Quantile Regression with Rank-Based Samples

by

Olawale Fatai Ayilara

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg

Abstract

Quantile Regression, as introduced by Koenker, R. and Bassett, G. (1978), provides a complete picture of the relationship between the response variable and covariates by estimating a family of conditional quantile functions. Also, it offers a natural solution to challenges such as; homoscedasticity and sometimes unrealistic normality assumption in the usual conditional mean regression. Most of the results for quantile regression are based on simple random sampling (SRS). In this thesis, we study the quantile regression with rank-based sampling methods. Rank-based sampling methods have a wide range of applications in medical, ecological and environmental research, and have been shown to perform better than SRS in estimating several population parameters. We propose a new objective function which takes into account the ranking information to estimate the unknown model parameters based on the maxima or minima nomination sampling designs. We compare the mean squared error of the proposed quantile regression estimates using maxima (or minima) nomination sampling design and observe that it provides higher relative efficiency when compared with its counterparts under SRS design for analyzing the upper (or lower) tails of the distribution of the response variable. We also evaluate the performance of our proposed methods when ranking is done with error.

Acknowledgment

Foremost, I would like to express my sincere gratitude to my advisor Dr. Mohammad Jafari Jozani for his constant moral and financial support during my M.Sc coursework and research. His patience, motivation, enthusiasm, and immense knowledge are inestimable. His guidance helped me throughout the course of the research and writing of this thesis.

Also, I would like to appreciate my thesis committee: Dr. Elif Acar and Dr. Lisa Lix for their encouragement and insightful comments.

My earnest thanks also goes to Dr. Mogbademu, Prof. Ajala, and Mr. Lanre, for their recommendations during the course of my application to the University of Manitoba.

I thank my fellow course-mates at the University of Manitoba: Faisal Atakora, Erfan Hoque, Philip Olabisi, Manqiong Chen, for the challenging discussions, sleepless nights we had working together to meet STAT 7080 deadlines, and for all the fun we have had in the last two years.

Last but not the least, I would like to thank my wonderful family, my inestimable jewel Funmilayo Ayilara (Nee Fatubarin), for giving birth to me and supporting me morally, financially and spiritually throughout my life.

Dedication Page

This work is dedicated to Almighty God.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we briefly describe the conditional mean and quantile regression models and motivate the problem of quantile regression analysis with rank based data. To this end, we give some preliminary results regarding quantile regression, its estimates as well as some of its properties. We also describe how resampling methods can be used to make inference under the quantile regression setting. Finally, we perform a numerical study to fit quantile regression in R and provide the outline of the thesis.

1.1 Motivation

Regression is commonly used as a statistical method to quantify the relationship between explanatory variables and a response variable by modeling the relationship between the conditional mean of the response variable and explanatory variables. For example, in a medical study one might be interested in examining the relationship between the spinal bone mineral density (BMD) and age of adolescent. Figure 1.1 shows the measurements of the BMD (Y) and age (X) of 261 north American

