proposed estimators depend on the first-order inclusion probabilities. The proposed estimators can be calculated even if we only know the auxiliary variables for the sampled units, as long as the population quantile of the auxiliary variable is known.

In Section 2, we define estimators of the distribution function that can be found in the literature, and which can be used to estimate a quantile. In Section 3, we introduce the proposed estimators for a quantile. In Section 4, we give regularity conditions under which the proposed estimators are consistent. In Section 5, we compare the proposed estimators with alternative estimators via simulation. We also investigate the empirical properties of a bootstrap variance estimator. This article concludes with some discussions in Section 6.

## 2. Estimators of Quantiles

An exhaustive review of estimators of the distribution function and quantiles can be found in Dorfman (2009).

By substituting the design weights in (3) with calibration weights, we obtain the following naïve estimator

$$\widehat{F}_w(t) = \frac{1}{\widehat{N}_w} \sum_{i \in s} w_i \delta(y_i \leq t), \tag{4}$$

where $\widehat{N}_w = \sum_{i \in s} w_i$. The $w_i$ denote the regression weights calibrated with respect to the population total of the auxiliary variable. The estimator of $Y_\alpha$ based on these calibration weights is given by $\widehat{Y}_{w;\alpha} = \widehat{F}_w^{-1}(\alpha)$.

The model-based estimator of the distribution function suggested by Chambers and Dunstan (1986) is based on the following heteroscedastic regression model

$$y_i = \beta x_i + \nu(x_i) u_i, \tag{5}$$

where $\beta$ is an unknown parameter, $\nu(x_i)$ is a known function of $x$ and the $u_i$ are independent and identically distributed random variables with zero mean. The distribution function estimator proposed by Chambers and Dunstan (1986) is

$$\widehat{F}_{cd}(t) = \left[ \sum_{i \in s} \delta(y_i \leq t) + \frac{1}{n} \sum_{j \in U-s} \sum_{i \in s} \delta\left( u_{ni} \leq \frac{t - b_n x_j}{\nu(x_j)} \right) \right], \tag{6}$$

with

$$b_n = \left[ \sum_{i \in s} \frac{x_i^2}{\nu^2(x_i)} \right]^{-1} \sum_{i \in s} \frac{y_i x_i}{\nu^2(x_i)}; \quad u_{ni} = \frac{y_i - b_n x_i}{\nu(x_i)}.$$

The Chambers and Dunstan (1986) estimator of $Y_\alpha$ is given by $\widehat{Y}_{cd;\alpha} = \widehat{F}_{cd}^{-1}(\alpha)$.

Rao et al. (1990) proposed the following estimator

$$\widehat{F}^{\bullet}_{rkm}(t) = \frac{1}{N} \left\{ \sum_{i \in s} \pi_i^{-1} \delta(y_i \leq t) + \left( \sum_{i \in U} \widehat{G}_i(t) - \sum_{i \in s} \pi_i^{-1} \widehat{G}_{ic}(t) \right) \right\}$$

with

$$\widehat{G}_i(t) = \frac{1}{\widehat{N}} \sum_{j \in s} \frac{1}{\pi_j} \delta\left( \widehat{u}_j \leq \frac{t - \widehat{R} x_i}{x_i^{1/2}} \right),$$

$$\widehat{G}_{ic}(t) = \left( \sum_{j \in s} \frac{\pi_i}{\pi_{ij}} \right)^{-1} \left[ \sum_{j \in s} \frac{\pi_i}{\pi_{ij}} \delta\left( \widehat{u}_j \leq \frac{t - \widehat{R} x_i}{x_i^{1/2}} \right) \right],$$

$$\widehat{u}_j = \frac{y_j - \widehat{R} x_j}{x_j^{1/2}}, \widehat{R} = \left[ \sum_{i \in s} \frac{x_i}{\pi_i} \right]^{-1} \sum_{i \in s} \frac{y_i}{\pi_i},$$

where $\pi_{ij}$ denotes the joint inclusion probability for the units $i$ and $j$. Since the estimator $\widehat{F}^{\bullet}_{rkm}(t)$ is not always a monotone nondecreasing function, Rao et al. (1990) proposed to use the following estimator

$$\widehat{F}_{rkm}(t) = \max\left\{\widetilde{F}_{rkm}(y_{(i)}) : y_{(i)} \leq t\right\}, \tag{7}$$

where the $y_{(i)}$'s are the order statistics of the sample $\{y_i, i \in s\}$ and $\widetilde{F}_{rkm}(y_{(i)})$ is defined by the following recursive formula

$$\widetilde{F}_{rkm}(y_{(i)}) = \max\left\{\widetilde{F}_{rkm}(y_{(i-1)}), \widehat{F}^{\bullet}_{rkm}(y_{(i)})\right\},$$

with $\widetilde{F}_{rkm}(y_{(1)}) = \widehat{F}^{\bullet}_{rkm}(y_{(1)})$. The Rao et al. (1990) estimator of $Y_\alpha$ is given by $\widehat{Y}_{rkm;\alpha} = \widehat{F}^{-1}_{rkm}(\alpha)$.

Silva and Skinner (1995) proposed the following estimator based on poststratification

$$\widehat{F}_{ps}(t) = \frac{1}{N}\sum_{g=1}^{G}\frac{N_g}{\widehat{N}_g}\sum_{i\in s}\frac{1}{\pi_i}\delta(y_i \leq t)\delta(i \in U_g), \tag{8}$$

where $U_1, \ldots, U_G$ are $G$ poststrata partitioning the population, $N_g$ is the size of $U_g$ and $\widehat{N}_g = \sum_{i\in s_g}\pi_i^{-1}$, with $g = 1, \ldots, G$. The estimator of $Y_\alpha$ is given by $\widehat{Y}_{ps;\alpha} = \widehat{F}^{-1}_{ps}(\alpha)$.

When the population quantile $X_\alpha$ of an auxiliary variable is known, Rao et al. (1990) proposed the following ratio estimator of $Y_\alpha$

$$\widehat{Y}_{r;\alpha} = \frac{\widehat{Y}_{\pi;\alpha}}{\widehat{X}_{\pi;\alpha}}X_\alpha, \tag{9}$$

where $\widehat{Y}_{\pi;\alpha}$ and $\widehat{X}_{\pi;\alpha}$ are respectively the Hájek estimators of $Y_\alpha$ and $X_\alpha$ (see (2)). Rao et al. (1990) also proposed a difference estimator and showed that $\widehat{Y}_{r;\alpha}$ has a smaller mean square error than the difference estimator.

Harms and Duchesne (2006) proposed an estimator of the distribution function based on a calibration constraint specified by the quantile of an auxiliary variable. This estimator is denoted by $\widehat{Y}_{cal;\alpha}$.

Note that the estimators $\widehat{Y}_{cd;\alpha}$, $\widehat{Y}_{rkm;\alpha}$ and $\widehat{Y}_{ps;\alpha}$ assume that the auxiliary variable is known for all the units of the population, whereas estimators $\widehat{Y}_{r;\alpha}$ and $\widehat{Y}_{cal;\alpha}$ only require the knowledge of $X_\alpha$.

## 3. Proposed Estimators for a Quantile

The proposed estimators are based upon the following idea, which can be illustrated for a median: if the distribution of the variable of interest is such that the mean equals the median, the median could be estimated by using an estimator for the mean. We propose to transform the variable of interest in such a way that the median equals the mean for the transformed variable. If the transformation is monotone increasing, the median of the variable of interest can be estimated by inverting the estimate for the mean of the transformed variable. This method can also be extended to the estimation of any quantile. The proposed estimators are given by (18) and (20) in Subsection 3.3. In order to justify

this approach, it is necessary to transform the variable (Subsection 3.1) and to use a regression estimator (Subsection 3.2).

