Quantile Model-Assisted Estimation Approach for the Estimation of a Population Total (Estimación de un total poblacional asistida por modelos de regresión cuantílica)

José Fernando Zea Castro
Statistician
Code: 196971

Universidad Nacional de Colombia
Faculty of Science
Department of Statistics
Bogotá, D.C.
November 25, 2011

Quantile Model-Assisted Estimation Approach for the Estimation of a Population Total (Estimación de un total poblacional asistida por modelos de regresión cuantílica)

José Fernando Zea Castro
Statistician
Code: 196971

1. Work submitted to Obtain the Degree of
Master of Science in Statistics

Supervisor
Leonardo Trujillo Oyola
PhD in Social Statistics

Universidad Nacional de Colombia
Faculty of Science
Department of Statistics
Bogotá, D.C.
November 25, 2011

Title in English

Quantile Model-Assisted Estimation Approach for the Estimation of a Population Total

Título en español

Estimación de un total poblacional asistida por modelos de regresión cuantílica

Abstract: A quantile model-assisted approach is used to estimate a finite population total. This estimator attempts to make efficient use of auxiliary information under the presence of influential points. The approach consists in minimizing a weighted sum of the distances between fitted and observed values. Firstly, an estimator for population coefficients of a quantile regression is obtained, then a GREG-like estimator for the total is presented. The performance of the proposed estimator is assessed empirically via simulation studies under scenarios such as: different distribution of the errors and distinct settings of extreme observations with a single auxiliary variable. The proposed estimator for the finite population total seems to have a good performance in terms of smaller bias and mean square error specially in skewed distribution, normal mixtures and in the presence of influential points.

Keywords: Quantile Regression, Generalized Regression Estimation,Model-Assisted Survey Sampling

Acceptation Note

Thesis Work

" mention"

Jury

Jury

Jury

Supervisor

Cosupervisor

Bogotá, D.C., November 2011

Dedicated to:

This thesis is dedicated to my mother who taught me the value of perseverance, has supported me in every moment of my life and for her unconditional love.

Acknowledgements

The author expresses his gratitude to his supervisor, Leonardo Trujillo who has offered their knowledge, experience and intuition, Thanks to my friends Mario Saavedra and Pedro Del Campo who helped to look through the thesis in the last stages, thanks to Andres Gutierrez who advices and work in the survey sampling area was of great help to complete this work. Free software developers deserve a sincere acknowledgment because they have made their indirect contribution to the development of this thesis (R project, Quantreg package, LaTeX and many others). Thanks to Camilo Cubides for his valuable thesis template. Finally thanks to God who has encouraged me during this thesis.

Contents

List of Tables

List of Figures

Introduction

The use of available 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 named as model-assisted estimators, they usually have better performance in terms of variance.

This include GREG estimators, see for example (Pfeffermann and Rao, 2011), calibration estimators (Deville and Sarndal, 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 or not the assumptions of the models hold.

A good model fit is required in order to achieve an efficient use of the auxiliary information (Sarndal et al., 1992). The existence of extreme observations such as influential points and/or outliers implies that the use of simple linear regression could be inappropriate to assist the estimation of a population total, thus the use of robust regression techniques appears as a feasible alternative. Authors that have already worked in the problem of dealing with outliers in survey data are (Chambers, 1986),(Lee, 1992) and (Chambers, 2000).

A model-assisted quantile regression estimator for the total (QREG) is proposed and evaluated. It is shown that the use of this proposed estimator reduces the effect of influential points in the estimation of population total and gives a smaller variance than the obtained with GREG estimator, aditionally it shows better performance when the normality assumption does not hold. The rapid increase in computational capabilities nowadays makes easy to implement this estimator in practice.

The aim of this work is to to adapt the weighted quantile regression to built a model-assisted quantile regression estimator of the total for a finite population. Accuracy and efficiency measures such as bias and mean square error of the proposed model are assesed through simulations in different scenarios. A proposed variance estimator is evaluated

The document is structured as follows: in the first section the objetives of this thesis are shown, in the second section a basic review of sampling theory is exposed, in the third section the essentials of quantile regression are presented, the in fourth section is proposed an estimator of quantile regression parameters based on survey data, in the fifth section an estimator of the population total is presented using quantile regression (QREG),in the sixth section the performance of the estimators is evaluated empirically through a

simulation study, in the seventh section a practical application is developed, finally some conclusions and proposals for future research are shown.

CHAPTER 1

Objetives

1.1 General Objetive

To present a new estimator for the population total using the design-based approach and the model-assisted survey through a quantile regression model.

1.2 Specific Objetives

- To obtain the total estimator through the minimization of the objective function of the quantile regression model and the use of the generalized difference estimator.
- To propose and to evaluate a variance estimation method..
- To study the estimator under different scenarios through efficiency measures such as relative bias, mean square error and relative efficiency of the estimator.

CHAPTER 2

Design-Based Estimator for the Total in Survey Sampling

Let us consider a finite population U with a set of N elements $u_1, ..., u_k, ..., u_N$, let y denote a variable and y_k the fixed value of y for the k-th element. In survey sampling the aim is to estimate different functions of the y_k's, $\theta = f(y_1, y_2, ...y_N)$ called population parameters. The total is a parameter of great interest in practice and it plays an important role in the estimation of parameters such as ratios, means, proportions, quantiles and regression parameters.

$$t_y = \sum_U y_k \tag{2.1}$$

For estimating a parameter we observe y_k in a sample s, $(s \subseteq U)$, when $s = U$ a census is made, nevertheless, in some cases censuses are expensive and impractical, then a sample s is drawn from U according to a probability sampling design $p(.)$, the sampling size $n(s)$ can be random or fixed $(n(s) = n)$.

Let be Ψ the set of all samples s and $p(s) = Pr(S = s)$ the probability of drawning the sample s under a given selection scheme. $p(s)$ defines, in a unique way a sampling design and it determines the essential properties of functions of the random set S named statistics which are often used to estimate population parameters.

The computation of $p(s)$ for all s in Ψ are cumbersome or even impossible. It is mainly a theoretical tool and it is not useful in practice for the estimation of a population parameter, To determine the efficiency measures (Mean Square Error, Relative Bias) of an estimator first and second inclusion probabilities are required- The probability that element k will be included in the sample or first order inclusion probability, denoted as π_k is obtained from the given design $p(.)$ and it is defined as follows:

$$\pi_k = Pr(k \in s) = Pr(I_k = 1) = \sum_{s \ni k} p(s) \tag{2.2}$$

Similarly, the inclusion probability of the k and l, second-order inclusion probabilities are given by:

$$\pi_{kl} = Pr(k, l \in s) = Pr(I_k = 1, I_l = 1) = \sum_{s \ni k,l} p(s) \tag{2.3}$$

The knowledge of the information of first and second order inclusion probabilities which are established with the sampling design allows to estimate the population total for a study variable y using the information of the selected sample elements. See (Sarndal et al., 1992).

2.1 Horvitz-Thompson Estimator

The Horvitz-Thompson estimator (HT), (Horvitz and Thompson, 1952) is defined by:

$$\widehat{t}_{y\pi} = \sum_{s} \frac{y_k}{\pi_k} \tag{2.4}$$

The HT estimator satisfy $E_p(\widehat{t}_{y\pi}) = t$, which means that HT is an unbiased estimator for the population total.

The variance of HT estimator is given by:

$$\sum \sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{2.5}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, (Sarndal et al., 1992)

2.2 Generalized Regression Estimator

Usually, auxiliary information associated with the k-th unit ($k = 1, 2, ...N$) is available prior the sample. A $p \times 1$ vector $\mathbf{x}_k$ contains the auxiliary information. The use of auxiliary information can improve estimates of the total in terms of accuracy with respect to the HT estimator which does not use auxiliary information in the estimation step.

It is assumed that the relationship between the variable of interest and the auxiliary information vector $\boldsymbol{x}_k = (x_{k1}, ...x_{kp})'$ looks as if it had been generated by a linear regression model, called ξ such that $y_k = \boldsymbol{x}_k'\boldsymbol{\beta} + \varepsilon_k$, with $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \dots & \beta_J \end{pmatrix}'$ and the next features:

$$\begin{cases} E_\xi(y_k) = \sum_{j=1}^{J} \beta_j x_{jk} & \text{k = 1,...N} \\ V_\xi(y_k) = \sigma_k^2 & \text{k = 1,...N} \end{cases}$$

The regression coefficients can be estimated following the principles of the estimation of total functions or using the least squares approach, minimizing for $\boldsymbol{B}$ the next expression:

$$D = \sum_{U} \left(\frac{y_k - \boldsymbol{x}_k'\boldsymbol{B}}{\sigma_k^2} \right)^2 \tag{2.6}$$

It is straightforward to show that the hypothetical finite population fit of the model would result in estimating β as:

$$\boldsymbol{B} = (\boldsymbol{x}\Sigma^{-1}\boldsymbol{x}')^{-1} = \sum_U \left(\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1}\left(\frac{\boldsymbol{x}_k y_k}{\sigma_k^2}\right) \tag{2.7}$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{pmatrix}$

The sample regression coefficient can be estimated analogously solving:

$$\min_{\hat{\boldsymbol{B}}} D = \min_{\hat{\boldsymbol{B}}} \sum_s \left(\frac{y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}}}{\sigma_k^2\pi_k}\right)^2 \tag{2.8}$$

$\hat{B}$ is estimated as:

$$\hat{\boldsymbol{B}} = \sum_s \left(\frac{\boldsymbol{x}_k\boldsymbol{x}_k^{'}}{{\sigma_k}^2\pi_k}\right)^{-1}\left(\frac{\boldsymbol{x}_k y_k}{\sigma_k^2\pi_k}\right) \tag{2.9}$$