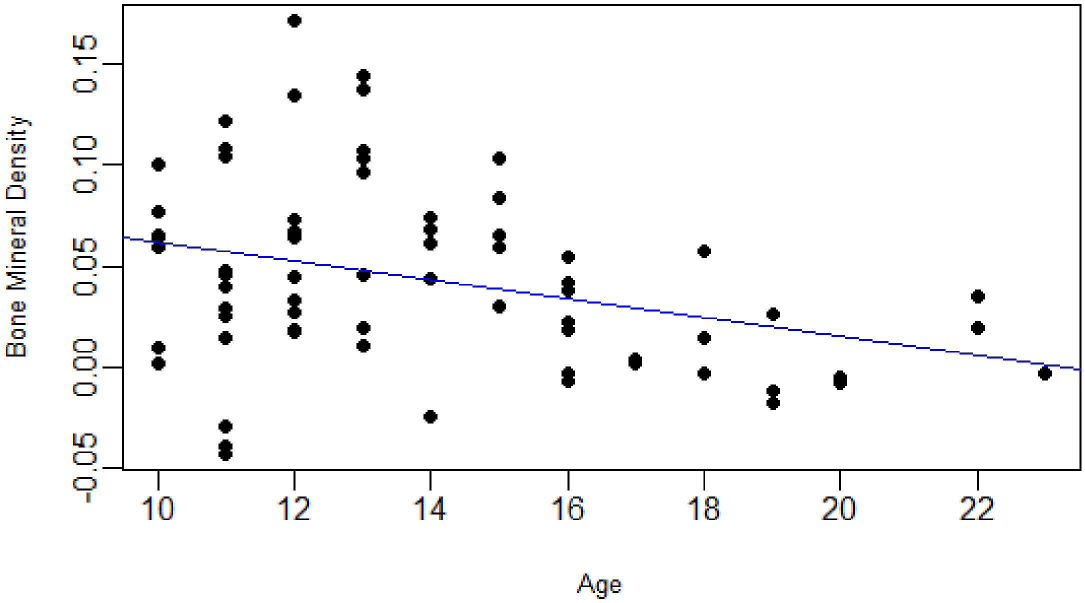

Figure 1.1: Bone mineral density as a function of the adolescent age for the baseline study with the least squares regression line.

adolescents (Bachrach, L.K. et al., 1999) as well as a fitted regression line between Y and X using a linear regression model. A classical linear regression model for this example can be written as follow

$$\mathbf{Y} = \mathbf{X}^\top \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\mathbf{X} = (1, X_1)^\top$ with X_1 being the age of adolescent, $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ is the vector of spinal BMD measurements and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ is the vectors of errors which is assumed to follow a normal distribution.

As we can see in Figure 1.1, the fitted linear regression model does not provide a complete picture of the relationship between BMD and age in the sense that the change in the shape of the distribution of BMD is not captured. This is because the focus of (1.1) is to describe the relationship between the expectation of the response

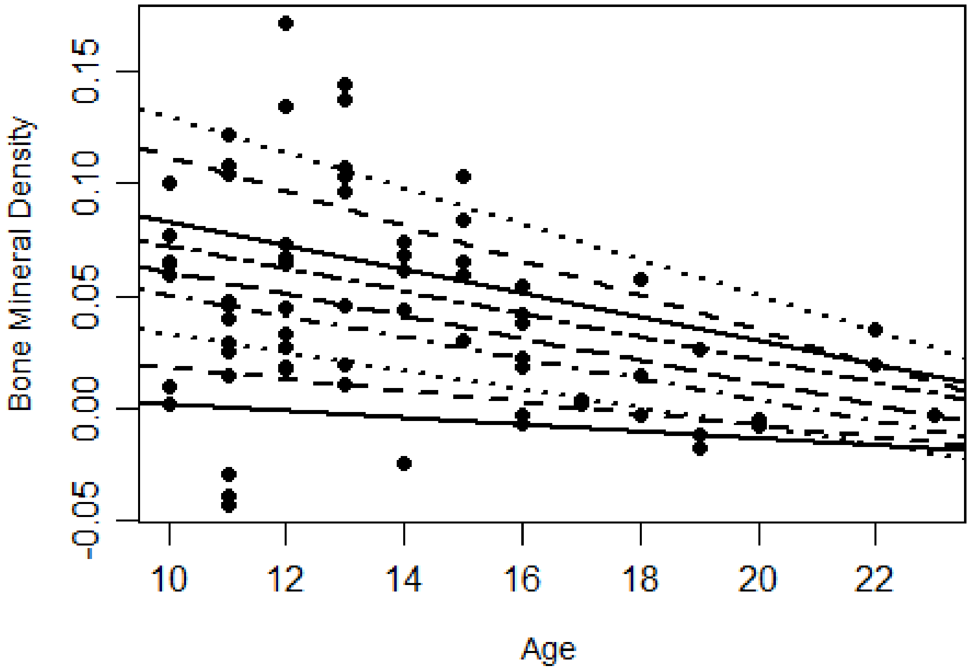

Figure 1.2: Bone mineral density as a function of the adolescent age for the baseline study with the least squares regression line and quantile regression line for $\tau = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$.

variable given the values of a set of covariates, which gives an incomplete picture of the relationship between the response variable and covariates in non-central locations, which is incisively where the interest of most researchers lies. A better descriptive model could be a set of quantile regression models as shown in Figure 1.2.

There are many real situations where one might be interested in studying the behaviour of the tails of the distribution of the response variable as a function of the explanatory variables. In such situations, modeling the conditional mean has inherent limitations which makes it inefficient to address research questions regarding non-central locations on the response distribution. As an example, consider the problem of investigating the impact of various demographic characteristics and maternal behavior on the birth weight of infants born in the United States (Abreveya, J., 2001). It is known that low birth weight is associated with several health problems. These problems have been linked with educational attainment and labor market outcomes, resulting in study factors that influence birth weight and public policy initiatives that might prove effective in reducing the number of infants with low birth weight. Most of the literature on the analysis of birth weight have employed traditional conditional mean regression methods. It has been recognized that the resulting estimates of various effects on the conditional mean of birth weights were not necessarily indicative of the size and nature of these effects on the lower tail of the birth-weight distribution (Koenker Roger, 2005). Several studies have used probit models in the case of binary response for the occurrence of low birth weights in an effort to directly concentrate on the lower tail.