### *3.1. A Transformation of the Variables*

We propose to transform the variable of interest such that the distribution of the transformed variable is approximately symmetric. Consider the midpoint distribution function $F^{\circ}(\cdot)$ (Nygård and Sandström 1985) defined by

$$F^{\circ}(y) = \frac{1}{2}[F(y^{-}) + F(y)]. \tag{10}$$

The quantity $F(y^{-})$ is the left-hand limit, that is, $F(y^{-}) = \lim_{t \to y^{-}} F(t)$. Alternatively, $F^{\circ}(y) = N^{-1} \sum_{i \in U} [\delta(y_i < y) + 0.5\delta(y_i = y)]$. Note that $0 < F^{\circ}(y_i) < 1$ for all $i \in U$. If the population quantile $Y_\alpha$ is the parameter of interest, we consider the following transformed values

$$y^{*}_{\alpha;i} = \Psi(y_i) + z_\kappa, \tag{11}$$

where $\Psi(y_i) = \phi^{-1}(F^{\circ}(y_i))$ and $\phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function $\phi(\cdot)$ of a normal $N(0, 1)$; that is,

$$\phi(y) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{y} \exp\left(\frac{-t^2}{2}\right) dt.$$

The quantity $z_\kappa = \phi^{-1}(\kappa)$ is the $\kappa$-th quantile of a normal $N(0, 1)$ distribution, with $\kappa = (\lceil \alpha N \rceil - 0.5)/N$. Note that $\kappa$ can be approximated by $\alpha$ for large populations, as $\kappa \to \alpha$ when $N \to \infty$. The quantity $\alpha$ is the level of the quantile $Y_\alpha$ considered.

In the definition of $\Psi(y_i)$, we use (10) instead of $F(t)$ because the function $\phi^{-1}(\cdot)$ is not defined on 0 and 1. Note that the transformation $\Psi(y_i)$ does not depend on the choice of $\alpha$. This function maps the quantiles of the distribution of $y$ with the quantile of the standardised normal distribution $N(0, 1)$. Note that $\Psi(y_i)$ can be estimated with or without auxiliary variables.

The following Lemma gives the relationship between the population quantile $Y_\alpha$ and the following population mean of the transformed variable

$$\overline{Y}^{*}_{\alpha} = \frac{1}{N} \sum_{i \in U} y^{*}_{\alpha;i}.$$

**Lemma 1** *We have that* $Y_\alpha = \Psi^{-1}(\overline{Y}^{*}_{\alpha})$, *where the function* $\Psi^{-1}(\cdot)$ *is the inverse of function* $\Psi(\cdot)$ *defined in (11)*

The proof is given in Appendix A.

The transformed values in (11) depend on population values, which would need to be estimated. We propose to estimate $y^{*}_{\alpha;i}$ by its substitution estimator given by

$$\widehat{y}^{*}_{\alpha;i} = \widehat{\Psi}(y_i) + z_\kappa,$$

where $\widehat{\Psi}(y_i) = \phi^{-1}(\widehat{F}^{\circ}(y_i))$. The function $\widehat{F}^{\circ}(\cdot)$ is the empirical midpoint estimator of the distribution function (10). This estimator is given by

$$\widehat{F}^{\circ}(y) = \frac{1}{2}[\widehat{F}(y^{-}) + \widehat{F}(y)], \tag{12}$$

where $\widehat{F}(\cdot)$ is a consistent estimator of $F(\cdot)$. In this article, we propose to use the Hájek-type estimator (3) in (12). However, we could use (6), (7) or (8) instead of (3). This may give a more efficient estimator.

The auxiliary variable may be transformed in the same way. When the values $x_i$ are known for the entire population, we propose to use the following transformation.

$$x^{*}_{\alpha;i} = \Psi_x(x_i) + z_\kappa, \tag{13}$$

where $\Psi_x(x_i) = \phi^{-1}(F^{\circ}_x(x_i))$, $F^{\circ}_x(x) = [F_x(x^{-}) + F_x(x)]/2$ and $F_x(t) = N^{-1}\sum_{i\in U}\delta(x_i \le t)$. Note that the values of $x^{*}_{\alpha;i}$ cannot be calculated if we only know the sampled values of the auxiliary variable, as the function $F_x(\cdot)$ is unknown in this situation. If this is the case, we propose the transformation

$$\widehat{x}^{*}_{\alpha;i} = \widehat{\Psi}_x(x_i) + z_\kappa, \tag{14}$$

where $\widehat{\Psi}_x(x_i) = \phi^{-1}(\widehat{F}^{\circ}_x(x_i))$ and $\widehat{F}^{\circ}_x(x) = [\widehat{F}_x(x^{-}) + \widehat{F}_x(x)]/2$. The function $\widehat{F}_x(\cdot)$ may be any estimator of the distribution function $F_x(t)$. In this article, we propose to use the Hájek (1971) estimator of $F_x(\cdot)$ (see (3)).

### 3.2. *The Regression Estimator*

We propose to estimate $\overline{Y}^{*}_{\alpha}$ using a regression estimator (e.g. Cassel et al. 1976, 1977), which uses the auxiliary information. This estimator is defined by

$$\bar{y}^{*}_{reg;\alpha} = \bar{y}^{*}_{\alpha} + \widehat{\beta}_x\left(\overline{X}^{*}_{\alpha} - \bar{x}^{*}_{\alpha}\right), \tag{15}$$

where $\bar{y}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}\widehat{y}^{*}_{\alpha;i}$, $\overline{X}^{*}_{\alpha} = N^{-1}\sum_{i\in U}x^{*}_{\alpha;i}$, $\bar{x}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}x^{*}_{\alpha;i}$, with

$$\widehat{\beta}_x = \left[\sum_{i\in s}\frac{1}{\pi_i q_i^2}\left(x^{*}_{\alpha;i} - \bar{x}^{*}_{\alpha}\right)^2\right]^{-1}\sum_{i\in s}\frac{1}{\pi_i q_i^2}\left(x^{*}_{\alpha;i} - \bar{x}^{*}_{\alpha}\right)\left(\widehat{y}^{*}_{\alpha;i} - \bar{y}^{*}_{\alpha}\right). \tag{16}$$

Note that the regression estimator $\bar{y}^{*}_{reg;\alpha}$ assumes that the auxiliary variable is known for the entire population. When we only know the values of the auxiliary variable for the sampled units, we propose to use the following regression estimator instead of (15):

$$\bar{y}^{*}_{regS;\alpha} = \bar{y}^{*}_{\alpha} + \tilde{\beta}_x\left(\widehat{\overline{X}}^{*}_{\alpha} - \widehat{\bar{x}}^{*}_{\alpha}\right), \tag{17}$$

where $\widehat{\bar{x}}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}\widehat{x}^{*}_{\alpha;i}$ and $\widehat{\beta}_x$ is given by (16) after substituting $x^{*}_{\alpha;i}$ by $\widehat{x}^{*}_{\alpha;i}$. The control mean in (17) can be obtained as

$$\widehat{\overline{X}}^{*}_{\alpha} = \widehat{\Psi}_x(X_\alpha)\,.$$

This implicitly assumes that we know $X_\alpha$. The Estimator (9) and the estimator proposed by Harms and Duchesne (2006) are also based on this assumption.

We can observe that the estimators $\bar{y}^*_{reg;\alpha}$ and $\bar{y}^*_{regS;\alpha}$ are based upon a single auxiliary variable. The proposed regression estimators can be easily extended to several auxiliary variables (e.g. Särndal et al. 1992, 225). For this purpose, the various auxiliary variables may be transformed by using the transformations (13) or (14) suggested for the variable $x$.