The generalized regression estimator (GREG) denoted $\hat{t}_{yr}$ is formally defined as:

$$\hat{t}_{yr} = \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j(t_{xj} - \hat{t}_{xj\pi}) \tag{2.10}$$

where $\hat{\boldsymbol{B}}$ is the vector of estimated regression coefficients, $\hat{\boldsymbol{B}} = (\hat{B}_1, \dots \hat{B}_J)$ and J is the number of auxiliary variables. The estimator is slightly biased, Taylor linearization provides an approximate variance for $\hat{\boldsymbol{B}}$, (Sarndal et al, pp 225-226)

$$AV(\hat{t}_{yr}) = \sum\sum_U \Delta_{kl}\check{E}_k\check{E}_l \tag{2.11}$$

where $\check{E}_k = y_k - \boldsymbol{x}_k'\boldsymbol{\beta}$, a variance estimator is given by:

$$V^{(1)}(\hat{t}_{yr}) = \sum\sum_U \Delta_{kl}g_{ks}\check{e}_{ks}g_{ls}\check{e}_{ls} \tag{2.12}$$

where $g_{ks} = 1 + (\sum_u \boldsymbol{x}_k - \sum_s \boldsymbol{x}_k)'(\sum_s \boldsymbol{x}_k\boldsymbol{x}_k'/\sigma_k^2\pi_k)^{-1}\boldsymbol{x}_k/\sigma_k^2$. An alternative variance estimator is obtained as:

$$V^{(2)}(\hat{t}_{yr}) = \sum\sum_U \Delta_{kl}\check{e}_k\check{e}_l \tag{2.13}$$

CHAPTER 3

Quantile Regression

3.1 Population Quantile Estimation

(Koenker, 2005) defines the τ-th quantile $Q^{(\tau)}$ as:

$$Q^{(\tau)} = \inf\{y : F(y) \geq \tau\} \tag{3.1}$$

where $F(y) = P(Y \leq y)$, $0 \leq \tau \leq 1$.

The simplest and most familiar measures used to describe a distribution are the mean for the central location, nevertheless restricting attention to the mean alone leads us to ignore other important properties that offer more insight into the distribution. For many researchers, attributes of interest often have skewed distributions, for which the mean and standard deviation are not necessarily the best measures of location and shape. To characterize the location and shape of asymmetric distributions quantiles are very useful. The p-th quantile of a distribution denoted Q^{τ} is the value of the inverse of the cumulative distribution function (CDF) at τ, that is a value of y such that $F(y) = \tau$ in other words the proportion of the population with an attribute below $Q^{(\tau)}$ is the τth quantile. (Hao and Naiman, 2007)

The above problem is equivalent to solve:

$$\min_{Q^{(\tau)}} E[d_\tau(y, Q^{(\tau)})] = \min_{Q^{(\tau)}} \left\{ (1-\tau) \int_{y<Q^{(\tau)}} |y - Q^{(\tau)}| + \tau \int_{y \geq Q^{(\tau)}} |y - Q^{(\tau)}| \right\} \tag{3.2}$$

where

$$d_\tau(y, Q^{(\tau)}) = \begin{cases} (1-\tau) \mid y - Q^{(\tau)} \mid & y < Q^{(\tau)} \\ \tau \mid y - Q^{(\tau)} \mid & y \geq Q^{(\tau)} \end{cases} \tag{3.3}$$

The value of $Q^{(\tau)}$ which minimize (3.2) is the τ quantile.

Given a random sample $\{y_1, y_2, \ldots, y_n\}$, the τth sampling quantile estimation is found through:

$$\hat{Q}(\tau) = inf\left\{y : \hat{F}(y) \geq \tau\right\} \tag{3.4}$$

where

$$\hat{F}(y) = \frac{\#(Y_i \geq y)}{n} \tag{3.5}$$

Analogous to the population CDF the empirical or sample CDF $\hat{F}$ is considered in association to a sample, the empirical CDF gives the proportion of the sample values than is less than or equal to any given y.

The τ-th sample quantile can be alternatively found solving:

$$\min_{Q^{(\tau)}} \left\{ \tau \sum_{y_k \geq Q^{(\tau)}} \mid y_k - Q^{(\tau)} \mid + (1-\tau) \sum_{y_k < Q^{(\tau)}} \mid y_k - Q^{(\tau)} \mid \right\} \tag{3.6}$$

The τth sample quantile $\hat{Q}^{(\tau)}$ is the τth quantile of the corresponding of the empirical CDF $\hat{F}$

Note that 3.6 is analogous to 3.2 where the integrals are replaced by averages of the sample data. The above expression can be seen as an estimate moments of $E[d_\tau(y, Q^{(\tau)})]$.

3.2 Quantile Regression Model

(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 than the linear regression over a wide range of error distributions and under the presence of influential points.

Let us consider $y_1, \ldots y_k, \ldots y_N$ the values of a study variable 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 model ξ such that $y_k = \boldsymbol{B}'^{(\tau)}\boldsymbol{x}_k + \varepsilon_k$, 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 in (3.6), the parameters to be estimated change from one-parameter problem $Q^{(\tau)}$ to a p-parameters problem $\boldsymbol{B}'^{(\tau)}$

When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

Parameters of quantile regression are found solving the next minimization problem:

$$\min_{\boldsymbol{B}^{(\tau)}} f(\boldsymbol{B}^{(\tau)}) = \min_{\boldsymbol{B}^{(\tau)}} \tau \sum_{y_k \geq \boldsymbol{B}'^{(\tau)}\boldsymbol{x}_k} \mid y_k - \boldsymbol{B}'^{(\tau)}\boldsymbol{x}_k \mid + (1-\tau) \sum_{y_k < \boldsymbol{B}'^{(\tau)}\boldsymbol{x}_k} \mid y_k - \boldsymbol{B}'^{(\tau)}\boldsymbol{x}_k \mid \tag{3.7}$$

The former optimization problem is solved through linear programming techniques, some details to found quantile regression parameters are developed below:

(3.7) can be written as follows:

$$\min_{\boldsymbol{B'(\tau)}} f(\boldsymbol{B'}^{(\tau)}) = \min_{\boldsymbol{B'}^{(\tau)}} \sum_{k=1}^{N} \rho_\tau(\varepsilon_k) \tag{3.8}$$

where $\varepsilon_k = y_k - \boldsymbol{B'}^{\tau} \boldsymbol{x}_k$, and

$$\rho_\tau(\varepsilon) = \begin{cases} (\tau - 1)\varepsilon & \varepsilon < 0 \\ \tau\varepsilon & \varepsilon \geq 0 \end{cases} \tag{3.9}$$

This is a optimization problem without restrictions, the function to maximize is non-differentiability, therefore to solve this problem is used non-differentiable optimization. (Mora, 2005) define:

$$p^+(\varepsilon) = \begin{cases} 0 & \varepsilon < 0 \\ \varepsilon & \varepsilon \geq 0 \end{cases} \tag{3.10}$$

$$p^-(\varepsilon) = \begin{cases} -\varepsilon & \varepsilon < 0 \\ 0 & \varepsilon \geq 0 \end{cases} \tag{3.11}$$

Thus (3.9) can be written as:

$$\rho_\tau(\varepsilon) = \tau p^+(\varepsilon) + (1 - \tau)p^- \tag{3.12}$$

(3.7) is reexpresed as:

$$\min_{\boldsymbol{B'}^{(\tau)}} f(\boldsymbol{B'}^{(\tau)}) = \min_{\boldsymbol{B'}^{(\tau)}} \sum_{k=1}^{N} \tau p^+(\varepsilon_k) + (1 - \tau)p^-(\varepsilon_k) \tag{3.13}$$

Let $u_k = p^+(\varepsilon_k)$, $v_k = p^-(\varepsilon_k)$, note that $u_k - v_k = e_k$, $u_k \geq 0$,$v_k \geq 0$ for $k = 1, 2, \cdots N$

(3.13) is rewritten as:

$$\min_{(u,v,\boldsymbol{B'}^{(\tau)})} f(\boldsymbol{B'}^{(\tau)}) = \min_{(u,v,\boldsymbol{B'}^{(\tau)})} \sum_{k=1}^{N} \tau u_k + (1 - \tau)v_k \tag{3.14}$$

subject to:

$$u_k - v_k = \boldsymbol{B'}^{(\tau)} x_k,\ k = 1, 2, \ldots, N$$

$$u_k \geq 0, v_k \geq 0$$

The former linear programming problem with $2N + p$ equalities is expressed in matrix form as:

$$\min_{(u,v,\boldsymbol{B'}^{(\tau)})} f(\boldsymbol{B'}^{(\tau)}) = \min_{(u,v,\boldsymbol{B'}^{(\tau)})} \left[\begin{array}{ccc} \mathbf{0}_p & \tau \mathbf{1}_N^{'} & (1-\tau)\mathbf{1}_N^{'} \end{array} \right] \left[\begin{array}{c} \boldsymbol{B'^{(\tau)}} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{array} \right] \tag{3.15}$$

subject to:

$$\left[\begin{array}{ccc} \boldsymbol{X} & I_N & (1-\tau)\mathbf{1}_N^{'} \end{array} \right] \left[\begin{array}{c} \boldsymbol{B'^{(\tau)}} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{array} \right]$$

To solve the above linear programming problem several methods exists such as the modified algorithm of Barrodale and Roberts *BB method* for l1 regression which is described in detail in (Koenker and D'Orey, (1987, 1994)), this algorithm is implemented in quantreg package, (Koenker, 2011). More details of the estimation methods for quantile regression parameters can be found in (Mora, 2005) and (Hao and Naiman, 2007).

