Quantile Model-Assisted Estimation Approach for Survey Data

Zea, José Fernando
Department of Statistics
National University of Colombia
Bogota, Colombia
Email: jfzeac@unal.edu.co

Trujillo, Leonardo
Department of Statistics
National University of Colombia
Bogota, Colombia
Email: ltrujilloo@bt.unal.edu.co

Introduction

The use of auxiliary information to improve the accuracy of a total estimator has a long history in sampling theory. The linear regression model and its multiple variants have been used to assist the estimation of totals in survey sampling. Parametric and nonparametric models have been used in order to incorporate the auxiliary information. Estimators using these type of models are commonly name as model-assisted estimators. Estimators using the available auxiliary information in the population usually have better performance in terms of a small variance.

This kind of estimators include GREG type estimators (see for example, Pfefferman and Rao, 2011), calibration estimators (Deville and Särndal, 1992), calibration estimators based on neural networks (Montanari and Ranalli, 2005), local polynomial regression estimators (Breidt and Opsomer, 2000) and Wilcoxon rank based estimators (Gutierrez and Breidt, 2009), among many others. Model assisted estimators are approximately unbiased irrespective of whether the assumptions of the models hold or not. Nevertheless, it is required that the model fits reasonably good in order to achieve an efficient use of the auxiliary information (Särndal, Swensson and Wretman, 1992). The existence of extreme observations such as influential points and/or outliers implies that the use of simple linear regression could be inappropriate in order to assist the estimation of a total population. Thus, the use of robust regression techniques appears as an alternative. Other authors that have already worked in the problem of dealing with outliers in survey data are Chambers (1986, 2000) and Lee (1995).

In particular, in this document, we will show that the use of quantile regression models reduces the effect of influential points on the model fitting and therefore gives a smaller variance. Also, the proposed estimator shows better performance than the GREG estimator when the normality assumption does not hold. The proposed estimator considers a median regression model in order to assist the total estimation. The rapid increase in computational capabilities nowadays makes that the use of the proposed approach is easy to implement.

The document is structured as follows. The first section summarizes the essentials of quantile regression. Then, in the second section we propose an estimator of the quantile regression parameters based on survey data. In the third section, an estimator of the population total is presented using median regression. Then, the performance of the estimators is evaluated empirically through a simulation study. The paper ends with some conclusions and proposals for future research.

Quantile Regression

Koenker and Bassett (1978) introduced the quantile regression as a robust alternative to the least squares estimation for the linear model. Quantile regression has shown higher efficiency over a wide range of error distributions and under the presence of influential points.

Koenker (2005) defines the τ-th quantile $Q^{(\tau)}$ as $Q^{(\tau)} = \inf\{y : F(y) \geq \tau\}$ where, $F(y) = P(Y \leq y)$, $0 \leq \tau \leq 1$. The τ-th quantile can be found solving the equation

$$(1) \quad \min_{Q^{(\tau)}} \left\{ \tau \int_{y \geq Q^{(\tau)}} |y - Q^{(\tau)}| + (1-\tau) \int_{y < Q^{(\tau)}} |y - Q^{(\tau)}| \right\}$$

On the other hand, the τ-th sample quantile (in particular for the median, $\tau = 0.5$) can be alternatively found solving:

$$(2) \quad \min_{Q^{(\tau)}} \left\{ \tau \sum_{y_k \geq Q^{(\tau)}} \mid y_k - Q^{(\tau)} \mid + (1-\tau) \sum_{y_k < Q^{(\tau)}} \mid y_k - Q^{(\tau)} \mid \right\}$$

Let us consider $y_1, y_2, \ldots y_N$ the values of the variable of study for each element k in the population. Suppose there are p auxiliary variables denoted x_1,...x_p. For the k-th element, the auxiliary vector $\boldsymbol{x}_k = (x_{1k}, ..., x_{pk})'$ is defined for $k = 1, 2, ...N$. Suppose there is a superpulation model ξ such that $y_k = \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k + \varepsilon_k$, for $k = 1, 2, \ldots, N$. The estimation of the parameters of the quantile regression is carried out in an analogous way to the quantile estimation above. The minimization problem can be expressed now as:

$$(3) \quad \min_{\boldsymbol{B}^{(\tau)}} f(\boldsymbol{B}^{(\tau)}) = \min_{\boldsymbol{B}^{(\tau)}} \tau \sum_{y_k \geq \boldsymbol{B}^{(\boldsymbol{\tau})\prime}\boldsymbol{x}_k} \mid y_k - \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k \mid + (1-\tau) \sum_{y_k < \boldsymbol{B}^{(\boldsymbol{\tau})\prime}\boldsymbol{x}_k} \mid y_k - \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k \mid$$