### *3.3. The Proposed Estimators*

Based on Lemma 1, we propose to estimate the quantile $Y_\alpha$ by

$$\widehat{Y}_{reg;\alpha} = \widehat{\Psi}^{-1}\left(\bar{y}^*_{reg;\alpha}\right). \tag{18}$$

As $\widehat{\Psi}^{-1}(y) = \widehat{F}^{\circ -1}(\phi(y))$, an alternative expression for the proposed estimator is

$$\widehat{Y}_{reg;\alpha} = \widehat{F}^{\circ -1}(\widehat{\alpha}_{reg}), \tag{19}$$

where $\widehat{\alpha}_{reg} = \phi\left(\bar{y}^*_{reg;\alpha}\right)$. This estimator consists in inverting a midpoint distribution function $\widehat{F}^{\circ}(\cdot)$ at the value $\widehat{\alpha}_{reg}$, which is adjusted to take into account the auxiliary variable. Note that if we invert the midpoint distribution function (12) at the value $\alpha$ and if we use the estimator (3), we obtain an estimator which is approximately equal to the Hájek-type estimator (2) when $\widehat{F}^{\circ}(\cdot)$ is given by (3).

When we only know the values of the auxiliary variable for the sampled units and when the population quantile $X_\alpha$ is known, we propose to use a different estimator given by

$$\widehat{Y}_{regS;\alpha} = \widehat{\Psi}^{-1}\left(\bar{y}^*_{regS;\alpha}\right) = \widehat{F}^{\circ -1}(\widehat{\alpha}_{regS}), \tag{20}$$

where $\widehat{\alpha}_{regS} = \phi\left(\bar{y}^*_{regS;\alpha}\right)$ and $\bar{y}^*_{regS;\alpha}$ is defined by (17).

The proposed estimators are not affected by outliers, because $\widehat{y}^*_{\alpha;i}$ and $x^*_{\alpha;i}$ are implicitly based upon the ranks of $y$ and $x$ (see (11)). Note that $\widehat{Y}_{reg;\alpha} = X_\alpha$ when $y_i = x_i$. The efficiency of the proposed estimators depends on the correlation between $y^*_i$ and $x^*_i$ rather than the correlation between $y_i$ and $x_i$.

It is worth investigating some properties of the Estimator (19) under equal probability sampling ($\pi_i = n/N$). In this case, it can be shown that

$$\bar{y}^*_{reg;\alpha} \doteq z_k - \widehat{\beta}_x \frac{1}{n}\sum_{i\in s}\Psi_x(x_i).$$

Thus, $\bar{y}^*_{reg;\alpha}$ increases monotonically when $\alpha$ increases, because $z_\kappa$ is a monotone function of $\alpha$, and $\widehat{\beta}_x$ and $\Psi_x(x_i)$ do not depend on $\alpha$. Hence, $\widehat{Y}_{reg;\alpha_1} \leq \widehat{Y}_{reg;\alpha_2}$ when $\alpha_1 \leq \alpha_2$. This is a desirable property of an estimator of a quantile. Provided that $\widehat{\beta}_x > 0$, we have that $\widehat{\alpha}_{reg} > \alpha$ when $\sum_{i\in s}\Psi_x(x_i)$ is negative; that is, when the sample contains small $x_i$ values. In this case, the estimate based on $\alpha$ (e.g. (2) with (3)) is likely to have a negative error. By using a level $\widehat{\alpha}_{reg}$ larger that $\alpha$, we should reduce this error. Furthermore, as the adjustment, $\widehat{\beta}_x n^{-1}\sum_{i\in s}\Psi_x(x_i)$, does not depend on $\alpha$, the proposed estimators are likely to be good for some $\alpha$, but not for any $\alpha$. The simulation study in Section 5 investigates this features.

The rescaled bootstrap variance estimator (Rao et al. 1992) can be used to estimate the variance of the proposed estimators. A confidence interval for the point estimator can be

also computed using the rescaled bootstrap confidence interval (the histogram approach). In Subsection 5.1, we evaluate the empirical performance of this variance estimator and this confidence interval.

## 4. Design Consistency

Consider the following regularity conditions:

$$|\widehat{Y}_\alpha - Y_\alpha| = O_p(n^{-1/2}), \tag{21}$$

$$\left|\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right| = O_p(n^{-1/2})\cdot \tag{22}$$

Conditions (21) and (22) mean that $\widehat{Y}_\alpha$ and $\bar{y}^*_{reg;\alpha}$ are $\sqrt{n}$-consistent. Isaki and Fuller (1982) and Robinson and Särndal (1983) gave conditions under which (22) holds. Francisco and Fuller (1991) established the consistency of $\widehat{Y}_\alpha$. Furthermore, the fact that the $y^*_{\alpha;i}$ can be considered as values generated from a normal distribution speaks in favour of (22).

As $\widehat{F}^{\circ -1}(\cdot)$ is a nondifferentiable function, we need to assume that this function converges to a differentiable function in order to prove the consistency. We assume that there exists a quantile function $Q(\cdot)$ which is twice differentiable, and such that

$$\sup_{|\epsilon|<o(n^{-1/2})} |\widehat{F}^{\circ -1}(\alpha+\epsilon) - \widehat{F}^{\circ -1}(\alpha) - Q(\alpha+\epsilon) + Q(\alpha)| = o_p(1). \tag{23}$$

This condition can be justified by Bahadur (1966) Lemma (see also Serfling 1980, Lemma E, p. 97).

**Theorem 1** *Under assumptions (21), (22) and (23), the proposed estimator $\widehat{Y}_{reg;\alpha}$ is $\sqrt{n}$-consistent, as $|\widehat{Y}_{reg;\alpha} - Y_\alpha| = O_p(n^{-1/2})$.*

The proof of Theorem 1 is given in the Appendix B. In addition, $\widehat{Y}_{reg;\alpha}$ is asymptotically unbiased when $|\widehat{Y}_{reg;\alpha} - Y_\alpha|$ is uniformly bounded, as in this situation, the convergence in probability of $\widehat{Y}_{reg;\alpha}$ to $Y_\alpha$ implies that the expectation of $\widehat{Y}_{reg;\alpha}$ converges to $Y_\alpha$ (Lehmann 1999, 53).

It can be shown that the second estimator (20) is also consistent by assuming that (22) holds for $\bar{y}^*_{regS;\alpha}$.

## 5. Simulation

In this section, the proposed estimators $\widehat{Y}_{reg;\alpha}$ and $\widehat{Y}_{regS;\alpha}$ (see (18) and (20)) are compared numerically with alternative estimators described in Section 2. The alternative estimators considered are: $\widehat{Y}_{\pi;\alpha}$ (see (2)), $\widehat{Y}_{w;\alpha}$ (see (4)), $\widehat{Y}_{cd;\alpha}$ (Chambers and Dunstan 1986), $\widehat{Y}_{rkm;\alpha}$ (Rao et al. 1990), $\widehat{Y}_{ps;\alpha}$ (Silva and Skinner 1995), $\widehat{Y}_{r;\alpha}$ (see (9)) and $\widehat{Y}_{cal;\alpha}$ (Harms and Duchesne 2006).

The proposed Estimators (19) and (20) are based on the midpoint distribution function (12), which could be based on any estimator of $F(\cdot)$. For example, we can use the Estimators (3), (6), (7) or (8). The Estimators (6), (7) and (8) use auxiliary information and are therefore expected to be more accurate than (3). In our simulation study, we considered the worst-case scenario when the proposed estimators are based upon the Hájek-'type'

*Table 1. Descriptive statistics of the variables of interest of the populations considered: $\rho$ is the population correlation coefficient between y and x, $\rho^*$ is the population correlation coefficient between $y^*$ and $x^*$, and $\gamma_y$ and $\gamma_x$ are respectively the population skewness coefficients of y and x.*

| Pop. | $Y_{0.05}$ | $Y_{0.25}$ | $Y_{0.5}$ | $Y_{0.75}$ | $Y_{0.95}$ | $\rho$ | $\rho^*$ | $\gamma_y$ | $\gamma_x$ |
|---|---|---|---|---|---|---|---|---|---|
| Sugar | 34886 | 57585 | 80009 | 117159 | 204745 | 0.89 | 0.84 | 2.4 | 2.3 |
| MUN-1 | 6 | 10 | 16 | 31 | 84 | 0.61 | 0.70 | 8.2 | 1.2 |
| MUN-2 | 6 | 10 | 16 | 31 | 84 | 0.69 | 0.87 | 8.2 | 1.4 |
| ES-SILC | 13368 | 17970 | 22000 | 27700 | 42524 | 0.69 | 0.62 | 1.8 | 3.1 |
| HMT | 0.55 | 1.25 | 2.23 | 3.86 | 7.53 | 0.76 | 0.78 | 2.0 | 1.4 |

distribution function $\widehat{F}_\pi(t)$ defined by (3). In terms of simplicity, the proposed estimators should be obviously based upon (3).