Also, sometimes the homoscedasticity assumption in the conditional mean regression fails (e.g., Fig 1.1), which makes it ill-equipped to fit a good model or capture necessary information in model fitting. In areas such as epidemiology, economics and environmental studies where heavy tailed distributions are common, leading to a prevalence of outliers, the conditional mean can then become inappropriate and misleading because it is heavily influenced by outliers.

Quantile regression introduced by Koenker, R. and Bassett, G. (1978) offers a natural solution to these aforementioned challenges posed by the conditional mean regression. A more complete picture of covariate effects can be provided by estimating a family of conditional quantile functions. While the classical regression model specifies change in the conditional mean of the response variable associated with changes in the explanatory variables, the quantile regression model specifies changes in the conditional quantile.

The possibility of using multiple quantiles allows us to have a complete understanding of how the distribution of the response variable is affected by explanatory variables, including information about the change in the shape of the distribution. The nature of quantile regression and the ability to deal with different types of distributions enable us to eliminate dependence upon the normality assumptions and deal with the problem with more realistic solutions.

Quantile regression has been used to study the effects of demographics and maternal behavior on the distribution of birth outcomes (Koenker Roger, 2005). Austin, P.C. and Schull, M.J. (2003) assessed the association between hospital transport interval and patient and system characteristics using quantile regression. They demonstrated that richer inferences can be drawn from the data using quantile regression. Wei, Y. et al. (2006) compared estimated reference curves for height using the penalized likelihood approach of Cole, T.J. and Green, P.J. (1992) with quantile regression curves based on data used for modern Finnish reference charts.

Most of the literature on quantile regression are based on simple random sampling (SRS). However, this sampling technique is not economical in situations where the actual measurement of the variable of interest is difficult, destructive and expensive, such as BMD measurement, determination of age structure of fish. Also, there are cases where one might be interested in studying the effect of explanatory variables on the tails of the distribution of the response variable. In such cases, SRS might not necessarily give a good sample from the tail of the distribution of the response variable and it will be beneficial to work with more structured sampling designs to get more directed samples from the underlying population. In this thesis, the goal is to study quantile regression based on some of the variation of ranked set sampling (RSS) (McIntyre, G.A., 1952). RSS is a sampling approach that could lead to improved statistical inference for many situations where the measurements are difficult or expensive to obtain but sampling units can be easily and cheaply ordered by some means before taking the actual measurement on them. In this chapter, we briefly review quantile regression and demonstrate it using a SRS data, collected by Bachrach, L.K. et al. (1999), to study bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian Youth. A formal definition of RSS and some of its properties are presented in chapter 2.

1.2 Quantiles

Suppose Y is a continuous random variable with a cumulative distribution function (CDF), $F(y) = P(Y \leq y)$, $y \in \mathbb{R}$. For any $0 < \tau < 1$, the τ^{th} quantile of Y, defined by $Q_\tau(Y)$, is given as

$$Q_\tau(Y) = F^{-1}(\tau) = \inf\{y : F(y) > \tau\}.$$

Examples of $Q_\tau(Y)$ are the first, second and third quartiles corresponding to the choices of $\tau = 0.25, 0.50$ and 0.75, respectively. One can easily show that $Q_\tau(Y)$ is a non-decreasing function of τ, i.e $Q_{\tau_1}(Y) \leq Q_{\tau_2}(Y)$ for $\tau_1 < \tau_2$. To see this, we note that $\{y : F(y) \geq \tau_1\} \supset \{y : F(y) \geq \tau_2\}$. This implies that $\inf\{y : F(y) \geq \tau_1\} \leq \inf\{y : F(y) \geq \tau_2\}$, which means $Q_{\tau_1}(Y) \leq Q_{\tau_2}(Y)$.