This is essentially an optimization problem that could be solved using linear programming techniques. More details can be found in Koenker and D'Orey (1987, 1994), Koenker (2005), Mora (2005) and Hao and Naiman (2007). More advanced techniques in quantile regression including weights is developed in detail in Koenker (2005). Suppose a vector of weights $w_1, ...w_N$ is available in the population U. The estimators for the weighted quantile regression parameters are estimated by solving:

$$(4) \quad \min_{\boldsymbol{B}^{(\tau)}} f(\boldsymbol{B}^{(\tau)}) = \min_{\boldsymbol{B}^{(\tau)}} \tau \sum_{y_k \geq \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k} w_k \mid y_k - \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k \mid + (1-\tau) \sum_{y_k < \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k} w_k \mid y_k - \boldsymbol{B}^{(\tau)\prime}\boldsymbol{x}_k \mid$$

Estimation of the Quantile Regression Parameters for Survey Data

A sampling estimator for $\hat{\boldsymbol{B}}^{(\tau)}$ is proposed following the ideas for the weighted quantile regression in the previous section above. The quantile regression parameters for data coming from a sample survey s can be estimated using a median regression ($\tau = 0.5$) and solving:

$$(5) \quad \min_{\hat{\boldsymbol{B}}^{(\tau)}} f(\hat{\boldsymbol{B}}^{(\boldsymbol{\tau})}) = \min_{\hat{\boldsymbol{B}}^{(\tau)}} \tau \sum_{s \cap \left\{ y_k \geq \hat{\boldsymbol{B}}^{(\boldsymbol{\tau})\prime}\boldsymbol{x}_k \right\}} \frac{\mid y_k - \hat{\boldsymbol{B}}^{(\boldsymbol{\tau})\prime}\boldsymbol{x}_k \mid}{\pi_k} + (1-\tau) \sum_{s \cap \left\{ y_k < \hat{\boldsymbol{B}}^{(\tau)\prime}\boldsymbol{x}_k \right\}} \frac{\mid y_k - \hat{\boldsymbol{B}}^{(\tau)\prime}\boldsymbol{x}_k \mid}{\pi_k}$$

π_k is the first order inclusion probability of the unit k. There is not a closed expression for $\hat{\boldsymbol{B}}^{(\tau)}$ and then the Taylor linearization method cannot be used in order to get a variance estimator. Other different values of τ could be used in order to achieve a more complete understanding of how the response distribution is affected by the covariables.

Quantile Regression Based Estimator

The quantile regression based estimator for a population total can be obtained in an analogous way of the GREG estimator. A linear relation between the study variable y and a vector of covariables $\boldsymbol{x}$ is asumed such as $y_k = \boldsymbol{x}_k'\boldsymbol{B}^{(\tau)}+E_k$, where $E_k = y_k - y_k^0$ and $y_k^0 = \boldsymbol{x}_k'\boldsymbol{B}^{(\tau)}$. Then, the total population can be expressed as $t_y = \sum_U(\hat{y}_k + E_k) = \sum_U \boldsymbol{x}_k'\boldsymbol{B}^{(\tau)} + \sum_U(y_k - \boldsymbol{x}_k'\boldsymbol{B}^{(\tau)})$. Then, analogously to the least squares approach, a quantile regression based estimator for the total population is also:

$$\hat{t}_y = \sum_U \boldsymbol{x}_k'\hat{\boldsymbol{B}}^{(\tau)} + \sum_s \frac{y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}}^{(\tau)}}{\pi_k} = \hat{t}_{y\pi} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x}})'\hat{\boldsymbol{B}}^{(\tau)} \quad (6)$$

where $\hat{\boldsymbol{B}}^{(\tau)}$ is estimated through quantile regression.

We will denote as $\hat{t}_{yqr}$, the quantile regression estimator (QREG) obtained considering a median regression. In other words, $\hat{t}_{yqr} = \hat{t_{y\pi}} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x}})'\hat{\boldsymbol{B}}^{(0.5)}$. This expression can be written as $\hat{t}_{yqr} = \sum_s \frac{E_k}{\pi_k} + \left(\sum_U \boldsymbol{x_k'}\right)\boldsymbol{B} = \sum_s \breve{E}_k + \left(\sum_U \boldsymbol{x_k'}\right)\boldsymbol{B}$. Then, the variance can be written as $V(\hat{t}_{yqr}) = \sum\sum_U \Delta_{kl}\breve{E}_k\breve{E}_l$, and the variance estimator $\hat{V}(\hat{t}_{yqr}) = \sum\sum_s \breve{\Delta}_{kl}\breve{e}_k\breve{e}_l$, where $e_k = y_k - \hat{y}_k$, $\hat{y}_k = \boldsymbol{x}_k'\hat{\boldsymbol{B}}^{(\tau)}$ and $\breve{e_k} = e_k/\pi_k$.