The simulation study is based on several populations which are briefly described as follows. The sugar population consists of $N = 338$ sugar cane farms where $y$ denotes the gross value of canes and $x$ is the total cane harvested. The sugar population was used by Chambers and Dunstan (1986), Rao et al. (1990) and Silva and Skinner (1995). The population of municipalities (Särndal et al. 1992, 652) consists of $N = 284$ municipalities, where the variable of interest is the population size of the municipalities in 1985. We considered two auxiliary variables: (i) the number of conservative seats in municipal council (population MUN-1); and (ii) the total number of seats in municipal council (population MUN-2). We considered the Hansen et al. (1983) population (population HMT), which is $N = 14{,}000$ units generated from a bivariate gamma population (see also Rao et al. 1990). Finally, the last population is based on a random subset of $N = 2{,}000$ individuals from the 2012 Spanish Statistics on Income and Living Conditions (ES-SILC) Survey (Eurostat 2012). The ES-SILC provides information on income, poverty, social inclusion and living conditions for a sample of households and individuals. We considered the equivalised net income as the variable of interest and the tax on income contributions as the auxiliary variable. A brief descriptive analysis of the various populations is given in Table 1.

For each simulation, 1,000 samples were selected to compute the empirical relative bias $RB = (E[\widehat{Y}_\alpha] - Y_\alpha)/Y_\alpha$ and the empirical relative root mean square error $RRMSE = MSE[\widehat{Y}_\alpha]^{1/2}/Y_\alpha$ of an estimator $\widehat{Y}_\alpha$, where $E[\cdot]$ and $MSE[\cdot]$ denote respectively the empirical expectation and mean squared error. Simple random sampling and stratified random sampling were used to select the samples. The population quantiles $Y_{0.05}$, $Y_{0.25}$, $Y_{0.5}$, $Y_{0.75}$, and $Y_{0.95}$ are the parameters of interest.

Table 2 reports the empirical relative bias (RB) under simple random sampling. The RBs of the proposed estimators are of a reasonable range compared with the RBs of the alternative estimators, which can be larger than 10 percent in some cases. With the MUN-1 and MUN-2 populations, some estimators of $Y_{0.25}$ can have a large positive RB. Note that the proposed estimators tend to have large RB when the skewness of $y$ is large and $\alpha$ is small or large. With $\alpha = 0.05$ or 0.95, the proposed estimators and the alternative estimators can have large positive RB, especially when $\alpha = 0.95$. For example, this is the case of the estimator $\widehat{Y}_{cal;\alpha}$ for the Sugar, MUN-1 and MUN-2 populations and when $\alpha = 0.95$. The simulation results indicate that the estimator $\widehat{Y}_{cal;\alpha}$ can be severely biased. The estimators $\widehat{Y}_{w;\alpha}$ and $\widehat{Y}_{\pi;\alpha}$ have similar RBs. Studies from the existing literature

*Table 2. RB (%) of estimators of $Y_\alpha$ under simple random sampling.*

| Population | $\alpha$ | $\widehat{Y}_{\pi;\alpha}$ | $\widehat{Y}_{w;\alpha}$ | $\widehat{Y}_{cd;\alpha}$ | $\widehat{Y}_{ps;\alpha}$ | $\widehat{Y}_{rkm;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{r;\alpha}$ | $\widehat{Y}_{cal;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Sugar | 0.05 | 0.1 | 0.9 | −1.8 | −1.3 | −0.6 | 6.7 | −5.1 | 2.8 | 4.4 |
| ($n = 30$) | 0.25 | 0.1 | −5.1 | −0.5 | 0.3 | −0.6 | 2.0 | −1.2 | 1.7 | 1.8 |
| | 0.50 | −2.1 | −2.1 | 6.9 | 0.8 | 0.1 | 2.5 | 0.1 | 1.7 | 2.1 |
| | 0.75 | −0.3 | −6.0 | 10.7 | 0.3 | 0.1 | 2.5 | −0.8 | 2.2 | 3.1 |
| | 0.95 | 3.2 | 1.5 | 3.0 | 4.6 | −1.3 | 17.8 | 2.4 | 8.7 | −2.7 |
| Sugar | 0.05 | −5.3 | −1.5 | −8.1 | −1.1 | −1.6 | 2.5 | −6.7 | −0.2 | 0.6 |
| ($n = 60$) | 0.25 | −1.9 | −2.0 | −2.3 | 0.5 | −0.4 | 0.9 | −0.6 | 0.9 | 0.8 |
| | 0.50 | −0.9 | −1.0 | 4.3 | 1.0 | 0.1 | 1.3 | 0.8 | 1.1 | 1.4 |
| | 0.75 | −1.8 | −2.2 | 5.7 | 0.4 | −0.4 | 0.8 | 1.3 | 0.8 | 1.0 |
| | 0.95 | −1.9 | 1.7 | 10.8 | 2.6 | 2.6 | 6.8 | 2.7 | 9.0 | 7.1 |
| MUN-1 | 0.05 | −3.2 | −3.2 | −29.9 | 4.4 | −3.7 | 6.6 | 0.3 | 3.4 | 8.8 |
| ($n = 50$) | 0.25 | 5.1 | −5.4 | 8.6 | 8.9 | 4.7 | 9.7 | 12.4 | 11.9 | 13.5 |
| | 0.50 | −2.2 | −6.5 | 31.3 | 2.5 | −1.0 | 3.7 | −4.6 | 1.5 | −0.3 |
| | 0.75 | 0.3 | −6.7 | 22.4 | 1.3 | −0.4 | 3.4 | −2.9 | 0.2 | −0.3 |
| | 0.95 | 4.2 | 4.2 | 5.2 | 4.9 | 4.8 | 19.6 | 10.4 | 23.4 | 24.3 |
| MUN-2 | 0.05 | −2.9 | −2.9 | 17.2 | 15.1 | −3.3 | 7.3 | −8.8 | −2.8 | −6.0 |
| ($n = 50$) | 0.25 | 5.3 | −5.4 | 21.9 | 17.3 | 4.8 | 9.8 | 5.4 | 23.5 | 17.1 |
| | 0.50 | −1.9 | −6.3 | 18.0 | 13.7 | −0.4 | 4.1 | −4.2 | −0.2 | −2.9 |
| | 0.75 | 0.1 | −6.7 | 2.6 | 18.3 | −1.2 | 1.9 | 0.5 | 23.1 | 10.5 |
| | 0.95 | 4.5 | 4.5 | −6.7 | 28.9 | 4.5 | 19.8 | 5.5 | 20.5 | 21.4 |
| HMT | 0.05 | −1.0 | 0.7 | −53.9 | 0.4 | −0.3 | 2.3 | 1.9 | 0.6 | 0.9 |
| ($n = 200$) | 0.25 | −0.3 | 0.3 | −4.2 | 0.2 | 0.3 | 1.0 | 0.7 | −0.3 | 0.9 |
| | 0.50 | −0.1 | 0.4 | 10.4 | 0.4 | 0.4 | 0.9 | 0.4 | −0.1 | 0.9 |
| | 0.75 | −0.7 | −0.1 | 11.2 | 0.0 | 0.1 | 0.4 | 0.0 | −0.5 | 0.7 |
| | 0.95 | −1.3 | −1.2 | 10.0 | 0.4 | 0.6 | 2.1 | 0.0 | −1.2 | 2.8 |
| ES−SILC | 0.05 | −1.3 | 0.5 | −8.9 | 0.3 | 0.4 | 2.2 | −0.8 | 0.5 | 1.7 |
| ($n = 100$) | 0.25 | −0.3 | −0.3 | −3.5 | 0.1 | 0.1 | 0.4 | −0.2 | 0.2 | 0.4 |
| | 0.50 | −0.3 | −0.3 | 0.8 | 0.1 | 0.0 | 0.4 | −0.1 | 0.2 | 0.3 |
| | 0.75 | −0.5 | −0.5 | 5.3 | −0.1 | −0.1 | 0.4 | 0.4 | 0.3 | 0.4 |
| | 0.95 | 2.6 | −0.3 | 11.1 | −0.3 | 0.1 | 2.4 | 0.8 | −3.0 | 2.3 |