1.2.1 Conditional Quantile

Suppose Y is the response variable, and $\mathbf{X}$ is the p-dimensional explanatory variable. Let $F_Y(y \mid \mathbf{X} = \mathbf{x}) = P(Y \leq y \mid \mathbf{X} = \mathbf{x})$ denote the conditional CDF of Y given $\mathbf{X}$. Then, the τ^{th} conditional quantile of Y is defined as

$$Q_\tau(Y \mid \mathbf{X}) = \inf\{y : F_Y(y \mid \mathbf{X}) > \tau\}.$$

Assuming a linear regression between $Q_\tau(Y \mid \mathbf{X})$ and $\mathbf{X}$, the linear quantile regression model for a given conditional quantile τ can be formulated as follows

$$Q_\tau(Y \mid \mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}(\tau), \tag{1.2}$$

where $0 < \tau < 1$, $Q_\tau(\cdot \mid \cdot)$ denotes the conditional quantile function for the τ^{th} quantile and $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau))^\top$ is the vector of quantile coefficients that depends on τ.

The classical regression model in (1.1) has only one conditional mean expressed by one equation. In contrast, (1.2) shows that the quantile regression model can have several conditional quantiles. As a result, several equations can be expressed in the form of (1.2). For example, if the quantile regression model specifies 10 quantiles, the 10 equations will yield 10 sets of coefficients for the covariates, associated with specific choices of quantiles.

1.3 Quantile Regression Estimation

Suppose we observe $(Y_i, \mathbf{X}_i, i = 1, ..., n)$ is a random sample of size n from the underlying population. Using the least squares method one can estimate the parameter of the model (1.1) by minimizing the sum of squared residuals;

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2. \tag{1.3}$$

Minimization of (1.3) can be achieved by taking the partial derivatives of (1.3) with respect to $\boldsymbol{\beta}$, and setting the partial derivatives equal to zero. This results in the least squares estimation of $\boldsymbol{\beta}$ as follow

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y,$$

where $\mathbf{X}$ is an $n \times (p+1)$ design matrix with full column rank, that is $rank(\mathbf{X}) = p + 1 \leq n$ and Y is the vector of the response variable. For quantile regression and to model $Q_\tau(Y \mid \mathbf{X})$ as a function $\mathbf{X}$, one needs to replace the squared error loss in (1.3) by the following loss function

$$\rho_\tau(U_i) = (\tau - I(U_i < 0))U_i, \quad U_i = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}, \tag{1.4}$$

which is also known as the check function. One can easily show that the check function (1.4) can be written in the following form,

$$\rho_\tau(U_i) = |U_i|(\tau I(U_i > 0) + (1 - \tau) I(U_i < 0)), \tag{1.5}$$

Note that ρ_τ can be considered as an asymmetric absolute error loss function as shown in Figure 1.3. That is, a weighted sum of absolute deviations, where a weight $(1 - \tau)$ is assigned to the negative deviations and a weight τ is used for the positive deviations. When $\tau = 0.5$, the quantile loss function is equivalent to the absolute error loss function. Examples of ρ_τ for different values of τ are shown in Figure 1.3

Quantiles can be obtained as solutions to minimization problems associated with suitable choices of the check function. Without loss of generality, assume that Y is a continuous random variable. The expected value of the sum of deviations from a given center a with the check function can be written as $E[\rho_\tau(Y-a)]$. We show that a specific choice of the check function in (1.5) guarantees the modeling of the τ^{th} quantile of Y as function of $\mathbf{X}$. To see this, we have

$$\begin{aligned} Q_\tau(Y) &= \underset{a}{\operatorname{argmin}}\, E[\rho_\tau(Y-a)] \\ &\equiv \underset{a}{\operatorname{argmin}} \left[(1-\tau)\int_{-\infty}^{a} |y-a| f(y)dy + \tau \int_{a}^{+\infty} |y-a| f(y)dy\right]. \end{aligned}$$

Differentiating $E[\rho_\tau(Y-a)]$ with respect to (w.r.t.) a and setting the partial derivatives equal to zero, we have

$$\begin{aligned} \frac{\partial}{\partial a} Q_\tau(Y) &= \frac{\partial}{\partial a}\left[(1-\tau)\int_{-\infty}^{a} (a-y) f(y)dy + \tau \int_{a}^{+\infty} (y-a) f(y)dy\right] \\ &= (1-\tau)\int_{-\infty}^{a} f(y)dy - \tau \int_{a}^{+\infty} f(y)dy = 0, \end{aligned}$$

which results in

$$(1-\tau)F(a) - \tau(1-F(a)) = 0,$$

or equivalently,

$$F(a) = \tau.$$

That is, what minimizes $E[\rho_\tau(Y-a)]$ in a is equal to $Q_\tau(Y)$, the τ^{th} quantile of Y. Following the same method, the necessary coefficients of the linear quantile regression in (1.2) can be obtained as

$$\widehat{\boldsymbol{\beta}(\tau)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}). \tag{1.6}$$