Simulation

In order to evaluate the accuracy and the efficiency of the proposed estimator, some simulations were used to analyze different models changing the distribution of the residuals and considering different variance structures. For the correctly specified variance structure models, the quantile regression model did not take into account the variance structure. Additionally, different scenarios with different types of outliers were made in order to compare the performance of the quantile regression estimator with some well-know estimators such as the Horvitz-Thompson (HT) estimator and the GREG estimator. It is shown, after the simulation results, that the QREG estimator is more efficient than the other two estimator under non-normal distribution assumptions and under the presence of influential points.

A Monte Carlo simulation was carried out considering $M = 5,000$ repeated SI samples with $n = 100$ from a population of $N = 1,000$ elements. A superpopulation model $y_k = 10 + 2x_k + \varepsilon_k$ such that $E_\xi(y_k) = 10 + 2x_k$ and $V_\xi(y_k) = \sigma_k$ was specified for all $k = 1, 2, ...N$. The values of the auxiliary variable were generated from an exponencial distribution with parameter equal to 1. The ε_k were assumed independent and normally distributed. In each sample, three estimators: the HT estimator t_π, the GREG estimator $\hat{t_{yreg}}$ and the QREG estimator $\hat{t_{yqr}}$ were calculated. The simulations were carried out by using the statistical software R 2.12.2. The algorithms are available under request to the authors.

The performance of the estimators is evaluated in terms of their Relative Bias (RB) and their Mean Square Error (MSE). For purpose of comparison between an estimator $\hat{t_y}$ with the QREG estimator, we used the relative efficiency (RE) defined by $\boldsymbol{RE}(\hat{t_y}, \hat{t}_{yqr}) = \frac{MSE(\hat{t_y})}{MSE(\hat{t}_{yqr})}$. Values of RE bigger than one indicates that precision is gained using the QREG estimator. Ratios close to one suggest that one is not losing efficiency with the estimator proposed in this paper.

The model specifications considered in this simulation were as follows:

1. M_1: *Linear model with correctly specified variance structure, normal, uncorrelated and homoscedastic errors* ($E_\xi(y_k) = 10 + 2x_k$ and $V_\xi(y_k) = 1$).

2. M_2: *Linear model with correctly specified variance structure, normal, uncorrelated and heteroscedastic errors.* The superpopulation model ξ has the same $E_\xi(y_k)$ than M_1 but $V_\xi(y_k) = \sqrt{x_k}$.

3. M_3: *Linear model with incorrectly specified variance structure, normal, uncorrelated and heteroscedastic errors.* The superpopulation model is the same as M_2 but the model is incorrectly specified for the GREG estimator ignoring the variance structure.

4. M_4: *Linear model with correctly specified variance structure and non-normal, uncorrelated and homoscedastic errors.* This model assumes an exponential distribution with parameter equal to one.

5. M_5: *Linear model with five percent of contaminated data with a mixture of normal distributions for the residuals.* The errors were generated with a mixture of normal distributions with a zero mean for the 95% of the data and a mean of 5% for the remaining data. The variances for both distributions were equal to 1.

Apart from these first five scenarios, different configurations of extreme observations were considered defining new values for x_k and y_k as:

$$(7) \quad x_k^* = \begin{cases} x_k & 99\ \%\ \text{of no contaminated points} \\ \delta_k & 1\ \%\ \text{of contaminated points} \end{cases}, \; y_k^* = \begin{cases} y_k & 99\ \%\ \text{of no contaminated points} \\ \omega_k & 1\ \%\ \text{of contaminated points} \end{cases}$$

with δ_k and ω_k defined for every particular case below.

6. M_6: *Linear model with normal residuals and the presence of 1% of outlier points in the x-axis.* Ten points out of $n = 100$ were randomly contaminated increasing their values in the x-axis and the values in the y-axis were simulated on the range of the original data according to (7). Four different scenarios under this particular model were analyzed for different configurations of δ_k and ω_k as follows:

   - M_{6a}: $\delta_k \sim U(\overline{x} + 3s, \overline{x} + 4s)$ and $\omega_k \sim U(\min(y), Q^{(0.05)}(y))$.
   - M_{6b}: δ_k simulated as in M_{6a} and $\omega_k \sim U(Q^{(0.95)}(y), \max(y))$.
   - M_{6c}: $\delta_k \sim U(\overline{x} + 5s, \overline{x} + 6s)$ and ω_k simulated as M_{6a}.
   - M_{6d}: δ_k simulated as M_{6c} and ω_k simulated as M_{6b}.

