On estimating quantiles using auxiliary information

Berger, Yves G.
University of Southampton, Southampton Statistical Sciences Research Institute
Southampton SO17 1BJ, United Kingdom
E-mail: Y.G.Berger@soton.ac.uk

Muñoz, Juan F.
University of Granada, Department of Quantitative Methods in Economics and Business
Granada, 18071, Spain
E-mail: jfmunoz@ugr.es

Estimation of quantiles may be of considerable interest when measuring income distribution and poverty lines. For instance, the median is regarded as a more appropriate measure of location than the mean when variables, such as income, expenditure, etc, exhibit highly skewed distributions. In sample surveys, auxiliary information is often used at the estimation stage to increase the precision of estimators of means. The use of auxiliary information has been studied extensively for estimation of means, but it has no obvious extensions to the estimation of quantiles. A novel method for estimating quantiles using auxiliary information is proposed. The proposed estimator is based upon the regression estimator of a transformed variable of interest. Simulation studies support our findings and show that the proposed estimator can be more accurate than or as accurate as alternative estimators (Chambers & Dunstan, 1986; Rao et al. 1990; Silva & Skinner 1995) which can be computationally more intensive.

1. Introduction

In sample surveys, auxiliary information is often used at the estimation stage to increase the precision of estimators of means (e.g. Cassel, Särndal & Wretman, 1976, 1977). The use of auxiliary information has been studied extensively for estimation of means, but it has no obvious extensions to the estimation of quantiles. In this paper, we propose a novel estimator for quantiles that takes auxiliary information into account.

We consider a finite population $U=\{1,\ldots,i,\ldots N\}$ containing N units. Let $y_1,\ldots,y_N$ denote the values of the variable of interest, y, and $x_1,\ldots,x_N$ denote the values of an auxiliary variable, x. We assume that $\overline{X}=(1/N)\sum_{i\in U}x_i$ is known. An estimator of a quantile of y which uses the association between y and x is usually accurate because the variable x contains more information than the variable y, as $\overline{X}$ is known or the values of x may be known for all the units of the population.

The aim is to estimate the population quantile

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

where $F(t)=N^{-1}\sum_{i\in U}\delta\{y_i\le t\}$ is the population distribution function and $\delta\{\cdot\}$ is the indicator function which takes the value 1 if its argument is true and 0 otherwise. The function $F^{-1}(\cdot)$ is the inverse of the function $F(\cdot)$. Throughout this paper, we define the inverse of any function $G(\cdot)$ by $G^{-1}(\alpha)=\inf\{t:G(t)\ge\alpha\}$.

Assume that a sample s, of size n, is selected from U according to a sampling design, and that the values y_i are known for all the sampled units $i\in s$. Following the definition of Y_α given in (1), the customary estimator for Y_α is obtained by substituting $F(t)$ by an estimator into (1). There exists a wide range of

estimators for the distribution function using auxiliary information (e.g. Chambers & Dunstan, 1986; Rao et al. 1990; Silva & Skinner 1995) or an estimator which does not use auxiliary information. For example, the Hájek type estimator of Y_α is defined by

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

where

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

and $\hat{N} = \sum_{i \in s} \pi_i^{-1}$, where π_i denotes the inclusion probability of unit i.

2. Transformation of the variables

Consider the transformed values

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

where $\Psi(y_i) = \phi^{-1}(F^\circ(y_i))$ and $\phi(\cdot)$ is the cumulative distribution function of a normal $N(0,1)$ distribution. The function $F^\circ(y_i)$ is the mid-point distribution function (Nygård & Sandström, 1985) of the variable of interest defined by

$$F^\circ(y_i) = \frac{1}{2}[F(y_{i-1}) + F(y_i)].$$

The quantity z_κ is a known offset term which is given by the κ-th quantile of a normal $N(0,1)$ distribution; that is, $z_\kappa = \phi^{-1}(\kappa)$, where $\kappa = (\lceil \alpha N \rceil - 0.5)/N$. Note that κ can be approximated by α as $\kappa \to \alpha$ when $N \to \infty$.

We propose to transform the auxiliary information the same way. Consider

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

where $\Psi_x(x_i) = \phi^{-1}(F_x^\circ(x_i))$, $F^\circ(y_i) = [F_x(x_{i-1}) + F_x(x_i)]/2$ and $F_x(t) = N^{-1} \sum_{i \in U} \delta\{x_i \le t\}$.

As the transformed values in (4) depend on population values, they would need to be estimated. We propose to estimate $y^*_{\alpha;i}$ by its substitution estimator given by

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

where $\hat{\Psi}(y_i) = \phi^{-1}(\hat{F}^\circ(y_i))$. The function $\hat{F}^\circ(y_i)$ is an empirical mid-point estimator of the distribution function of the variable of interest. This function is given by

$$\hat{F}^\circ(y_i) = \frac{1}{2}[\hat{F}(y_{i-1}) + \hat{F}(y_i)],$$

where $\hat{F}(y_i)$ is an estimator of the distribution function.