Additionally advanced techniques in quantile regression including weighted quantile regression, which it 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:

$$\min_{\boldsymbol{B}^{(\tau)}} f(\boldsymbol{B}^{(\tau)}) = \min_{\boldsymbol{B}^{(\tau)}} \ \tau \sum_{y_k \geq \boldsymbol{B'}^{(\tau)} \boldsymbol{x}_k} w_k \mid y_k - \boldsymbol{B'}^{(\tau)} \boldsymbol{x}_k \mid + (1-\tau) \sum_{y_k < \boldsymbol{B'}^{(\tau)} \boldsymbol{x}_k} w_k \mid y_k - \boldsymbol{B'}^{(\tau)} \boldsymbol{x}_k \mid \tag{3.16}$$

CHAPTER 4

Quantile Model-Assisted Estimation Approach for Survey Data

4.1 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. The quantile regression parameters for data coming from a sample survey s can be estimated using a median regression ($\tau = 0.5$) and solving:

$$\min_{\hat{\boldsymbol{B}}^{(\tau)}} f(\hat{\boldsymbol{B}}^{(\boldsymbol{\tau})}) = \min_{\hat{\boldsymbol{B}}^{(\tau)}} \tau \sum_{s \cap \left\{y_k \geq \hat{\boldsymbol{B}}'^{(\boldsymbol{\tau})} \boldsymbol{x}_k\right\}} \frac{\mid y_k - \hat{\boldsymbol{B}}'^{(\boldsymbol{\tau})} \boldsymbol{x}_k \mid}{\pi_k} + (1-\tau) \sum_{s \cap \left\{y_k < \hat{\boldsymbol{B}}'^{(\tau)} \boldsymbol{x}_k\right\}} \frac{\mid y_k - \hat{\boldsymbol{B}}'^{(\tau)} \boldsymbol{x}_k \mid}{\pi_k} \tag{4.1}$$

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. In this thesis we use $\tau = 0.5$ to build an assisted estimator for the total.

4.2 Quantile Regression Based Estimator for the Population Total

4.2.1 Total Estimation

The quantile regression based estimator (QREG) 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 assumed such as

$$y_k = \boldsymbol{x}_k' \boldsymbol{B}^{(\tau)} + E_k \tag{4.2}$$

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:

$$t_y = \sum_U y_k^0 + E_k = \sum_U \boldsymbol{x}_k' \boldsymbol{B}^{(\tau)} + \sum_U (y_k - \boldsymbol{x}_k' \boldsymbol{B}^{(\tau)}) \tag{4.3}$$

The difference estimator for the total, (Sarndal et. al, pp. 222) is defined by:

$$\hat{t}_{dif} = \sum_U \boldsymbol{x}_k' \boldsymbol{B}^{(\tau)} + \sum_s \frac{y_k - \boldsymbol{x}_k' \boldsymbol{B}^{(\tau)}}{\pi_k} = \hat{t}_{y\pi} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x}})' \boldsymbol{B}^{(\tau)} \tag{4.4}$$

We propose to estimate the unknown regression parameter $\boldsymbol{B}^{(\tau)}$ in 4.4 through a quantile regression model. We will denote this estimator as $\hat{t}_{yqr}$, the quantile regression estimator (QREG) which is obtained considering a median regression ($\tau = 0.5$). In other words:

$$\hat{t}_{yqr} = \sum_U \boldsymbol{x}_k' \hat{\boldsymbol{B}}^{(\tau)} + \sum_s \frac{y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}^{(\tau)}}{\pi_k} = \hat{t}_{y\pi} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x}})' \hat{\boldsymbol{B}}^{(\tau)} \tag{4.5}$$

Then the QREG estimator can be expressed finally as:

$$\hat{t}_{yqr} = \hat{t}_{y\pi} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{x}})' \hat{\boldsymbol{B}}^{(0.5)} \tag{4.6}$$

Other alternative to write 4.5 is:

$$\hat{t}_{yqr} = \sum_s \frac{e_k}{\pi_k} + \left(\sum_U \boldsymbol{x_k'} \right) \hat{\boldsymbol{B}}^{(0.5)} \tag{4.7}$$

where $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}^{(\tau)}$

4.2.2 Variance Estimation

The exact design-based variance of the QREG does not have a closed form, therefore an analytical expression for estimating the variance of the estimator can not be obtained. (Wolter, 2007) claims that the sample median is not smooth enough, therefore a jackknife variance estimation cannot yield a consistent estimator, a quantile regression parameters are a generalization of a median,therefore is not recommended to use jackknife variance estimation, instead, bootstrap variance estimation or an alternative estimator should be used. Follow the procedure suggested by (Sarndal et al., 1992) for use bootstrap for variance estimation, an artificial population U^* is built using the sample data, U^* mimics the real and unknown population, a series of independent samples (bootstrap sampling) are drawn from U^* by the same design to the one by which s was drawn from U, for each bootstrap sampling and estimate $\hat{\theta}_a^*$ is calculated in the same way as $\hat{\theta}$, $(a = 1, ...A)$ was calculated, $V(\hat{\theta})$ is estimated by:

$$\hat{V}_{BS} = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{\theta}_a^* - \hat{\theta}^*)^2 \tag{4.8}$$

where

$$\hat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\theta_a^* \tag{4.9}$$

A natural way of estimate the variance of the QREG estimator is compute the variance of $\sum_s \frac{E_k}{\pi_k} = \sum_s \frac{y_k - \boldsymbol{x}_k'\boldsymbol{B}^{(\tau)}}{\pi_k}$ in (4.4).

The HT variance of this expression is:

$$V(t_{yqr}) = \sum\sum_{U} \Delta_{kl}\frac{E_k E_l}{\pi_k \pi_l} \tag{4.10}$$

where $E_k = y_k - B_0^{(0.5)} - B_1^{(0.5)}x_k$.

An ubised estimator for 4.10 is given by:

$$\hat{V}(t_{yqr}) = \sum\sum_{s} \check{\Delta}_{kl}\frac{\hat{e}_k \hat{e}_l}{\pi_k \pi_l} \tag{4.11}$$

where $e_k = y_k - \hat{B}_0^{(0.5)} - \hat{B}_1^{(0.5)}x_k$

Some concepts such as convergence in distribution to a constant, mean square convergence are required to analyze the theoretical properties of the previous variance estimator, the assessment of these properties for the proposed estimator is out of the scope of this thesis. Some simulations are made to show that QREG estimator is reasonable well in terms of variance.

A finite population of size $N = 1000$ was simulated from a superpopulation model, $y_k = 10 + 2x_k + \varepsilon_k$, the linear model has correctly specified variance structure with normal, uncorrelated and homoscedastic errors.

In each runs of the simulation ($M = 5000$), simple random sample were drawn with $n = 100$. The parameters $B_0^{(0.5)}$ and $B_1^{(0.5)}$ and t_{yqr} were estimated as in (4.1) and (4.6).

The performance of the QREG estimator was evaluated computing the relative bias (RB), It was obtained from the simulation: $RB = -0.647\%$, which is very close to zero, therefore the use of the estimator is validated and justified even under a scenarios where GREG estimator is preferred.

4.2.3 Representative Strategies

Definition 4.2.1. *Let $\boldsymbol{x}$ be an auxiliary information vector, a sampling strategy is called* ***representive*** *with respect to $\boldsymbol{x}$, if and only if*

$$\hat{t}(\boldsymbol{x}) = t_{\boldsymbol{x}} \tag{4.12}$$

or, if the estimator applied to the auxiliary variables reproduce exactly the population total of each auxiliary variable. (Gutierrez and Breidt, 2009)

Result 4.2.1. *Under any sampling p(s), the proposed population total estimator induces a representative strategy at least for a one solution of the quantile regression minimization problem, for that solution the pair $(\hat{t}_{yqr}, p(s))$ estimates the population total of the auxiliary variables with null variance.*

Proof.

To compute the quantile regression parameters the next minimization problem must be solved:

$$\min_{\hat{\boldsymbol{B}}^{(0.5)}} f(\hat{\boldsymbol{B}}^{(0.5)}) = \min_{\hat{\boldsymbol{B}}^{(0.5)}} \sum_s \frac{\mid x_k - \hat{B_0}^{(0.5)} + \hat{B_1}^{(0.5)} x_k \mid}{\pi_k} \tag{4.13}$$

Note that all the summands on the right side of the equality, $\mid x_k - \hat{B_0}^{(0.5)} + \hat{B_1}^{(0.5)} \mid$ are greater or equal to zero, It is straightforward to show that $f(\hat{\boldsymbol{B}}^{(0.5)})$ is minimized when $\mid x_k - \hat{B_0}^{(0.5)} + \hat{B_1}^{(0.5)} \mid = 0$ for all $k = 1, ...n$ which occurs when $\hat{B_0}^{(0.5)} = 0$ and $\hat{B_1}^{(0.5)} = 1$

Therefore

$$\begin{aligned}
\hat{t}_x &= \sum_U \hat{B}_0^{(0.5)} + \hat{B}_1^{(0.5)} x_k + \frac{\sum_s x_k - \hat{B}_0^{(0.5)} - \hat{B}_1^{(0.5)} x_k}{\pi_k} \\
&= \sum_U 0 + x_k + \frac{\sum_s x_k - 0 - x_k}{\pi_k} \\
&= x_k
\end{aligned}$$

4.2.4 Cochran Consistency

Definition 4.2.1. *An estimator is consistent for a finite population for a finite population under a given class of design if $s = U$ implies that $\hat{\theta} = \theta$. (Sarndal et al., 1992)*

Result 4.2.1. *Under SI designs quantile regression based estimator is Cochran-consistent.*

Proof

if $s = U$ and is used SI design then $\hat{t}_{x\pi} = t_{x\pi}$ and $\hat{t}_{y\pi} = t_{y\pi}$ and,

[0]The combination $(p,\hat{t})$ denoting an estimator $\hat{t}$ based on sample drawn accordingly to a design p is called a strategy

$$\begin{aligned}\hat{t}_{yqr} &= \hat{t}_{y\pi} + (\boldsymbol{t}_{\boldsymbol{x}} - \boldsymbol{t}_{\boldsymbol{x}\pi})'\hat{B}^{(0.5)} \\ &= t_y + (\boldsymbol{t}_{\boldsymbol{x}} - \boldsymbol{t}_x)'\hat{B}^{(0.5)} \\ &= t_y\end{aligned}$$

CHAPTER 5

Empirical Simulation

5.1 Monte Carlo Simulation

(Sarndal et al, pp 221) explains that "Monte Carlo simulation is often used when an exact description of the sampling distribution of a given estimator is difficult to obtain. To obtain the exact distribution, one would have to consider all samples s that are possible under the given design. For every s, one must know the probability $p(s)$ of drawing (s) and the value of the estimator of a parameter of interest, $\hat{\theta} = \hat{\theta}(s)$. It would be then be possible to calculate the exact values of the bias and the variance of $\hat{t}$. Nevertheless, this is ordinaly an impossible undertaking, since the number of possible samples is much too large. This is why Monte Carlo simulation is frequently used to study the statistical properties of estimators in surveys"