(Dorfman 2009) indicate that the Chambers and Dunstan estimator, $\widehat{Y}_{cd;\alpha}$, can have a large bias. This estimator is based on a superpopulation model. Dorfman (2009) indicates that when the superpopulation model holds, this estimator tends to be very accurate. When the super population model does not hold, the estimator has an inevitable bias. This is the reason why we observe a large RBs for this estimator in Table 2. The large RB corresponds to situations when the superpopulation model does not hold.

The efficiency of the estimators is measured by the empirical relative root mean square errors (RRMSE) which are reported in Table 3. We observe that the proposed estimators perform well in all situations expect when $\alpha = 0.95$. However, we observe that the alternative estimators also have large RRMSE in this situation. Note that the proposed estimators are based upon the Hájek distribution function (3). We notice a clear

*Table 3. RRMSE (%) of estimators of $Y_\alpha$ under simple random sampling.*

| Population | $\alpha$ | $\widehat{Y}_{\pi;\alpha}$ | $\widehat{Y}_{w;\alpha}$ | $\widehat{Y}_{cd;\alpha}$ | $\widehat{Y}_{ps;\alpha}$ | $\widehat{Y}_{rkm;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{r;\alpha}$ | $\widehat{Y}_{cal;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Sugar | 0.05 | 17.7 | 17.7 | 15.0 | 18.2 | 16.4 | 18.1 | 16.6 | 17.8 | 18.6 |
| ($n = 30$) | 0.25 | 11.6 | 12.5 | 6.6 | 9.7 | 9.2 | 9.3 | 10.7 | 9.4 | 9.4 |
| | 0.50 | 12.0 | 11.2 | 9.8 | 9.6 | 9.3 | 9.4 | 10.6 | 10.4 | 10.5 |
| | 0.75 | 14.3 | 13.1 | 15.3 | 10.9 | 10.3 | 11.4 | 11.2 | 12.6 | 12.5 |
| | 0.95 | 26.0 | 22.1 | 54.9 | 27.8 | 31.7 | 42.6 | 17.8 | 35.3 | 51.3 |
| Sugar | 0.05 | 13.8 | 13.1 | 12.9 | 12.5 | 12.2 | 12.1 | 14.0 | 12.6 | 12.9 |
| ($n = 60$) | 0.25 | 8.2 | 8.0 | 4.6 | 6.2 | 6.2 | 6.3 | 7.7 | 6.2 | 6.4 |
| | 0.50 | 8.3 | 7.6 | 6.0 | 6.5 | 6.1 | 6.2 | 7.3 | 7.0 | 7.1 |
| | 0.75 | 8.9 | 7.7 | 7.3 | 6.4 | 5.9 | 6.6 | 7.0 | 6.9 | 6.8 |
| | 0.95 | 12.4 | 12.0 | 29.0 | 14.3 | 13.7 | 18.2 | 12.9 | 27.1 | 28.1 |
| MUN-1 | 0.05 | 19.5 | 19.5 | 33.3 | 17.8 | 19.1 | 18.3 | 25.9 | 18.2 | 18.7 |
| ($n = 50$) | 0.25 | 12.2 | 12.1 | 13.3 | 14.0 | 12.0 | 14.0 | 18.7 | 15.8 | 18.4 |
| | 0.50 | 14.8 | 15.5 | 34.7 | 15.2 | 13.3 | 13.3 | 14.1 | 14.4 | 13.0 |
| | 0.75 | 17.1 | 15.3 | 29.5 | 14.9 | 12.4 | 17.8 | 14.0 | 14.4 | 13.5 |
| | 0.95 | 29.6 | 29.4 | 55.7 | 33.8 | 38.7 | 52.7 | 29.2 | 92.4 | 92.2 |
| MUN-2 | 0.05 | 18.6 | 18.6 | 21.4 | 22.5 | 18.2 | 17.3 | 18.2 | 19.4 | 16.8 |
| ($n = 50$) | 0.25 | 12.7 | 12.7 | 23.8 | 23.0 | 11.1 | 13.8 | 12.9 | 25.7 | 19.1 |
| | 0.50 | 14.4 | 14.9 | 22.5 | 26.1 | 12.3 | 11.9 | 12.4 | 12.6 | 11.0 |
| | 0.75 | 16.7 | 16.3 | 15.4 | 26.0 | 13.2 | 12.1 | 13.1 | 32.6 | 15.3 |
| | 0.95 | 28.0 | 28.0 | 26.7 | 77.9 | 28.0 | 58.4 | 23.7 | 76.9 | 83.7 |
| HMT | 0.05 | 11.7 | 11.5 | 55.1 | 11.3 | 12.1 | 11.4 | 19.6 | 11.8 | 12.7 |
| ($n = 200$) | 0.25 | 8.0 | 7.6 | 6.0 | 6.5 | 6.2 | 6.5 | 7.7 | 8.0 | 6.9 |
| | 0.50 | 7.5 | 6.8 | 11.2 | 5.9 | 5.8 | 5.9 | 6.4 | 7.5 | 6.3 |
| | 0.75 | 7.2 | 6.3 | 11.9 | 5.7 | 5.5 | 5.7 | 6.4 | 7.2 | 6.1 |
| | 0.95 | 9.9 | 9.1 | 11.9 | 9.6 | 8.9 | 9.9 | 9.3 | 9.9 | 11.0 |
| ES-SILC | 0.05 | 8.3 | 8.1 | 11.4 | 7.8 | 7.8 | 7.9 | 8.4 | 8.1 | 9.1 |
| ($n = 100$) | 0.25 | 3.7 | 3.6 | 4.4 | 3.4 | 3.3 | 3.4 | 3.9 | 3.6 | 3.6 |
| | 0.50 | 4.0 | 3.8 | 2.7 | 3.3 | 3.3 | 3.3 | 3.8 | 3.6 | 3.6 |
| | 0.75 | 4.7 | 4.2 | 6.1 | 4.0 | 3.6 | 3.8 | 4.7 | 4.1 | 4.1 |
| | 0.95 | 10.6 | 10.1 | 18.7 | 10.4 | 10.2 | 11.3 | 11.0 | 10.5 | 12.8 |

improvement between the proposed estimators and the Hájek estimator (2), because the RRMSEs of the proposed estimators are usually smaller than the RRMSEs of the Hájek estimator $\widehat{Y}_{\pi;\alpha}$. In other words, there is a clear improvement when using $\widehat{\alpha}_{reg}$ instead of $\alpha$, except when $\alpha = 0.95$ and 0.25 with the MUN-1 and MUN-2 populations. The proposed estimators can be more efficient than the alternative estimators, especially when $\alpha = 0.50$ and 0.75. We also observe that $\widehat{Y}_{reg;\alpha}$ is generally more efficient than $\widehat{Y}_{regS;\alpha}$.

We also conducted another series of simulations using stratified simple random sampling. The conclusions derived from this simulation study are similar. The results of this simulation study are not presented in this article.