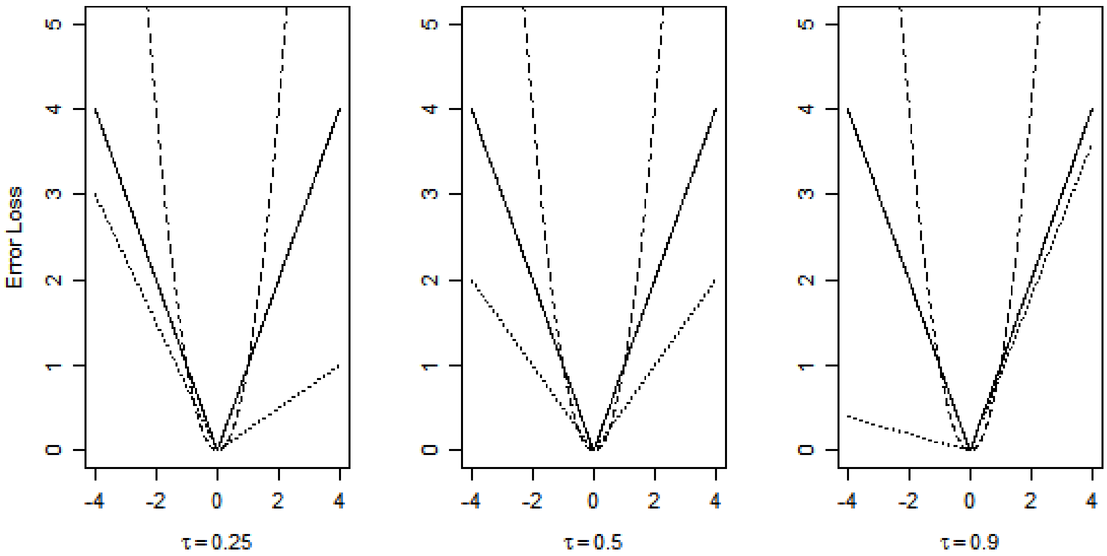

Figure 1.3: Absolute error loss (solid), squared (dashed) and quantile error loss function (dotted).

1.3.1 Difficulties in Estimation of Quantile Regression Parameters

Estimating the quantile regression parameters using the least squares or maximum likelihood method does not lead to closed form expressions for $\boldsymbol{\beta}(\tau)$ even in the simplest quantile regression model with minimal assumptions. Also, calculating such estimates is not easy. This is because, $\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$ is not differentiable everywhere so standard numerical algorithms do not work. To be more specific, $\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$ is not differentiable at points where one or more residuals $Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}$ are zero.

One way to look at the problem is to approximate the loss function $\rho_\tau(U)$ with a smooth function (Chen, C., 2003) and apply standard numerical algorithms.

Another way, which is the most popular, is to the minimize (1.6) by applying linear programming approach using the simplex method (Koenker Roger, 2005). This is the method that will be adopted for this thesis. In the next section, we briefly review the linear programming approach towards finding $\hat{\boldsymbol{\beta}}(\tau)$ in the quantile regression.

1.4 Linear Programming

A linear programming is an optimization method in achieving the best solution in a system of equations in which the objective and all constraints are linear. In other words, the objective of a general linear program is minimizing $c^\top \mathbf{x}$ subject to linear constraints such as $a_i^\top \mathbf{x} \geq b_i$, $i \in M_1, a_i^\top \mathbf{x} \leq b_i$, $i \in M_2, a_i^\top \mathbf{x} = b_i$, $i \in M_3$, $x_j \geq 0$, $j \in N_1$, $x_j \leq 0$, $j \in N_2$, where $c \in R^p$ is a cost vector, $\mathbf{x} \in R^p$ is a vector of decision variables, and constraints are given by $a_i \in R^p$ and $b_i \in R$ for $i \in \{1, 2, ..., m\}$. Index sets $M_1, M_2, M_3 \subseteq \{1, 2, ..., m\}$ and $N_1, N_2, \subseteq \{1, 2, ..., n\}$ are used to distinguish between different types of constraints. This type of problem was first solved formally using linear programming methods by Stigler, G. (1945).