A Monte Carlo simulation is typically carried out as follows. The finite population and the design are held fixed. A large number of samples is drawn from the given population according to the given design. Once drawn, a sample is replaced before the one is drawn, so that it is always the sample population that is sampled. The number of sample is denoted K. For every realized sampled sample, the estimate of the parameter of interest $\hat{\theta}$ and the variance estimate $\hat{V}(\theta)$ are calculated. If K is large enough, the distribution of the K estimates, named empirical sampling distribution, will closely approximate the exact sampling distribution. If we let $\hat{\theta}_j$ and $\hat{V}(\hat{\theta})_j$ denote the results obtained for the j-th sample, we can calculate

$$\overline{\hat{\theta}} = \frac{1}{K}\sum_{j=1}^{K}\hat{\theta}_j \tag{5.1}$$

which is an estimate of the expected value $E(\hat{\theta})$,

$$S^2_{\hat{\theta}} = \frac{1}{K-1}\sum_{j=1}^{K}(\hat{\theta}_j - \overline{\hat{\theta}})^2 \tag{5.2}$$

which is an estimate of the variance $V(\hat{\theta})$, and finally

$$\overline{\hat{V}} = \frac{1}{K}\sum_{j=1}^{K}\hat{V}(\hat{t})_j \tag{5.3}$$

The above procedure is mentioned throughout the document as the conventional Monte Carlo 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 such as exponential, normal mixture and lognornal distribution and under the presence of influential points.

A Monte Carlo simulation was carried out considering $M = 5,000$ [1] repeated SI samples with $n = 100$ from a population of $N = 1,000$ elements. A 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 exponential distribution with parameter equal to 1. The ε_k were assumed independent and normally distributed. In each sample, three estimators: the HT estimator $\hat{t}_\pi$, the GREG estimator $\hat{t}_{ygreg}$ and the QREG estimator $\hat{t}_{yqr}$ were calculated. The simulations were carried out by using the statistical software R 2.12.2 (R Development Core Team, 2010). The algorithms are available under request to the authors.

The performance of the total estimators is evaluated in terms of their Relative Bias (RB) and their Mean Square Error (MSE).

$$RB(\hat{t}) = \frac{100}{M}\sum_{m=1}^{M}\frac{\hat{t}_m - t}{t} \tag{5.4}$$

$$MSE(\hat{t}) = \frac{1}{M}\sum_{m=1}^{M}(\hat{t}_m - t)^2 \tag{5.5}$$

where t_m is computed in the m-th simulated sample.

For purpose of comparison between an estimator $\hat{t}_y$ with the QREG estimator,we used the relative efficiency (RE) defined by:

$$\boldsymbol{RE}(\hat{t_y}, \hat{t}_{yqr}) = \frac{MSE(\hat{t}_y)}{MSE(\hat{t}_{yqr})} \tag{5.6}$$

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 thesis.

[1] This number of runs provide stable predictions in the performance measures of the estimator

In scenarios where only a fraction of the simulated population data is contaminated by means of influential points, a modified version of the Monte Carlo Simulation is used. The procedure is performed with the following steps: firstly, extreme points are identified through traditional influence measures, (Chatterjee and Hadi, 1986). In particular, the DFFITS has a good performance at identifying influential points in the population (DFFITS identified 100% of the influential points correctly). Secondly, the simulation is carrying on by keeping the same proportion of contaminated data than in the simulated population. Third, in each sample the total is estimated through the π, GREG and QREG estimators. Finally, some performance measures are computed such as the Relative Bias (RB) and the Mean Square Error (MSE) in order to guarantee that every sample mimics the characteristics of the simulated population.

In the next section, the conventional Monte Carlo simulation it is used in scenarios with different configurations in the distribution (normal, exponential, lognormal and mixed normal distributions) and other assumptions such as homoscedastic and uncorrelated errors. Influence measures are used to identify influence points. Additionally, scenarios with contaminated data are generated explicitly with different fraction of contaminated data. In the scenarios with the presence of contaminated data, both the modified Monte Carlo simulation and conventional Monte Carlo simulation are used.

The aim of this research was to obtain a more efficient estimator under highly skewed distributions and the presence of extreme observations. The proposed estimator (QREG) is build using model assisted approach through quantile regression.

5.2 Scenarios with different distribution and specification of the errors

The scenarios with different distributions were simulated as follows:

- M_1: *Linear model with correctly specified variance structure, normal, uncorrelated and homoscedastic errors* ($E_\xi(y_k) = 10 + 2x_k$ and $V_\xi(y_k) = 1$).
- M_{2a}: *Linear model with correctly specified variance structure, normal, uncorrelated and heteroscedastic errors.* The model ξ has the same $E_\xi(y_k)$ than M_1 but $V_\xi(y_k) = \sqrt{x_k}$.
- M_{2b}: *Linear model with incorrectly specified variance structure, normal, uncorrelated and heteroscedastic errors.* The model is the same as M_2 but the model is incorrectly specified for the GREG estimator ignoring the variance structure.
- M_3: *Linear model with correctly specified variance structure and non-normal (exponential), uncorrelated and homoscedastic errors.* This model assumes an exponential distribution with parameter equal to one.
- M_4: *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.
- M_5: *Linear model with five percent of contaminated data with lognormal distributions, uncorrelated and homoscedastic errors.*

Table 5.1 reports the simulated relative bias (RB) and mean square error (MSE) for the estimators. In the three estimator the relative bias is negligible.

Under the scenario M_1, the proposed estimator features in a similar way than the GREG estimator. When the assumption of homoscedasticity is not satisfied the performance of the proposed estimator is similar to the GREG estimator in terms of bias and variance. It does not seem that it is relevant that the model is is either incorrectly specified (M_2a) or not (M_2b). A similar situation occurs under the scenario M_3 where the errors are exponential, although in this cases the performance of the QREG estimator is slightly better than the obtained for the GREG estimator in terms of its MSE.

In the case of the scenario M_4, where the errors are contaminated with a mixture of normals; the proposed estimator has a slightly lower performance in the classical Monte Carlo simulation. Nevertheless, in the modified Monte Carlo simulation where the proportion of contaminated data is fixed for each sample, the proposed estimator shows a good performance in terms of its MSE. In this scenario, the influential observations were previously identified in the population. The scenario M_5 where the errors are lognormal features well in terms of its MSE although in the modified Monte Carlo Simulation improves its performance slightly.

Scenario	MC version	Measure	$\hat{t}_{ypi}$	$\hat{t}_{ygreg}$	$\hat{t}_{yqr}$
M_1	Conventional	RB	0.00240	0.00037	0.00021
		MSE	(311044)	(9293)	(9388)
M_{2a}	Conventional	RB	0.00239	0.00021	-0.00018
		MSE	(340205)	(35804)	(36160)
M_{2b}	Conventional	RB	0.00239	0.00011	-0.00018
		MSE	(340205)	(35791)	(36160)
M_3	Conventional	RB	-0.00185	-0.00477	-0.00559
		MSE	(310422)	(9326)	(9256)
M_4	Conventional	RB	0.00098	-0.01033	-0.00366
		MSE	(361990)	(18511)	(19440)
	Modified	RB	-0.00840	-0.00192	-0.00308
		MSE	(273594)	(10724)	(9819)
M_5	Conventional	RB	-0.00625	-0.00853	-0.00756
		MSE	(334380)	(33846)	(33613)
	Modified	RB	-0.04932	-0.04680	-0.04597
		MSE	(327441)	(19750)	(19504)

TABLE 5.1. RB and MSE of the estimators under different specifications of the errors

Table 5.2 shows the relative efficiency of the HT and GREG estimators in comparison with the proposed estimator, QREG as defined in 5.6 The proposed estimator behaves well in the classic linear model and under the violation of the homoscedasticity assumption. The behavior of the QREG under a scenario with exponential errors is slightly better than the GREG estimator. Respect to the model with a mixture of normal errors or with lognormal errors (M_4 and M_5, respectively), the QREG estimator has a better performance than the other two estimators confirming what was obtained in Table 5.1 above. Under the classic Monte Carlo Simulation, both estimators QREG and GREG estimators show a similar efficiency in terms of MSE.

Scenarios	MC version	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$
M_1	Conventional	33.13	0.99
M_{2a}	Conventional	9.41	0.99
M_{2b}	Conventional	9.41	0.99
M_3	Conventional	33.54	1.01
M_4	Conventional	18.62	0.95
	Modified	27.86	1.09
M_5	Conventional	9.95	1.01
	Modified	16.79	1.01

TABLE 5.2. Relative Efficiency of the QREG Estimator under different specifications of the errors.

5.3 Scenarios with presence of influential points in the assisted model for the survey sampling estimator of the total

The occurrence of large units is common in population of businesses. If auxiliary information is available to identify these units, their impact on the survey estimates can be minimized on the survey design stage. This is typically accomplished by stratifying according to a size measure and by putting all the large units in a take-all stratum in which selection is done with certainty. However, some large units may still be unexpectedly selected in the sample due to imperfect auxiliary information at the time of stratification. (Pfeffermann and Rao, 2011)

In order to take into account what occurs in practice, some additional scenarios are considered with different configurations of extreme observations. Firstly, a linear model with correctly specified variance structure, normal, uncorrelated and homoscedastic errors, $E_\xi(y_k) = 10 + 2x_k$ and $V_\xi(y_k) = 1$ is simulated.

New values for x_k and y_k are generated as:

$$x_k^* = \begin{cases} x_k & \text{1-c \% of no contaminated points} \\ \delta_k & \text{c \% of contaminated points} \end{cases}, \; y_k^* = \begin{cases} y_k & \text{1-c \% of no contaminated points} \\ \omega_k & \text{c \% of contaminated points} \end{cases} \tag{5.7}$$

with δ_k and ω_k defined for every particular case below. The values considered for c in this simulation were 1, 3 and 5.