We now investigate the conditional relative biases of the proposed estimator $\widehat{Y}_{reg;\alpha}$ given the sample means of the auxiliary variable. For this purpose, the 1,000 selected samples were ordered according to the mean of the auxiliary variable. Then this ranking was used to create 20 groups of 50 observations each. Conditional relative biases were then obtained by calculating the *RB* for each of the 20 groups.

Figure 1 displays the conditional relative biases of the estimators of the first quartile under simple random sampling from the Sugar population. We observe that the Hájek-type estimator clearly exhibits the worst conditional performance with a linear trend as the group mean of $x$ increases. The conditional RB of the proposed estimator and the Rao et al. (1990) estimator does not seem to be correlated with the group mean of $x$. The Rao et al. (1990) estimator has a bias which is slightly smaller than the bias of the proposed estimator. Figure 2 displays the conditional relative biases of the estimators of the median under simple random sampling from the MUN-1 population. The conditional relative bias of the proposed estimator and the Rao et al. (1990) estimator does not seem to be correlated with the group mean of $x$.

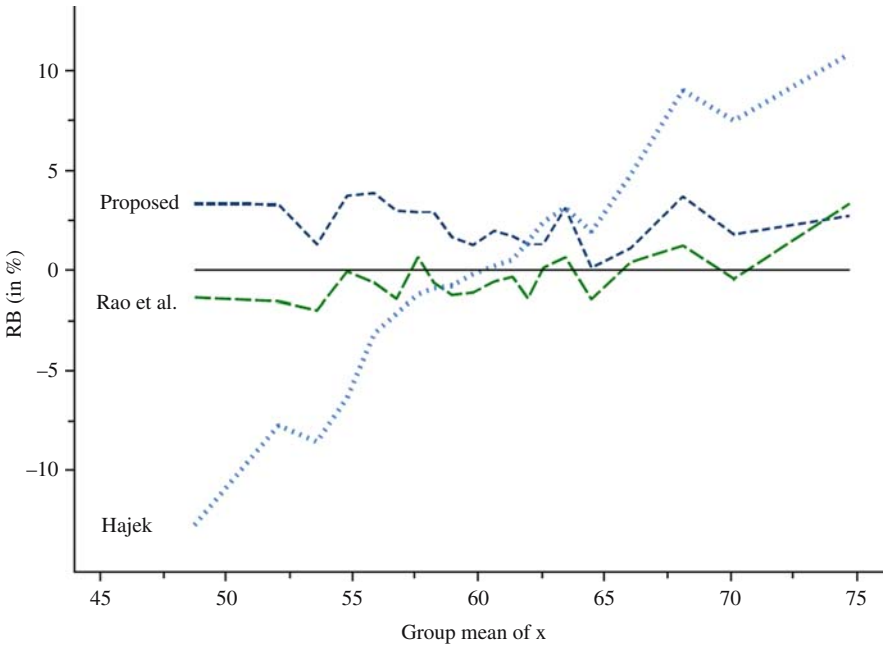


*Fig. 1. Conditional relative biases (%) of estimates of $Y_{0.25}$ under simple random sampling from the sugar population when $n = 30$.*

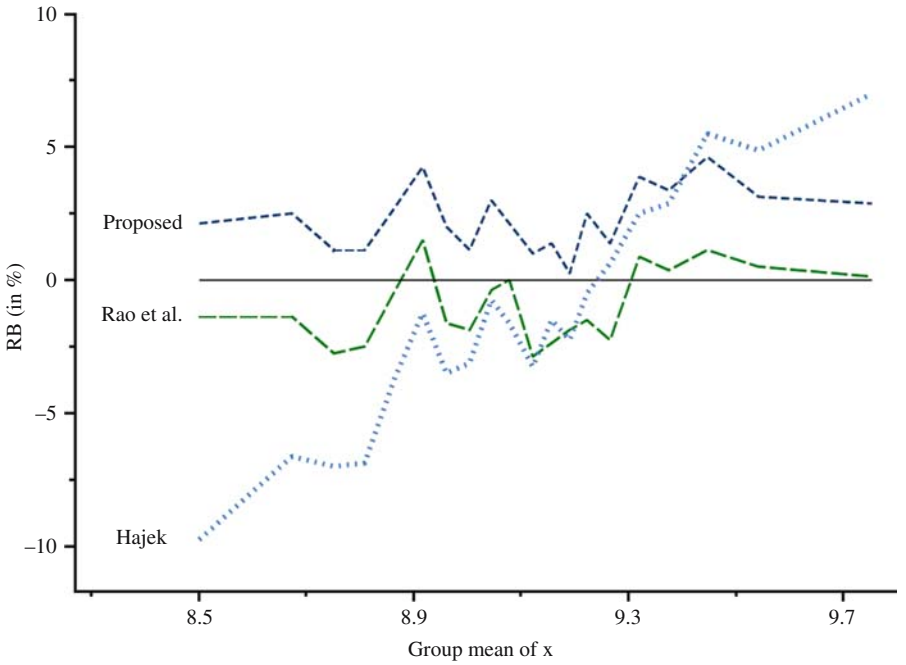


*Fig. 2. Conditional relative biases (%) of estimates of $Y_{0.5}$ under simple random sampling from the MUN-1 population when $n = 200$.*

The proposed estimator is biased and $\widehat{Y}_{rkm;\alpha}$ is approximately unbiased. This explains why $\widehat{Y}_{rkm;\alpha}$ shows under- and overestimation in Figures 1 and 2, otherwise $\widehat{Y}_{rkm;\alpha}$ would not be approximately unbiased. We observe an overestimation for all groups of mean for the proposed estimator, because this estimator has a small non-negligible bias.

The proposed transformation-based approach seems to perform well for estimating the central quantiles. In particular, results derived from simulation studies indicate that the proposed estimators have a good performance for the median. In this situation, the proposed estimators clearly outperform the Hájek estimator, especially when the conditional bias is taken into consideration. In addition, the proposed estimators perform well if they are compared to the various existing methods. For instance, although the proposed estimators can be slightly biased, they seem more efficient than the simpler alternatives $\widehat{Y}_{r;\alpha}$ (the ratio estimator) and $\widehat{Y}_{cal;\alpha}$ (Harms and Duchesne 2006). The values of RRMSE of the proposed estimators are comparable to the values of RRMSE of the more sophisticated estimator $\widehat{Y}_{rkm;\alpha}$ (Rao et al. 1990). These conclusions hold also in the situation where only population quantiles of the auxiliary variable are known. However, the proposed estimators can have large biases for the tail quantiles, specially when $\alpha = 0.95$. In this situation, the Hájek estimator appears more robust compared to all the more complex approaches.

### *5.1. Variance Estimation and Confidence Intervals*

We propose to estimate the variance of the proposed point estimators using the rescaled bootstrap variance estimator (Rao et al. 1992). Rao and Wu (1988) showed that the rescaled bootstrap variance estimator is a consistent estimator for the variance when the

*Table 4. Empirical relative bias (%) of the rescaled bootstrap variance estimators under simple random sampling when $n = 200$. The column $\rho$ gives the correlation between the auxiliary variable and the variable of interest.*

| Population | $\rho$ | $\frac{n}{N}$ | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | |
|---|---|---|---|---|---|---|---|---|
| | | | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ |
| ES-SILC | 0.69 | 0.01 | 8.0 | 7.3 | 12.6 | 9.5 | 18.0 | 18.9 |
| | | 0.05 | 13.3 | 11.7 | 24.4 | 23.8 | 14.8 | 11.9 |
| Log-Normal | 0.50 | 0.01 | 13.1 | 11.0 | 5.6 | 5.4 | 6.1 | 0.6 |
| | | 0.05 | 23.0 | 17.6 | 18.4 | 16.8 | 12.9 | 10.5 |
| | 0.70 | 0.01 | 2.1 | 6.7 | 14.4 | 12.0 | 8.4 | 6.5 |
| | | 0.05 | 17.5 | 10.5 | 15.1 | 13.3 | 14.2 | 18.2 |
| | 0.90 | 0.01 | 4.4 | 10.7 | 3.4 | 8.1 | 17.8 | 12.6 |
| | | 0.05 | 22.4 | 20.6 | 17.4 | 19.8 | 28.9 | 24.5 |
| HMT | 0.76 | 0.014 | 8.0 | 16.9 | 7.4 | 4.1 | 7.5 | 9.5 |

sampling fraction is small. A confidence interval can be computed using the rescaled bootstrap confidence interval (the histogram approach). In this section, we evaluate the empirical performance of this variance estimator and this confidence interval. A set of 10,000 independent simple random samples were selected.