7. M_7: *Linear models with normal residuals and the presence of 1% of outlier points in the y-axis*: Ten points out of $n = 100$ were randomly contaminated increasing their values in the y-axis and the values in the x-axis were simulated on the range of the original data according to (7). Four different scenarios under this particular model were analyzed for different configurations of δ_k and ω_k as follows:

   - M_{7a}: $\delta_k \sim U(\min(x), Q^{(0.05)}(x))$ and $\omega_k \sim U(\overline{y} + 3s, \overline{y} + 4s)$.
   - M_{7b}: $\delta_k \sim U(Q^{(0.95)}(y), \max(x))$ and ω_k simulated as M_{7a}.
   - M_{7c}: δ_k simulated as in M_{7a} and $\omega_k \sim U(\overline{y} + 5s, \overline{y} + 6s)$.
   - M_{7d}: δ_k simulated as M_{7b} and ω_k simulated as in M_{7c}.

8. M_8: *Linear models with normal residuals and the presence of* 1% *of outlier points in the y-axis and the x-axis.* Ten points out of $n = 100$ were randomly contaminated increasing their values in both the x-axis and the y-axis. Two different scenarios under this particular model were analyzed for different configurations of δ_k and ω_k as follows:

   - M_{8a}: $\delta_k \sim U(\overline{x} + 3s, \overline{x} + 4s)$ and $\omega_k \sim U(\overline{y} + 3s, \overline{y} + 4s)$.
   - M_{8b}: $\delta_k \sim U(\overline{x} + 5s, \overline{x} + 6s$ and $\omega_k \sim U(\overline{y} + 5s, \overline{y} + 6s)$.

Model	Measure	$\hat{t}_{\pi}$	$\hat{t}_{yreg}$	$\hat{t}_{yqr}$	Model	Measure	$\hat{t}_{\pi}$	$\hat{t}_{yreg}$	$\hat{t}_{yqr}$
M_1	RB	0.0024	0.00037	0.00021	M_{6d}	RB	-0.02661	-0.00553	-0.00181
	MSE	311044	9293	9388		MSE	296138	16407	9536
M_2	RB	0.0019	-0.00058	-0.00145	M_{7a}	RB	-0.02712	0.00246	-0.0023
	MSE	448529	139172	140531		MSE	296109	11093	9299
M_3	RB	0.0019	-0.00098	-0.00145	M_{7b}	RB	-0.02712	0.00088	-0.00223
	MSE	448529	139100	140531		MSE	296109	9631	9340
M_4	RB	-0.00185	-0.00477	-0.00559	M_{7c}	RB	-0.02704	0.00424	-0.00229
	MSE	310422	9326	9256		MSE	296109	13083	9299
M_5	RB	-0.0084	-0.00192	-0.00308	M_{7d}	RB	-0.02704	0.00389	-0.00222
	MSE	273594	10724	9819		MSE	296109	10664	9340
M_{6a}	RB	-0.02678	-0.00844	-0.00186	M_{8a}	RB	0.0167	0.00246	-0.00177
	MSE	296108	17172	9519		MSE	45193	25485	9240
M_{6b}	RB	-0.02661	-0.00321	-0.00181	M_{8b}	RB	-0.02649	-0.00032	-0.00193
	MSE	296138	10154	9547		MSE	299140	11852	9175
M_{6c}	RB	-0.02678	-0.01046	-0.00186					
	MSE	296108	37195	9508					

Table 1: RB and MSE of the estimators under the different scenarios considered.

Table 1 above summarizes the results of RB and MSE for the three estimators under the different scenarios considered in the simulation. Table 2 shows the Relative Efficiency of QREG with respect to HT and GREG.

Model	$\hat{t}_{\pi}$	$\hat{t}_{yreg}$	Model	$\hat{t}_{\pi}$	$\hat{t}_{yreg}$
M_1	33.13	0.99	M_{6d}	31.14	3.91
M_2 yM_3	3.19	0.99	M_{7a}	31.84	1.19
M_4	33.54	1.01	M_{7b}	31.7	1.03
M_5	27.86	1.09	M_{7c}	31.84	1.41
M_{6a}	31.11	1.8	M_{7d}	31.7	1.14
M_{6b}	31.02	1.06	M_{8a}	32.6	1.29
M_{6c}	14.39	5.62	M_{8b}	4.89	2.76

Table 2: Relative Efficiency of the QREG estimator.