We propose to estimate

$$\bar{Y}^*_\alpha = N^{-1} \sum_{i \in U} y^*_{\alpha;i}$$

using the regression estimator (e.g. Cassel, Särndal & Wretman, 1976, 1977), which makes use of the auxiliary information. This estimator is defined by

$$\bar{y}^*_{reg;\alpha} = \bar{y}^*_\alpha + \hat{\beta}_x(\bar{X}^*_\alpha - \bar{x}^*_\alpha)$$

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

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

3. Proposed estimator for a quantile

We have that $Y_\alpha = \Psi^{-1}(\bar{Y}^*_\alpha)$ where the function $\Psi^{-1}(\cdot)$ is the inverse of function $\Psi(\cdot)$ in (4). The proposed estimator for the α-th quantile Y_α is the substitution estimator given by

$$\hat{Y}_{reg;\alpha} = \hat{\Psi}^{-1}(\bar{y}^*_{reg;\alpha}).$$

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

$$\hat{Y}_{reg;\alpha} = \hat{F}^{\circ -1}(\hat{\alpha}_{reg}), \tag{5}$$

where $\hat{\alpha}_{reg} = \phi(\bar{y}^*_{reg})$. The proposed estimator consists in inverting a mid-point distribution function at the value $\hat{\alpha}_{reg}$ which is adjusted to take into account of the auxiliary variable. Note that the proposed estimator is not affected by outliers, because $y^*_{\alpha;i}$ and $x^*_{\alpha;i}$ are implicitly based upon the ranks of y_i and x_i.

4. Simulation study

In this section, the proposed estimator is compared numerically with alternative estimators (Chambers & Dunstan, 1986; Rao et al. 1990; Silva & Skinner 1995). In this simulation study, we consider the worst case scenario when the proposed estimator is based upon the Hájek type distribution function $\hat{F}_\pi(t)$ given by (3). Note that the accuracy of the proposed estimator can be improved by inverting the Rao *et al*. (1990) estimator for the distribution function (or any other estimator) rather than the Hájek type distribution function (3).

The simulation study is based upon 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 cane. The Sugar population was considered by several studies on estimation of distribution function and quantiles (Chambers & Dunstan, 1986; Rao et al., 1990; and Silva & Skinner, 1995). In order to analyse the effect of the correlation between the variable of interest and the auxiliary variable, we considered two cases for the auxiliary variable: (i) the total cane harvested (denoted by Sugar-1), and (ii) the area assigned for cane planting (denoted by Sugar-2).

The population of municipalities (Särndal, Swensson & Wretman, 1992, page 652) consists of 284 municipalities, where the variable of interest is the population in 1985. This population of municipalities was replicated four times to achieve a population size of $N = 1136$ individuals. We considered several auxiliary variables: (i) the number of conservative seats in municipal council (denoted by MUN-1), (ii) the total number of seats in municipal council (denoted by MUN-2), and (iii) the population in 1975 (denoted by MUN-3).

The 'Labor' population (Valliant, Dorfman & Royall, 2000, page 434) comprises $N = 478$ individuals, the variable of interest y is the usual amount of weekly wages and x is the variable age. We consider also a population of $N = 80$ factories (Murthy, 1967, page 228). For this population, y is the output for factories

and x is the number of workers. We consider the Hansen, Madow & Tepping (1983) population (denoted by HMT), which is $N = 14000$ units generated from a bivariate gamma population (Rao *et al.*, 1990). We also considered $N = 3114$ families extracted from the Spanish Household Panel Survey (SHPS), where y is the income and x is the expenditure.

Table 1: Empirical RRMSE of the estimators of quantiles. $RRMSE \times 100$ of estimators of Y_α, with α=0.25, 0.5 and 0.75, under simple random sampling. γ_y is the population skewness of y, γ_x is the population skewness of x and ρ is the correlation coefficient between y and x.