We used the ES-SILC and HMT populations defined in Section 5. In addition, we used artificial populations with variables of interest generated from log-normal distributions. Auxiliary variables correlated with the variable of interest are randomly generated. We consider the following correlation coefficients: 0.5, 0.7 and 0.9. The sample size considered is $n = 200$. The sampling fractions considered are $n/N = 0.01, 0.014$ and $0.05$.

In Table 4, we have the empirical relative biases of the rescaled bootstrap variance estimator. We observe larger relative biases when the sampling fraction is 0.05. The bias does not seem to be affected by the correlation or the level $\alpha$. In Table 5, we have the

*Table 5. Coverage rates (%) of the 95 percent rescaled bootstrap confidence interval (the histogram approach) under simple random sampling when $n = 200$.The column $\rho$ gives the correlation between the auxiliary variable and the variable of interest.*

| Population | $\rho$ | $\frac{n}{N}$ | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | |
|---|---|---|---|---|---|---|---|---|
| | | | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ | $\widehat{Y}_{reg;\alpha}$ | $\widehat{Y}_{regS;\alpha}$ |
| ES-SILC | 0.69 | 0.01 | 94.6 | 94.7 | 93.8 | 93.7 | 94.5 | 93.9 |
| | | 0.05 | 94.6 | 94.9 | 95.8 | 95.9 | 95.1 | 95.2 |
| Log-Normal | 0.50 | 0.01 | 96.0 | 95.7 | 94.7 | 94.4 | 93.7 | 93.4 |
| | | 0.05 | 96.2 | 96.4 | 95.7 | 95.6 | 96.3 | 96.4 |
| | 0.70 | 0.01 | 95.7 | 96.4 | 96.9 | 96.1 | 95.3 | 94.7 |
| | | 0.05 | 97.2 | 97.1 | 96.2 | 94.9 | 95.2 | 95.5 |
| | 0.90 | 0.01 | 95.4 | 96.0 | 94.2 | 94.9 | 96.4 | 95.7 |
| | | 0.05 | 95.7 | 95.5 | 96.0 | 96.4 | 95.3 | 95.8 |
| HMT | 0.76 | 0.014 | 93.3 | 94.8 | 94.3 | 93.5 | 94.6 | 94.3 |

observed coverage rates of the 95 percent rescaled bootstrap confidence interval. All the coverages observed are close to the nominal level of 95 percent. Based on this limited simulation study, it seems preferable to consider bootstrap confidence intervals rather than bootstrap variance, when measuring the accuracy of the proposed estimators.

## 6. Discussion

The proposed estimators are based on a regression estimator of the population mean, which is a technique widely used with survey data. The proposed approach can be applied to many standard surveys. It can be implemented with multistage sampling designs, as the proposed estimators are based upon first-order inclusion probabilities and a regression estimator. Alternative estimators proposed by Chambers and Dunstan (1986) and Rao et al. (1990) can be slightly more accurate than the proposed estimators. However, in order to compute these alternative estimators, it is necessary to know the auxiliary variable for the entire population. The Rao et al. (1990) estimator also requires the joint inclusion probabilities, which can be unknown. The proposed estimators are computationally simpler because they are free of joint inclusion probabilities, they are based on a regression estimator and they can be computed when the auxiliary variable is unknown for the nonsampled units. When the joint inclusion probabilities are known, the accuracy of the proposed estimators can also be improved by inverting the Rao et al. (1990) estimator of the distribution function (or any other estimators) rather than the Hájek-type estimator of the distribution function.

We have considered a regression estimator to take the auxiliary information into account. Other type of estimators based upon auxiliary information (Huang and Fuller 1978:, Deville and Särndal 1992) can also be used instead of a regression estimator. The proposed estimators can also be generalised to several auxiliary variables, since a regression estimator can be easily extended to accommodate this situation. In this article, the auxiliary variables are used to calibrate toward a population mean. This approach can be extended to calibration towards more complex population quantities such as means, quantiles, or variances (e.g. Owen 1991, Chaudhuri et al. 2008, Lesage 2011).

Chen and Wu (2002) proposed a pseudoempirical likelihood approach for estimating quantiles with auxiliary variables. Berger and De la Riva Torres (2015) proposed an empirical-likelihood approach for estimating quantiles with auxiliary variables. Empirical (and pseudoempirical) likelihood approaches are well suited for the estimation of quantiles with auxiliary variables, especially for the calculation of confidence intervals. It would be interesting to investigate how an empirical-likelihood approach could be used to derived confidence intervals for the proposed approach.

## Appendix A: Proof of Lemma 1

We have that

$$\overline{Y}^*_\alpha = \frac{1}{N}\sum_{i\in U} y^*_{\alpha;i} = \frac{1}{N}\sum_{i\in U} \phi^{-1}(F^\circ(y_i)) + z_\kappa, \tag{24}$$

$$F^\circ(y_i) = R_i, \tag{25}$$

where $R_i = N^{-1}(rank(y_i) - 0.5)$ and $rank(y_i)$ is the rank of observation $y_i$ in the population and $\phi^{-1}(\cdot)$ is the quantile function of a $N(0, 1)$ distribution. By substituting (25) into (24), we have that

$$\overline{Y}_{\alpha}^{*} = \frac{1}{N}\sum_{i\in U}\phi^{-1}(R_i) + z_{\kappa} = \frac{1}{N}(S_{<0.5} + S_{>0.5} + S_{0.5}) + z_{\kappa} \tag{26}$$

with

$$S_{<0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i < 0.5),$$

$$S_{>0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i > 0.5),$$

$$S_{0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i = 0.5).$$

It is clear that $S_{0.5} = 0$. Consider a unit $i$ such that $rank(y_i) < (N + 1)/2$. This implies that $R_i < 0.5$. Thus

$$S_{<0.5} = \sum_{r<(N+1)/2}\phi^{-1}((r - 0.5)/N), \tag{27}$$

$$\begin{aligned} S_{>0.5} &= \sum_{r<(N+1)/2}\phi^{-1}((N - r + 1 - 0.5)/N) \\ &= \sum_{r<(N+1)/2}\phi^{-1}(1 - (r - 0.5)/N). \end{aligned} \tag{28}$$

Substituting (27) and (28) into (26), we obtain

$$\overline{Y}_{\alpha}^{*} = \frac{1}{N}\sum_{r<(N+1)/2}\left\{\phi^{-1}((r - 0.5)/N) + \phi^{-1}(1 - (r - 0.5)/N)\right\} + z_{\kappa}. \tag{29}$$

As the normal distribution is symmetric, we have that $\phi^{-1}(p) = -\phi^{-1}(1 - p)$. Hence the sum in (29) equal zero. This implies that

$$\overline{Y}_{\alpha}^{*} = z_{\kappa}. \tag{30}$$

As $F^{\circ}(Y_{\alpha}) = N^{-1}(rank(Y_{\alpha}) - 0.5)$, $rank(Y_{\alpha}) = \lceil \alpha N \rceil$, and $\kappa = N^{-1}(\lceil \alpha N \rceil - 0.5)$, we have that

$$F^{\circ}(Y_{\alpha}) = \kappa. \tag{31}$$

We also have that

$$F^{\circ}(Y_{\alpha}) = \phi(\phi^{-1}(F^{\circ}(Y_{\alpha}))) = \phi(\Psi(Y_{\alpha})). \tag{32}$$

Equations (31) and (32) imply that

$$\phi(\Psi(Y_{\alpha})) = \kappa. \tag{33}$$

As $z_\kappa$ is the $\kappa$th quantile of a normal $N(0, 1)$ distribution, we have that $\phi(z_\kappa) = \kappa$, which combined with (33) gives

$$\phi(z_\kappa) = \phi(\Psi(Y_\alpha)) \cdot$$

The last expression implies

$$z_\kappa = \Psi(Y_\alpha), \tag{34}$$

as $\phi(\cdot)$ is a bijective function. Combining (30) with (34), we have that $\Psi(Y_\alpha) = \overline{Y}^*_\alpha$. The Lemma follows.