Quantile regression models are vey appropriate under the presence of influential points as it is mention previously. In the next scenarios models with a fraction of influential points are analyzed, other features such as the distance of the influential points to the mean of every axis and the location of the influential points in the plane are also considered. Influential points are generated by simulating outliers either in the x-axis, y-axis or both. Some of the corresponding scatterplots of the simulated scenarios are illustrated in appendix C. The simulation was as follows:

1. *Linear model with normal residuals and the presence of* 1%, 3% *and* 5% *outlier points in the x-axis.* The corresponding fraction of outlier points over the $N = 1000$

observations were randomly contaminated increasing their values in the x-axis. The values in the y-axis were simulated on the range of the original data according to (5.10). Four different scenarios under this particular model were analyzed for different percent of contamination and configurations of δ_k and ω_k as follows:

- $M_{1a} : \delta_k \sim U(\overline{x} + 3s_x, \overline{x} + 4s_x)$ and bottom y-axis: $\omega_k \sim U(\min(y), Q^{(0.05)}(y))$.
- $M_{1b} : \delta_k$ simulated as in M_{1a} and bottom y-axis: $\omega_k \sim U(Q^{(0.95)}(y), \max(y))$.
- $M_{1c} : \delta_k \sim U(\overline{x} + 5s_x, \overline{x} + 6s_x)$ and top y-axis ω_k simulated as M_{1a}.
- $M_{1d} : \delta_k$ simulated as M_{1c} and top y-axis: ω_k simulated as M_{1b}.

Where $\overline{x}$ and s_x are the mean and the standard deviation of the original x simulated values.

2. *Linear models with normal residuals and the presence of* 1%, 3% *and* 5% *outlier points in the y-axis*: The corresponding fraction of outlier points over the $N = 1000$ observations were randomly contaminated increasing their values in the y-axis. The values in the x-axis were simulated on the range of the original data according to (5.7). Four different scenarios under this particular model were analyzed for different configurations of δ_k and ω_k as follows:

   - $M_{2a} : \delta_k \sim U(\min(x), Q^{(0.05)}(x))$ and $\omega_k \sim U(\overline{y} + 3s, \overline{y} + 4s)$.
   - $M_{2b} : \delta_k \sim U(Q^{(0.95)}(y), \max(x))$ and ω_k simulated as M_{2a}.
   - $M_{2c} : \delta_k$ simulated as in M_{2a} and $\omega_k \sim U(\overline{y} + 5s, \overline{y} + 6s)$.
   - $M_{2d} : \delta_k$ simulated as M_{2b} and ω_k simulated as in M_{2c}.

   Where $\overline{y}$ and s_y are the mean and the standard deviation of the original y simulated values.

3. *Linear models with normal residuals and the presence of* 1%, 3% *and* 5% *outlier points in both the y-axis and the x-axis.* The corresponding fraction of outlier points over the $N = 1000$ observations were randomly contaminated increasing their values both in 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:

   Where x and y values are simulated as in scenarios with presence of outliers (1) and (2).

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

Table 5.3 reports the RB and MSE for the considered estimators. In all the cases, the RB of each estimator is negligible and in many cases the RB of the QREG is the lowest between the estimators considered. The MSE of the QREG and GREG estimator increases as the proportion of the influential points does (in this simulation, from 1% to 5%). However, the MSE of the QREG estimator increases more slowly than the GREG estimator does. Also, the QREG estimator is robust to the distance of the outliers to the center of the distribution in terms to MSE. In contrast, the GREG estimator is influenced for the distance that the outliers are located to the center of the distribution.

Scenarios		1%			3%			5%		
Type	Degree	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	$\hat{t}_{yqr}$	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	$\hat{t}_{yqr}$	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	$\hat{t}_{yqr}$
x-axis	M_{1a}	-0.027 (296,108)	-0.008 (17,172)	-0.002 (9,519)	-0.014 (286,872)	-0.01 (57,790)	-0.003 (10,527)	0.022 (300,947)	0.01 (111,874)	-0.006 (10,904)
	M_{1b}	-0.027 (296,108)	-0.01 (37,195)	-0.002 (9,508)	-0.014 (286,872)	-0.007 (123,970)	-0.004 (10,802)	0.022 (300,947)	0.025 (202,759)	-0.005 (11,826)
	M_{1c}	-0.027 (296,138)	-0.003 (10,154)	-0.002 (9,547)	-0.013 (286,948)	-0.005 (15,268)	-0.003 (10,623)	0.022 (301,084)	0.000 (21,155)	-0.006 (11,032)
	M_{1d}	-0.027 (296,138)	-0.006 (16,407)	-0.002 (9,536)	-0.013 (286,948)	-0.004 (39,544)	-0.004 (10,899)	0.022 (301,084)	0.01 (60,284)	-0.005 (11,954)
y-axis	M_{2a}	-0.027 (296,109)	0.002 (11,093)	-0.002 (9,299)	-0.014 (287740)	0.009 (27,243)	-0.004 (10,057)	0.025 (302,446)	0.035 (57,364)	0.002 (10,486)
	M_{2b}	-0.027 (296,109)	0.004 (13,083)	-0.002 (9,299)	-0.014 (287,740)	0.015 (43,449)	-0.004 (10,057)	0.025 (302,446)	0.048 (100,924)	0.002 (10,486)
	M_{2c}	-0.027 (296,109)	-0.003 (13,710)	-0.003 (13,214)	-0.014 (287,740)	-0.006 (22,595)	-0.004 (21,116)	0.025 (302,446)	-0.003 (30,231)	-0.001 (28,405)
	M_{2d}	-0.027 (296,109)	-0.004 (14,510)	-0.003 (13,214)	-0.014 (287,740)	-0.007 (24,809)	-0.003 (21,116)	0.025 (302,446)	-0.005 (33,832)	-0.001 (28,405)
x and y-axis	M_{3a}	-0.026 (299,140)	0 (11,852)	-0.002 (9,175)	-0.016 (296,342)	-0.003 (28,429)	-0.004 (9,629)	0.024 (322,344)	-0.017 (57,477)	-0.005 (9,686)
	M_{3b}	-0.026 (299,140)	0.002 (25,485)	-0.002 (9,240)	-0.015 (296,342)	0.000 (108,251)	-0.004 (9,928)	0.023 (322,344)	-0.026 (230,703)	-0.006 (10,653)

TABLE 5.3. RB and (MSE) of the $\hat{t}_{y\pi}$, $\hat{t}_{greg}$ and $\hat{t}_{qreg}$ estimators under different types of influential points in the assisted model (modified MC simulation

Table 5.4 confirms that QREG estimator is considerably better in terms of MSE than the GREG estimator. When contaminated points are further to the mean of the x-axis and/or y-axis, the performance of the QREG increases with respect to the GREG estimator in terms of MSE. For instance, under the scenario M_{1b}, the relative efficiency of the QREG estimator is 3.9 what indicates that the QREG estimator is approximately four times more efficient than the GREG estimator in terms of MSE. The relative efficiency of the QREG estimator decreases to 1.8 under the M_{1a} scenario when the outliers are simulated nearer to the mean of the x-axis. It is remarkable than the relative efficiency of the QREG with respect to GREG estimator increases when there is a higher proportion of contaminated data.

It is remarkable that the modified Monte Carlo Simulation is proper to simulate the set of samples with the same proportion of contaminated data that in the population. This version of Monte Carlo Simulation works better than the conventional method in scenarios with outliers because the conventional method does not reflect the presence of influential points in several of the K samples.

Table 5.5 shows the evaluation of the proposed estimator with respect to the π estimator and the GREG estimator using the conventional Monte Carlo Simulation. It is important to emphasize that this simulation does not mimic the characteristics of the simulated population, nevertheless this simulation is made in order to show that the QREG estimator does not have a considerable lose in efficiency even in the cases when the sample is not similar to the simulated population.

Using the conventional Monte Carlo simulation is observed that the relative bias of the QREG estimator is negligible in all scenarios and many cases lower that the bias of the GREG estimator.

Scenarios		1%		3%		5%	
Location	Degree	$t_{y\pi}$	t_{ygreg}	$t_{y\pi}$	t_{ygreg}	$t_{y\pi}$	t_{ygreg}
x-axis	M_{1a}	31.1	1.8	27.2	5.5	27.6	10.3
	M_{1b}	31.1	3.9	26.6	11.5	25.5	17.1
	M_{1c}	31.0	1.1	27.0	1.4	27.3	1.9
	M_{1d}	31.0	1.7	26.3	3.6	25.2	5.0
y-axis	M_{2a}	31.8	1.2	28.6	2.7	28.8	5.5
	M_{2b}	31.8	1.4	28.6	4.3	28.8	9.6
	M_{2c}	22.4	1.0	13.6	1.1	10.6	1.1
	M_{2d}	22.4	1.1	13.6	1.2	10.6	1.2
x and y-axis	M_{3a}	32.6	1.3	30.8	3.0	33.0	5.9
	M_{3b}	32.0	2.8	29.8	10.9	30.3	21.7

TABLE 5.4. Relative Efficiency of the QREG Estimator under different types of influential points in the assisted model.

The MSE of the QREG and GREG estimator increase as the proportion of contaminated data also does and in analogous way under a greater distance of the outliers to the center of the simulated data.