			RRMSE (in %)				
		α	Hájek (3)	Rao *et al.*	Chamber & Dunstan	Silva & Skinner	Proposed (14)
Sugar-1	$n = 30$	0.25	12.0	9.4	6.6	10.0	9.5
$\rho = 0.89$	$\gamma_y = 2.4$	0.50	12.1	9.2	9.3	9.9	8.8
	$\gamma_x = 2.3$	0.75	15.4	10.9	14.9	11.7	12.9
Sugar-2	$n = 30$	0.25	11.7	5.9	5.6	8.3	7.7
$\rho = 0.98$	$\gamma_y = 2.4$	0.50	11.9	6.6	7.5	7.1	8.2
	$\gamma_x = 2.3$	0.75	14.3	7.9	10.5	7.8	10.9
MUN-1	$n = 200$	0.25	7.7	7.4	10.0	11.0	10.9
$\rho = 0.61$	$\gamma_y = 8.2$	0.50	8.1	7.1	28.0	7.5	7.0
	$\gamma_x = 1.2$	0.75	7.0	5.3	16.5	5.1	5.3
MUN-2	$n = 200$	0.25	7.3	7.3	21.5	18.7	10.8
$\rho = 0.69$	$\gamma_y = 8.2$	0.50	8.1	6.9	15.5	15.3	6.2
	$\gamma_x = 1.4$	0.75	7.0	5.8	5.7	21.1	4.5
MUN-3	$n = 200$	0.25	7.3	6.2	9.5	10.0	10.4
$\rho = 0.998$	$\gamma_y = 8.2$	0.50	8.0	3.8	5.0	4.6	5.8
	$\gamma_x = 8.5$	0.75	6.6	2.5	0.5	2.2	4.0
Labor	$n = 100$	0.25	9.3	9.3	8.6	9.8	10.0
$\rho = 0.19$	$\gamma_y = 1.2$	0.50	8.4	8.6	10.8	8.3	8.9
	$\gamma_x = 0.5$	0.75	7.4	7.8	10.0	7.3	7.6
SHPS	$n = 500$	0.25	3.4	2.8	10.5	2.9	2.8
$\rho = 0.60$	$\gamma_y = 4.7$	0.50	2.7	2.3	3.0	2.4	2.3
	$\gamma_x = 2.4$	0.75	2.9	2.4	9.9	2.4	2.4
HMT	$n = 200$	0.25	8.3	6.1	6.1	6.4	6.6
$\rho = 0.76$	$\gamma_y = 2.0$	0.50	7.6	5.8	11.1	6.0	5.9
	$\gamma_x = 1.4$	0.75	7.1	5.7	11.6	5.9	5.9
Factories	$n = 30$	0.25	6.0	3.1	23.6	3.9	4.0
$\rho = 0.92$	$\gamma_y = 0.1$	0.50	8.7	3.4	8.2	4.5	4.6
	$\gamma_x = 1.3$	0.75	6.5	3.3	11.4	3.7	3.4

For each simulation, 1000 samples were selected to compute the empirical relative root mean square error $RRMSE = 100\% \times MSE[\hat{Y}_\alpha]^{1/2} / Y_\alpha$, where $MSE[\cdot]$ denote the mean square error. Simple random sampling is used to select the samples. The population quartiles $Y_{0.25}$, $Y_{0.5}$ and $Y_{0.75}$ are the parameters of interest.

The efficiency of the estimators is measured by the empirical relative root mean square errors (RRMSE) which are reported in Table 1. The proposed estimator is less efficient than other estimators when correlation is close to 1 (Sugar-2 and MUN-3 populations). The proposed estimator is the value of the inverse distribution function taken on $\hat{\alpha}_{reg}$ instead of α; where $\hat{\alpha}_{reg}$ is defined in (14). As the distribution function is based upon the Hájek type distribution function given by (3), we notice a clear improvement of using $\hat{\alpha}_{reg}$ instead of α, because the RRMSEs of the proposed estimator is usually smaller than of the RRMSEs of the Hájek estimator, except with estimator of $Y_{0.25}$ of the MUN populations.

Alternative estimators proposed by Chamber & Dunstan (1986), Rao *et al*. (1990) and Silva & Skinner (1995) can slightly more accurate than the proposed estimator. The Rao *et al*. (1990) estimator is usually the most accurate and is only slightly more accurate than the proposed estimator. However the Rao *et al*. (1990) estimator requires joint-inclusion probabilities which are not necessary for the proposed estimator. Note that with the population MUN-2, the proposed estimator is almost as accurate as the Rao *et al*. (1990) estimator, and the other estimators (Chamber & Dunstan, 1986; Silva & Skinner, 1995) have larger RRMSE.

5. Discussion

The proposed estimator for the quantile α is the value of an inverse mid-point distribution taken on $\hat{\alpha}_{reg}$ instead of α; where $\hat{\alpha}_{reg}$ takes into account of the auxiliary variable and is defined in (14). When the distribution function is the Hájek type mid-point distribution function given by the mid-point version of (3), the simulation study shows that there is a clear improvement of using $\hat{\alpha}_{reg}$ instead of α.

The proposed estimator is based on the regression estimator which is a technique widely used for survey data. The proposed estimator can demonstrate clear improvements compared with existing estimators in term of accuracy and simplicity. It can be easily extended for more than one auxiliary variable and to other technique of calibration. The proposed estimator is computationally simpler because it is free of joint inclusion probabilities.

REFERENCES

Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63, 615-20.

Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1977). Foundation of inference in survey sampling. John Wiley, New York.

Chambers, R. L. & Dunstan, R. (1986). Estimating distribution functions from survey data. Biometrika 73, 597-604.

Hansen, M. H., Madow, W. G & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Amer. Statist. Assoc. 78, 776-93.

Murthy, M. N. (1967). Sampling theory and methods. Calcutta: Statistical Publishing Society.

Nygård, F. & Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. J. Offic. Statist. 4, 399-412.

Rao, J. N. K., Kovar, J. G. & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika 77, 365-375.

Silva, P. L. D. & Skinner, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification. J. Offic. Statist. 11, 277-94.

RÉSUMÉ

Nous proposons une nouvelle approche pour estimer un quantile. Cette approche utilise l'information auxiliaire et est basée sur une transformation de la variable d'intérêt. Nous présenterons également quelques résultats de simulation.