Conclusions and Areas of Further Work

The aim of this research was to obtain a more efficient estimator under high skewed distributions and the presence of extreme observations. The proposed QREG estimator was based on quantile regression using auxiliary information. According to the simulation results, the QREG estimator has similar performance than the GREG estimator in terms of the mean square error under normal distribution of the residuals ($M1$ - $M3$). All the three estimators have a negligible relative bias. Therefore, supposing regular conditions and even in the absence of homoscedasticity, the GREG and the QREG estimators lead to similar results in terms of MSE and RB. If a non-normal distribution is considered for the residuals (M_4 with an exponential distribution and M_5 with a mixture of normal distributions for the residuals), the QREG performs slightly better than the GREG estimator.

In the scenarios with extreme points, the QREG estimator is considerably better in terms of MSE than the GREG estimator specifically in the scenarios in which the contaminated points are further to the mean of the x-axis and when the contamination in the y-axis induces a dramatic change in the slope of the model. For instance, under scenarios M_{6c} and M_{6d} and according to table 2, the MSE of the QREG estimator is less than one third than the MSE of the GREG estimator. In scenarios with outliers in the y-axis, the QREG estimator works better although the reduction in terms of MSE

is not so extreme. Finally, in scenarios considering extreme points both in the x-axis and the y-axis, the QREG estimator has better performance than the GREG estimator. All the considered scenarios in the simulation are not uncommon in the practice of survey sampling.

Areas of further research are the consideration of other nonparametric methods such as local polynomial quantile regression in the case of not clear patterns. Also, when the conditional densities of the response are heterogeneous, it would be useful to consider weighted quantile regression for the assisted estimator. Theoretical properties of the QREG estimator such as asymptotic unbiasedness, consistency, sufficiency need to be verified. Wang and Opsomer (2011) study the theoretical properties of survey estimators of quantile populations that consider non-differentiable functions of estimated quantities.

References

- Breidt, F. and Opsomer, J. D. (2000). *Local polynomial regression estimators in survey sampling*, Annals of Statistics,**28**, 1026-1053.
- Cassel, C.M. and Särndal, C.E. and Wretman, J. (1976). *Some results on generalized difference estimation and generalized regression estimation for finite populations*, Biometrika,**63**, 615-620.
- Chambers, R. (1986). *Outlier robust finite population estimators*, Journal of the American Statistical Association, **81**, 1063-1069.
- Chambers, R. (1986). Winsorization for identifying and treating outliers in Business Surveys, *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 717-726.
- Deville, J.C. and Särndal,C.E. (1992). *Calibration estimators in survey sampling*, Journal of the American Statistical Association, **87**, 376-382.
- Gutierrez, H. A. and Breidt, F. (2009). *Estimation of the population total using the generalized difference estimator and Wilcoxon ranks*, Revista Colombiana de Estadística,**32**, 123-143.
- Hao L. and Naiman D. Q.(2007). *Quantile Regression*, Sage Publications, Thousand Oaks.
- Koenker, R. and Bassett, G. (1978). *Quantile regression*, Econometrica, **46**, 33-50
- Koenker, R. and D'Orey, V.(1987, 1994). *Computing regression quantiles*, Applied Statistics, **36**,383-393, and **43**, 410-414.
- Koenker, R.(2005). *Quantile Regression*, Cambridge University Press, New York.
- Lee, H.(1995). Outliers in Business Surveys in B. Cox, D.A Binder, Christianson A., Colledge M.J., Kott P.S. (Eds) *Business Survey Methods*, Chapter 26. John Wiley Sons. New York, USA, pp 503-526.
- Montanari, G. and Ranalli, G.(2005). *Nonparametric model calibration estimation in survey sampling*, Journal of the American Statistical Association,**100**, 100(472), 1429-1442
- Mora, H.(2005). *Métodos numéricos para la estimación de parámetros en regresión cuantílica*, Revista Colombiana de Estadística, **28**, 221-231
- Pfeffermann, D. and Rao, C.R. (2011). *Sample Surveys: Inference and Analysis*, Handbook of Statistics 29B, Elsevier, Oxford, UK.
- Särndal, K.E and Swensson, B. and Wretman. B.(1992). *Model Assisted Survey Sampling*, Springer Verlag, New York, USA.
- Wang, J.C. and Opsomer, J.D (2011). *On the asymptotic normality and variance estimation for nondifferentiable survey estimators*, Biometrika,**98**, 91-106