Scenarios		1%			3%			5%		
Type	Degree	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	t_{yqr}	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	t_{yqr}	$\hat{t}_{y\pi}$	$\hat{t}_{ygreg}$	t_{yqr}
x-axis	M_{1a}	-0.001	0.065	0.013	-0.002	0.118	-0.002	-0.01	0.137	-0.015
		(313,482)	(82,934)	(91,310)	(328,863)	(201,876)	(252,369)	(339,622)	(279,033)	(392,818)
	M_{1b}	-0.001	0.161	0.023	-0.002	0.243	0.014	-0.01	0.23	0.015
		(313,482)	(135,853)	(161,850)	(328,863)	(300,735)	(453,612)	(339,622)	(365,291)	(682,501)
	M_{1c}	-0.009	0.019	0.004	0.004	0.036	0.004	0.012	0.042	0.006
		(308,833)	(16,996)	(17,790)	(318,791)	(29,261)	(32,755)	(329,315)	(37,681)	(44993)
	M_{1d}	-0.009	0.081	0.015	0.004	0.12	0.02	0.012	0.113	0.035
		(308,833)	(42,105)	(47,949)	(318,791)	(84,218)	(115,553)	(329,315)	(101,086)	(16,1949)
y-axis	M_{2a}	-0.013	-0.029	-0.016	0.008	-0.03	0.004	0.023	-0.024	0.025
		(335,708)	(87,565)	(89,375)	(400,776)	(235,949)	(25,6091)	(465,943)	(358,493)	(403,387)
	M_{2b}	-0.018	-0.038	-0.021	0.011	-0.041	0.007	0.035	-0.032	0.036
		(388,503)	(160,419)	(163,996)	(564,605)	(446,107)	(486,173)	(732,824)	(681,850)	(771,821)
	M_{2c}	-0.013	-0.017	-0.015	0.008	0.001	0.006	0.023	0.016	0.021
		(335,708)	(59,589)	(59,322)	(400,776)	(153,596)	(154,101)	(465,943)	(230,664)	(230,616)
	M_{2d}	-0.018	-0.022	-0.02	0.011	0.002	0.01	0.035	0.025	0.032
		(388,503)	(119,936)	(119,025)	(564,605)	(333,576)	(333,276)	(732,824)	(513,769)	(511,525)
x and y-axis	M_{3a}	-0.016	-0.037	-0.013	0.012	-0.073	-0.008	0.042	-0.088	-0.003
		(412,195)	(35,730)	(37,427)	(701,842)	(100,327)	(116,900)	(1,033,804)	(155,800)	(200,225)
	M_{3b}	-0.026	-0.116	-0.024	0.02	-0.207	-0.021	0.073	-0.211	-0.019
		(631,125)	(86,600)	(98,656)	(1,596,861)	(280,829)	(402,483)	(2,796,772)	(442,469)	(783,392)

TABLE 5.5. RB and (MSE) the QREG using conventional Monte Carlo simulation

In Table 5.6 is shown that the behavior of the QREG estimator with respect to GREG estimator is similar in terms of MSE in the scenarios where there are 1% degree of contaminated data. However, when the proportion of contaminated data is increased, the GREG estimator performs better than the QREG estimator. It is important to notice that the classical Monte Carlo simulation cannot replicate the behavior of the population data in

the different samples simulated. For this reason, we prefer to get conclusions from the modified Monte Carlo simulation approach.

Scenarios		1%		3%		5%	
Location	Degree	$t_{y\pi}$	t_{ygreg}	$t_{y\pi}$	t_{ygreg}	$t_{y\pi}$	t_{ygreg}
x-axis	M_{1a}	3.43	0.91	1.30	0.80	0.86	0.71
	M_{1b}	1.94	0.84	0.72	0.66	0.50	0.54
	M_{1c}	17.36	0.96	9.73	0.89	7.32	0.84
	M_{1d}	6.44	0.88	2.76	0.73	2.03	0.62
y-axis	M_{2a}	3.76	0.98	1.56	0.92	1.16	0.89
	M_{2b}	2.37	0.98	1.16	0.92	0.95	0.88
	M_{2c}	5.66	1.00	2.60	1.00	2.02	1.00
	M_{2d}	3.26	1.01	1.69	1.00	1.43	1.00
x and y-axis	M_{3a}	11.01	0.95	6.00	0.86	5.16	0.78
	M_{3b}	6.40	0.88	3.97	0.70	3.57	0.56

TABLE 5.6. Relative Efficiency of the QREG using conventional Monte Carlo simulation

5.4 Evaluation of Parameters

In this section, the results from the Monte Carlo simulation are used in order to show that the QREG estimator of the regression coefficients have a good performance in terms of low relative bias and mean square error. A finite population of size $N = 1,000$ is simulated from a model ξ. The relationship between y_k and x_k is assumed that follows the model ξ such that $y_k = 10 + 2x + \epsilon_k$ and

$$\begin{cases} E_\xi(y_k) = 10 + 2x & \text{k = 1,...N} \\ V_\xi(y_k) = \sigma^2 & \text{k = 1,...N} \end{cases} \tag{5.8}$$

Two cases were simulated: in the first case, the values of x are generated from an exponential distribution with parameter equal to one. The second case is similar to the last one but 1% of the 1,000 points are contaminated increasing their values in the x-axis and keeping the values in the y-axis on the range of the initial simulated data. The new values x_k^* were simulated as follows:

$$x_k^* = \begin{cases} x_k & \text{99 \% of no contaminated points} \\ \delta_k & \text{1 \% of contaminated points} \end{cases} \tag{5.9}$$

$$y_k^* = \begin{cases} y_k & \text{99 \% of no contaminated points} \\ \omega_k & \text{1 \% of contaminated points} \end{cases} \tag{5.10}$$

where $\delta_k \sim U\left(\overline{x} + 5s, \overline{x} + 4s\right)$ and $\omega_k \sim U\left(Q^{0.95}(y), max(y)\right)$

The figure at the right of (A.2) shows the corresponding scatterplot for the second scenario. It is assumed that ε are independent and identically distributed as $N(0, \sigma^2)$.

In each simulation, random samples according to a SI design were drawn. Each sample was of size $n = 100$. The parameters of a linear regression were estimated using least squares and the quantile regression parameters were estimated through the modified version of the Barrodale and Roberts algorithm for l1-regression, described in detail in (Koenker and D'Orey, (1987, 1994)). This process was repeated M = 1,000 times. The simulation was written in the statistical software R-2.12.2, (R Development Core Team, 2010). The quantreg package, (Koenker, 2011), was used to estimate the quantile regression model. The performance of an estimator $\hat{B}$ was evaluated in the simulation in terms of its relative bias (RB) and its mean square error (MSE):

$$RB(\hat{B}) = \frac{100}{M} \sum_{m=1}^{M} \frac{\hat{B}_m - B}{B} \tag{5.11}$$

$$MSE(\hat{B}) = \frac{1}{M} \sum_{m=1}^{M} (\hat{B}_m - B)^2 \tag{5.12}$$

where $\hat{B}_m$ was computed in the m-th simulated sample.

TABLE 5.7. Quantile Regression Parameter

	Relative Bias				Mean Squere Error			
	Quantile Regression		Least Square		Quantile Regression		Least Square	
Parameters	b_o	b	b_o	b	b_o	b	b_o	b
Not Contaminated	0.212	-1.264	0.020	-0.204	0.031	0.018	0.019	0.010
Contaminated	-0.916	0.332	0.502	-0.478	0.596	0.003	1.065	0.005

Table (5.7) above shows the relative bias of the sampling estimators of B_0 and B_1 based in weighted quantile regression is smaller than the relative bias based in least square estimation. Under contaminated data, the quantile regression parameters estimators show a better efficiency that the linear regression parameters in terms of MSE. Under not contaminated data the MSE of quantile regression parameter estimator show a good performance in comparison to linear regression parameters.

CHAPTER 6

Practical Application

MU284 is a dataset with 284 municipalities of Sweden. The municipal tax receipts in 1985 (RMT85) is the study variable, CS82 (number of Conservative Party seats in the municipal council) is considered as an auxiliary variable. The below graphic shows the relationship between the two variables in the population:

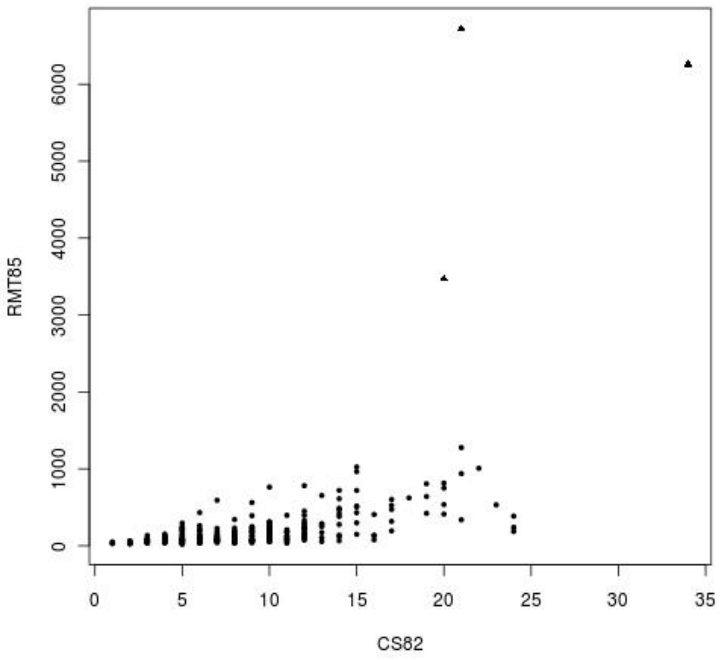

FIGURE 6.1. Scatterplot of a study variable and an auxiliary variable from MU284 dataset

Note that there are three influential points. A Monte Carlo simulation is performed, 5000 samples are drawing the population with a sample size of $n = 100$. It is expected in practice that a significant fraction of the influential points of the population appears when is drawn a sample.

To ensure that there is a similar proportion of influential points in the samples than in the population, a point is selected in each sample.

The simulation results suggest that the QREG estimator has a better performance that the GREG estimator and the HT estimator in terms of MSE.

Model	π-estimator	GREG-estimator	QREG-estimator
RB	-1.05	-0.81	-1.13
MSE	36492024	42836054	28026580

TABLE 6.1. Comparison between estimators for the MU284 Data Application

The relative efficiency of the QREG estimator with respect to GREG estimator is 1.528, additionally, QREG estimator is approximately unbiased.

CHAPTER 7

Conclusions and Areas of Further Work

Quantile regression as a robust alternative to the least squares estimation for the linear model and has shown higher efficiency than the linear regression over a wide range of error distributions and under the presence of influential points.

In this thesis a quantile model-assisted approach was used in order to estimate a finite population total through a proposed estimator (QREG estimator). The performance of the QREG estimator was evaluated empirically via simulation studies under different scenarios, The estimator showed a good performance in terms of a smaller bias and mean square error under several scenarios such as skewed distribution and under the presence of influential points both in the auxiliary information and in the variable of interest.