## Appendix B: Proof of Theorem 1

As $\phi(\cdot)$ is twice differentiable, a first-order Taylor expansion implies that

$$\phi\left(\bar{y}^*_{reg;\alpha}\right) - \phi\left(\overline{Y}^*_\alpha\right) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) + O_p\left(\left|\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right|^2\right), \tag{35}$$

where $f(y)$ is the density of a $N(0, 1)$ distribution. Equation (30) implies that $\phi\left(\overline{Y}^*_\alpha\right) = \phi(z_\kappa) = \kappa$. Thus, as $\kappa \to \alpha$ as $N \to \infty$, $\lim_{N\to\infty} \phi\left(\overline{Y}^*_\alpha\right) = \alpha$ and we have that

$$\phi(\bar{y}^*_{reg;\alpha}) - \alpha = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) + O_p(n^{-1}), \tag{36}$$

because $\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha = O_p(n^{-1/2})$.

As $Q(\alpha)$ is twice differentiable, a first-order Taylor expansion implies that

$$Q\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - Q(\alpha) = \left(\phi\left(\bar{y}^*_{reg;\alpha}\right) - \alpha\right) Q'(\alpha) + O_p\left(\left|\phi\left(\bar{y}^*_{reg;\alpha}\right) - \alpha\right|^2\right),$$

where $Q'(\alpha) = \partial Q(\alpha)/\partial\alpha$. Assumption (22) and (36) imply that

$$Q\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - Q(\alpha) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}), \tag{37}$$

as $f\left(\overline{Y}^*_\alpha\right)$ is bounded. Using assumption (23), Equation (37) implies that

$$\widehat{F}^{\circ -1}\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - \widehat{F}^{\circ -1}(\alpha) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}). \tag{38}$$

As $\widehat{F}^{\circ -1}\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) = \widehat{Y}_{reg;\alpha}$ and $\widehat{F}^{\circ -1}(\alpha) = \widehat{Y}_\alpha$, equation (38) becomes

$$\widehat{Y}_{reg;\alpha} = \widehat{Y}_\alpha + \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1})$$

which implies

$$\widehat{Y}_{reg;\alpha} - Y_\alpha = \widehat{Y}_\alpha - Y_\alpha + \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}).$$

Thus, the last expression combined with the conditions (21) and (22) implies that $|\widehat{Y}_{reg;\alpha} - Y_\alpha| = O_p(n^{-1/2})$.

## 7. References

Bahadur, R.R. 1966. "A Note on Quantiles in Large Samples." *The Annals of Mathematical Statistics* 37: 577–580.

Berger, Y.G. and C.J. Skinner. 2003. "Variance Estimation of a Low-Income Proportion." *Journal of the Royal Statistical Society Series C* 52: 457–468. DOI: http://dx.doi.org/10.1111/1467-9876.00417.

Berger, Y.G. and O. De la Riva Torres. 2015. "An Empirical Likelihood Approach for Inference Under Complex Sampling Design. To Appear in Journal of Royal Statistical Society, Senes B, 22p.".

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1976. "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations." *Biometrika* 63: 615–620. DOI: http://dx.doi.org/10.1093/biomet/63.3.615.

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1977. *Foundation of Inference in Survey Sampling*. New York: Wiley.

Chambers, R.L. and R. Dunstan. 1986. "Estimating Distribution Functions From Survey Data." *Biometrika* 73: 597–604. DOI: http://dx.doi.org/10.1093/biomet/73.3.597.

Chaudhuri, S., M.S. Handcock, and M.S. Rendall. 2008. "Generalized Linear Models Incorporating Population Level Information: An Empirical-Likelihood-Based Approach." *Journal of the Royal Statistical Society – Series B (Statistical Methodology)* 70: 311–328. DOI: http://dx.doi.org/10.1111/j.1467-9868.2007.00637.x.

Chen, J. and C. Wu. 2002. "Estimation of Distribution Function and Quantiles Using Model-Calibrated Pseudo Empirical Likelihood Method." *Statistica Sinica* 12: 1223–1239.

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: http://dx.doi.org/10.1080/01621459.1992.10475217.

Dorfman, A.H. 2009. "Inference on Distribution Functions and Quantiles." In *Handbook of Statistics 29B Sample Surveys: Inference and Analysis*, edited by D. Pfeffermann and C.R. Rao, pp. 371–395. Amsterdam, North-Holland: Elsevier.

Eurostat. 2003. *"Laeken" Indicators-Detailed Calculation Methodology, Directorate E: Social Statistics, Unit E-2: Living Conditions, DOC.E2/IPSE/2003*. Available at: http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken%20Indicators%20-%20calculation%20algorithm.pdf.

Eurostat. 2012. *European Union Statistics on Income and Living Conditions (EU-SILC)*. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc.

Francisco, C.A. and W.A. Fuller. 1991. "Quantile Estimation With a Complex Survey Design." *Annals of Statistics* 19: 454–469.

Hájek, J. 1971. Comment on a paper by D. Basu. In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston.

Hansen, M.H., W.G. Madow, and B.J. Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78: 776–793. DOI: http://dx.doi.org/10.1080/01621459.1983.10477018.

Harms, T. and P. Duchesne. 2006. "On Calibration Estimation for Quantiles." *Survey Methodology* 32: 37–52.

Huang, E.T. and W.A. Fuller. 1978. "Nonnegative Regression Estimation for Survey Data." In Procceding of the Social Statistics Section of the American Statistical Association, Washington DC, 300–303.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. DOI: http://dx.doi.org/10.1080/01621459.1982.10477770.

Lehmann, E.L. 1999. *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Lesage, E. 2011. "The Use of Estimating Equations to Perform a Calibration on Complex Parameters." *Survey Methodology* 37: 103–108.

Nygård, F. and A. Sandström. 1985. "The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations." *Journal of Official Statistics* 4: 399–412.

Osier, G. 2009. "Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques." *Journal of the European Survey Research Association* 3: 167–195.

Owen, A.B. 1991. "Empirical Likelihood for Linear Models." *The Annals of Statistics* 19: 1725–1747.

Rao, J.N.K., J.G. Kovar, and H.J. Mantel. 1990. "On Estimating Distribution Functions and Quantiles From Survey Data Using Auxiliary Information." *Biometrika* 77: 365–375. DOI: http://dx.doi.org/10.1093/biomet/77.2.365.

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83: 231–241. DOI: http://dx.doi.org/10.1080/01621459.1988.10478591.

Rao, J.N.K., C.F.J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18: 209–217.

Robinson, P.M. and C.-E. Särndal. 1983. "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling." *Sankhya B* 45: 240–248.

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

Serfling, N. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Silva, P.L.D., Nascimento and C.J. Skinner. 1995. "Estimating Distribution Functions With Auxiliary Information Using Poststratification." *Journal of Official Statistics* 11: 277–294.

Verma, V. and G. Betti. 2011. "Taylor Linearization Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics." *Journal of Applied Statistics* 38: 1549–1576. DOI: http://dx.doi.org/10.1080/02664763.2010.515674.

Received October 2013
Revised October 2014
Accepted November 2014