A quantile regression based estimator for a population total was proposed using survey data. Firstly, estimators for the quantile regression parameters were proposed using survey data. Then, a quantile regression estimator was built in order to take into account auxiliary information available for the whole population and using the general form of the generalized difference estimator. Monte Carlo simulations were used in order to assess the accuracy and the precision of these proposed estimators in terms of bias and means square error. A univariate quantile regression model was simulated to evaluate the performance of the regression parameters, the estimators of the slope and the intercept parameters in a quantile regression were more efficient compared with the classical least squares estimator under contaminated data. Regarding the quantile regression based estimator for a population total, the proposed (QREG) estimator worked properly in those scenarios where the classical linear regression holds and it has better performance than the GREG estimator in the presence of influential points, e.g. x-axis outliers, y-outliers or both and under errors with skewed distributions such as exponential, lognormal and normal mixtures,

This QREG estimator could be used in in several scenarios common in survey sampling practice in order to get the estimation of a population total when auxiliary information is available and the classical lineal model does not fit properly (errors with skewed distributions and/or presence of influential points).

Areas of further research are the consideration of nonparametric methods such as a locally polynomial quantile regression in the case of not clear patterns in a scatter-plot, the use of nonlinear quantile regression and the use of quantile regression to assist discrete data models. (Koenker, 2005) We have also considered in this thesis, the specific case when there is only one auxiliary variable so it is necessary to study the inclusion of several

auxiliary variables in the model. Theoretical properties of the QREG estimator such as asymptotic unbiasedness, consistency and sufficiency need to be assessed from a theoretical point, with respect to this Wang and Opsomer (2011) has studied the theoretical properties of survey estimators of quantile populations that consider non-diferentiable functions of estimated quantities.

APPENDIX A

Some Influential Point Scenarios

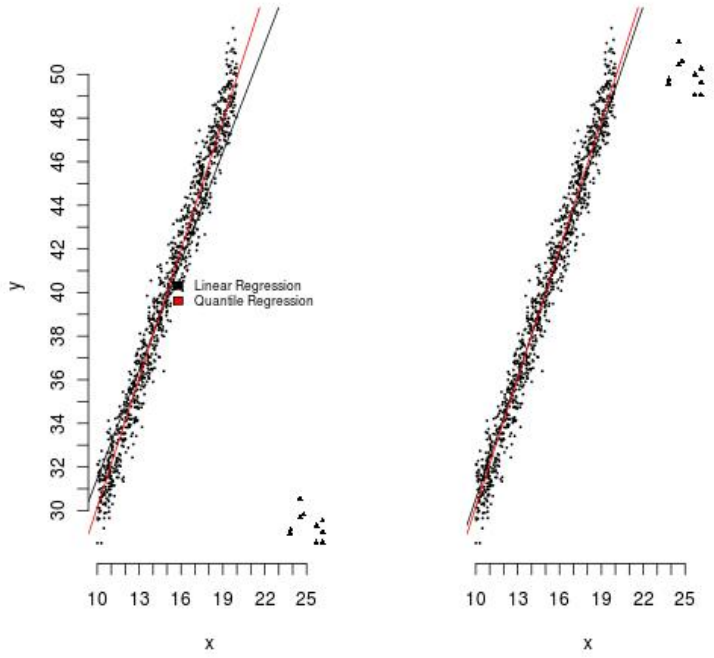

FIGURE A.1. Linear model with normal residual and the presence of 1% of $U(\overline{x}+3s, \overline{x}+4s)$ outlier points in the x-axis (M_{1a} and M_{1b})

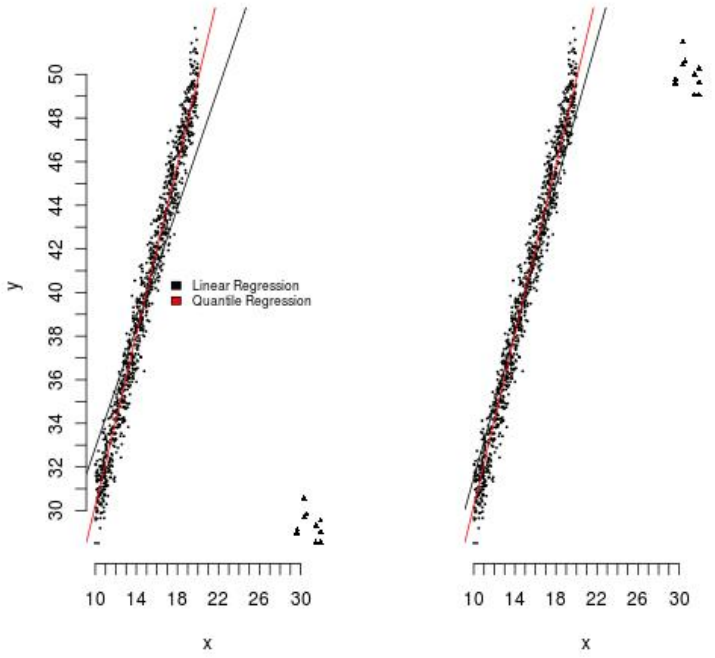

FIGURE A.2. Linear model with normal residual and the presence of 1% of $U(\overline{x}+5s, \overline{x}+6s)$ outlier points in the x-axis (M_{1c} and M_{1d})

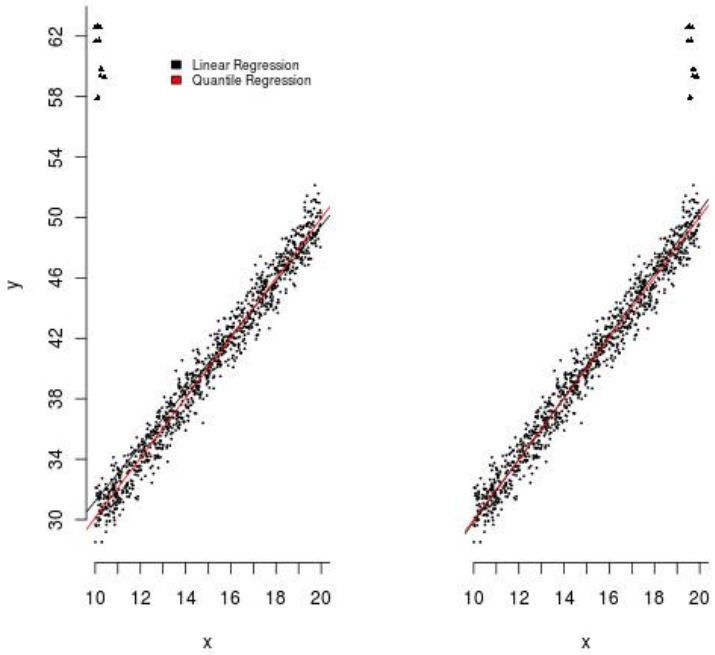

FIGURE A.3. Linear model with normal residual and the presence of 1% of $U(\overline{y}+3s, \overline{y}+4s)$ outlier points in the y-axis (M_{2a} and M_{2b})

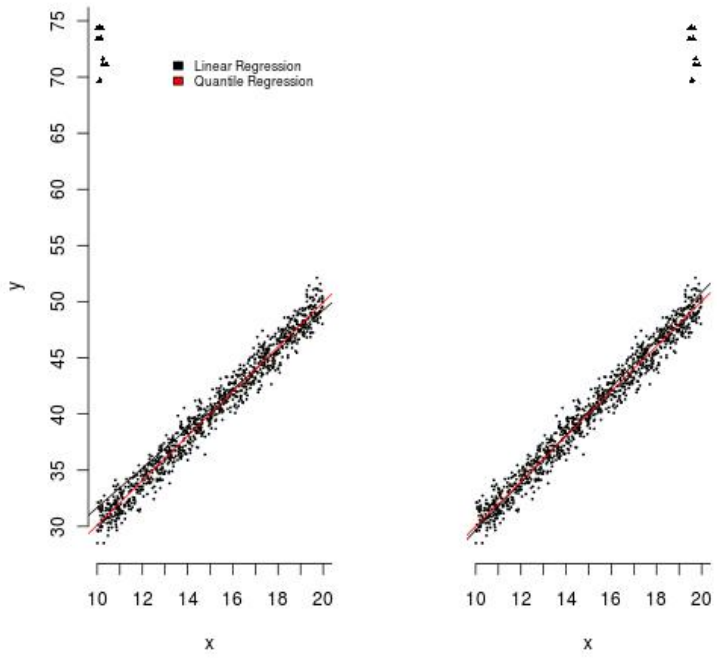

FIGURE A.4. Linear model with normal residual and the presence of 1% of $U(\overline{y}+5s, \overline{y}+6s)$ outlier points in the y-axis (M_{2c} and M_{2d})

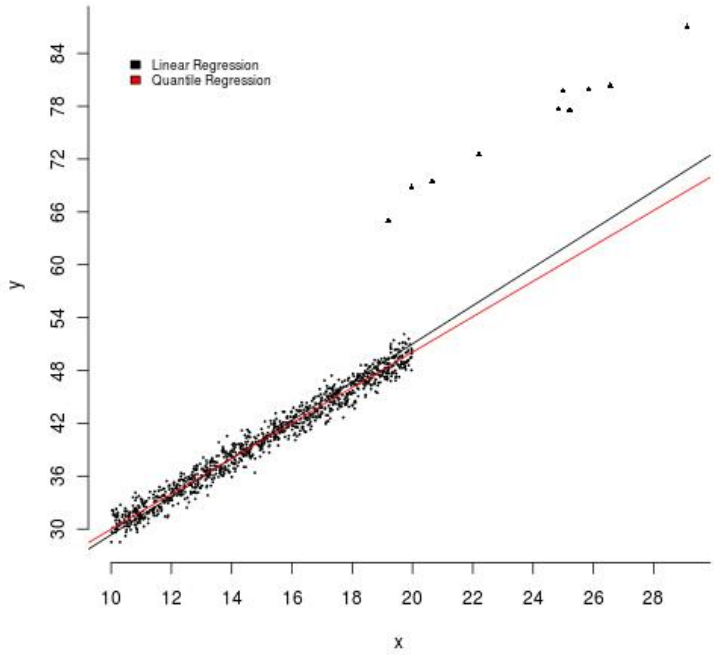

FIGURE A.5. Linear model with normal residual and the presence of 1% of $U(\overline{x}+3s, \overline{x}+4s)$ and $U(\overline{y}+3s, \overline{y}+4s)$ outlier points in the x-axis and the y-axis (M_{3a})

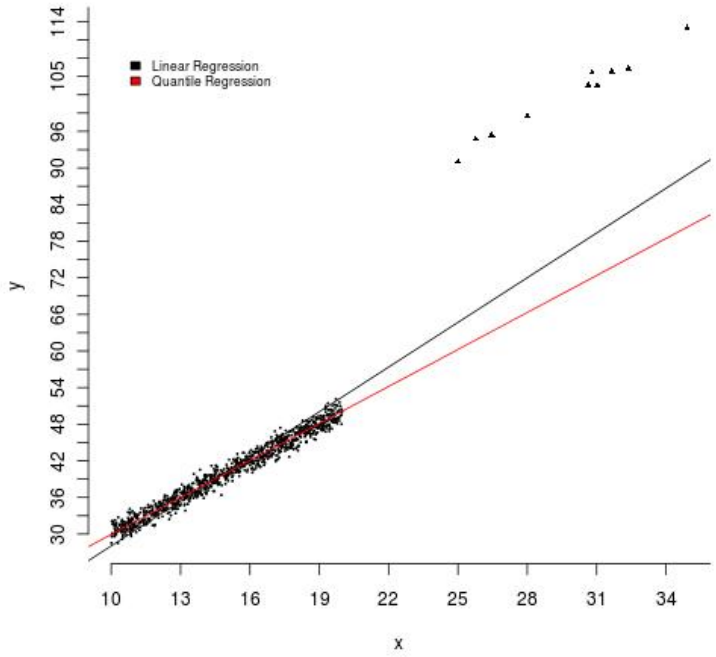

FIGURE A.6. Linear model with normal residual and the presence of 1% of $U(\overline{x}+3s, \overline{x}+4s)$ and $U(\overline{y}+3s, \overline{y}+4s)$ outlier points in the x-axis and the y-axis (M_{2b})

APPENDIX B

R Code: Evaluation of the QREG estimator through modified Monte Carlo method

```
Make_Total_Simulations=function(y,x,WEIGHTS=rep(1,N),Number_Simulations,N,n,influential_indicator){

is.wholenumber <- function(x, tol = .Machine$double.eps^0.5){ abs(x - round(x)) < tol }

Number_influential_points=sum(influential_indicator)
Prop_influential_points=Number_influential_points/N

n_influentials=Prop_influential_points*n
n_noinfluentials=n-n_influentials

if (!is.wholenumber(n_influentials)) {
n_influentials=round(n_influentials)
warning("The fraction of contaminated values in the sample is not a integer, it will be rounded")
}

if (!is.wholenumber(n_noinfluentials))
{
n_noinfluentials=round(n_noinfluentials)
warning("The fraction of noncontaminated values in the sample is not a integer, it would be rounded")
}


#Seed sets
set.seed(489342)
seeds=runif(Number_Simulations+1000, min=0, max=100000)
seeds=ceiling(seeds)
seeds=seeds[!duplicated(seeds)]

set.seed(489342)
sam=sample(1:length(seeds),Number_Simulations)
seeds=seeds[sam]
#table(duplicated(seeds))

MATRIX=matrix(F,N,Number_Simulations)

for(j in 1:Number_Simulations){
set.seed(seeds[j])
subindex_noinfluentials=which(influential_indicator==0)   #All positions of no inluential values
subindex_noinfluentials=sample(subindex_noinfluentials,n_noinfluentials)
```

```
MATRIX[subindex_noinfluentials,j]=T
}

for(j in 1:Number_Simulations){
set.seed(seeds[j])
subindex_influentials=which(influential_indicator==1)   #All positions of  inluential values
subindex_influentials=subindex_influentials[sample(1:length(subindex_influentials),n_influentials)]
MATRIX[subindex_influentials,j]=T
}


test_pi=numeric(Number_Simulations)
for(j in 1:Number_Simulations){
test_pi[j]=(N/n)*sum(y[MATRIX[,j]])
}

mean_test_pi=mean(test_pi)
var_test_pi=var(test_pi)

mean_test_ratio=mean(test_ratio)
var_test_ratio=var(test_ratio)

BETA=matrix(NA,nrow=2,ncol=Number_Simulations)
for(j in 1:Number_Simulations){
REG=lm(y[MATRIX[,j]]~x[MATRIX[,j]],weights=rep(N/n,n)*WEIGHTS[MATRIX[,j]])
BETA[,j]=matrix(c(as.numeric(REG$coefficients)),ncol=1)
}

X=matrix(c(rep(1,N),x),ncol=2)

test_greg=numeric(Number_Simulations)
for(j in 1:Number_Simulations){
test_greg[j]= (N/n*sum(y[MATRIX[,j]]) ) + ((t((matrix(apply(X,MARGIN=2,FUN=sum),ncol=1)) -
 ((N/n)*matrix(apply(X[MATRIX[,j],],MARGIN=2,FUN=sum),ncol=1)) ))%*%BETA[,j]  )
}

mean_test_greg=mean(test_greg)
var_test_greg=var(test_greg)


BETA_qr=matrix(NA,nrow=2,ncol=Number_Simulations)
for(j in 1:Number_Simulations){
QREG=rq(y[MATRIX[,j]]~x[MATRIX[,j]],method="fn",weights=rep(N/n,n))
BETA_qr[,j]=matrix(c(as.numeric(QREG$coefficients)),ncol=1)
}

X=matrix(c(rep(1,N),x),ncol=2)

test_qr=numeric(Number_Simulations)
for(j in 1:Number_Simulations){
#X[MATRIX[,j] y[MATRIX
test_qr[j]= (N/n*sum(y[MATRIX[,j]]) ) + ((t((matrix(apply(X,MARGIN=2,FUN=sum),ncol=1)) -
 ((N/n)*matrix(apply(X[MATRIX[,j],],MARGIN=2,FUN=sum),ncol=1)) ))%*%BETA_qr[,j]  )


}

mean_test_qr=mean(test_qr)
var_test_qr=var(test_qr)
```

```
################################    Summaries ################################################

mean_test=c(sum(y),mean_test_pi,mean_test_ratio,mean_test_greg,mean_test_qr)
var_test=c(NA,var_test_pi,var_test_ratio,var_test_greg,var_test_qr)

RB_test_pi=( sum ( (test_pi-sum(y))/sum(y) ) / Number_Simulations ) *100
MSE_test_pi=( sum( (test_pi-sum(y))^2  )  )/Number_Simulations

RB_test_ratio=( sum ( (test_ratio-sum(y))/sum(y) ) / Number_Simulations ) *100
MSE_test_ratio=( sum( (test_ratio-sum(y))^2  )  )/Number_Simulations

RB_test_greg=( sum ( (test_greg-sum(y))/sum(y) ) / Number_Simulations ) *100
MSE_test_greg=( sum( (test_greg-sum(y))^2  )  )/Number_Simulations

RB_test_qr=( sum ( (test_qr-sum(y))/sum(y) ) / Number_Simulations ) *100
MSE_test_qr=( sum( (test_qr-sum(y))^2  )  )/Number_Simulations

Relative_Bias=c(NA,RB_test_pi,RB_test_ratio,RB_test_greg,RB_test_qr)
MSE=c(NA,MSE_test_pi,MSE_test_ratio,MSE_test_greg,MSE_test_qr)

MEASURES=matrix(,nrow=5,ncol=5)
MEASURES[1,]=mean_test
MEASURES[2,]=var_test
MEASURES[3,]=sqrt(var_test)/mean_test*100
MEASURES[4,]=Relative_Bias
MEASURES[5,]=MSE
colnames(MEASURES)=c("TOTAL","pi-estimator","ratio estimator","greg-estimator","qr-estimator")
rownames(MEASURES)=c("Total","VarMontecarlo","Cv MC","Relative Biased","Mean Squared Error")

MEASURES
}
```

Bibliography

Breidt, F. and Opsomer, J. D. (2000), 'Local polynomial regression estimators in survey sampling', *Annals of Statistics* **28**, 1026–1053.

Chambers, R. (1986), 'Outlier robust finite population estimators', *Journal of the American Statistical Association* **81**, 1063–1069.

Chambers, R. (2000), 'Winsorization for identifying and treating outliers in business surveys', *Proceedings of the Second International Conference on Establishment Surveys* pp. 717–726.

Chatterjee, S. and Hadi, A. (1986), 'Influential observations, high leverage points, and outliers in linear regression', *Statistical Science* **1**, 379–393.

Deville, J. and Sarndal, C. (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 EstadÃ-stica* **32**, 123–1433.

Hao, L. and Naiman, D. Q. (2007), *Quantile Regression*, Sage Publications.

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

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Koenker, R. (2011), *quantreg: Quantile Regression.* R package version 4.67. *http://CRAN.R-project.org/package=quantreg

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.

Lee, H. (1992), *Outliers in business surveys. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson,A., Colledge, M.J., Kott, P.S. (Eds.), Business Survey Methods*, John Wiley.

Montanari, G. and Ranalli, G. (2005), 'Nonparametric model calibration estimation in survey sampling', *Journal of the American Statistical Association* **100**, 1429–1442.

Mora, H. (2005), 'Metodos numericos para la estimacion de parametros en regresion cuantilica', *Revista Colombiana de EstadÃstica* **28**, 221–231.

Pfeffermann, D. and Rao, C. (2011), *Sample Surveys: Inference and Analysis, Handbook of Statistics 29B*, Elsevier.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, RSarndal92 Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org/

Sarndal, C., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

Wolter, K. (2007), *Introduction to Variance Estimation*, Springer.