An equality constraint $a_i^\top \mathbf{x} = b_i$ is equivalent to the pair of constraints $a_i^\top \mathbf{x} \leq b_i$ and $a_i^\top \mathbf{x} \geq b_i$. Also a constraint of the form $a_i^\top \mathbf{x} \leq b_i$ can be written as $(-a_i)^\top \mathbf{x} \geq -b_i$. Each occurrence of an unconstrained variable $\mathbf{x}_j$ can be replaced by $\mathbf{x}_j{}^+ + \mathbf{x}_j{}^-$ where $\mathbf{x}_j{}^+$ and $\mathbf{x}_j{}^-$ are two new variables with $\mathbf{x}_j{}^+ \geq 0$ and $\mathbf{x}_j{}^- \leq 0$. We can thus write every linear program in the following general standard form

$$\min_{c}\{c^\top \mathbf{x} : A\mathbf{x} \geq b, \mathbf{x} \geq 0\}, \tag{1.7}$$

where $\mathbf{x}, c \in R^p, b \in R^m$ and $A \in R^{m \times p}$.

One can easily write any linear programming problem into the standard form by replacing each inequality constraint of the form $a_i^\top \mathbf{x} \leq b_i$ or $a_i^\top \mathbf{x} \geq b_i$ by a constraint $a_i^\top \mathbf{x} + z_i = b_i^*$ or $a_i^\top \mathbf{x} - z_i = b_i^*$ where z_i is a new so called *slack variable*, and an additional constraint $z_i \geq 0$. The general form is typically used to discuss the theory of the linear programming, while the standard form is often used when designing algorithms for linear programming to solve problems numerically.

1.4.1 Linear Program Duality

Every linear program has an inverse or dual formulation. Suppose our original objective, known as the *primal problem*, is a minimization problem. Corresponding to any primal problem, we may formulate a dual linear program in which minimizing with respect to the original variables turns into maximizing with respect to the Lagrange multiplier of the dual formulation. The concept of duality can be examined more generally following Berman, A. (1973). To this end, consider the linear programming problem

$$\min\{c^\top \mathbf{x} : A\mathbf{x} \geq b, \mathbf{x} \geq 0\}. \tag{1.8}$$

By introducing slack variable z we can rewrite (1.8) as follows

$$\min\{c^\top \mathbf{x} : A\mathbf{x} - z = b, \mathbf{x}, z \geq 0\}. \tag{1.9}$$

Let $X = \{(\mathbf{x}, z) : \mathbf{x}, z \geq 0\} \subseteq R^{m+p}$. Using the Lagrangian multiplier method, optimization of (1.9) can be written as

$$\begin{aligned} L((\mathbf{x}, z), \lambda) &= c^\top \mathbf{x} - \lambda^\top (A\mathbf{x} - z - b) \\ &= (c^\top - \lambda^\top A)\mathbf{x} + \lambda^\top z + \lambda^\top b, \end{aligned}$$

which has a finite minimum over $\mathbf{x}$ if and only if

$$\lambda \in Y = \{u \in R^m : c^\top - u^\top A \geq 0, u \geq 0\}.$$

for $\lambda \in Y$, the minimum of $L((\mathbf{x}, z), \lambda)$ is attained when both $(c^\top - \lambda^\top A)\mathbf{x} = 0$ and $\lambda^\top z = 0$. Thus, we have

$$g(\lambda) = \inf_{(\mathbf{x},z)\in X} L((\mathbf{x}, z), \lambda) = \lambda^\top b,$$

and so, we obtain the dual problem in the following maximization form:

$$\max\{b^\top \lambda : A^\top \lambda \leq c, \lambda \geq 0\}.$$

The dual can be determined analogously as $\max\{b^\top \lambda : A^\top \lambda \leq c\}$.

For the quantile regression problem, the primal formulation of the basic linear quantile regression optimization can be written as

$$\min_{\boldsymbol{\beta}\in R^p} \left[\tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| \right]. \tag{1.10}$$

Splitting the residual vector into positive and negative parts, $e_i = u_i - v_i$, with $u_i = e_i I(e_i > 0)$ and $v_i = |e_i| I(e_i < 0)$, we obtain an equivalent (1.10) expression of the form

$$\min_{\boldsymbol{\beta}\in R^p} \left[\tau \sum_{i=1}^n |e_i| I(e_i > 0) + (1-\tau) \sum_{i=1}^n |e_i| I(e_i < 0) \right] = \min_{\boldsymbol{\beta}\in R^p} \left[\tau \sum_{i=1}^n u_i + (1-\tau) \sum_{i=1}^n v_i \right],$$

which can be written as

$$\min_{\{\boldsymbol{\beta}(\tau),u,v\}} \left\{ \tau 1_n^\top U + (1-\tau) 1_n^\top V \right\},$$

where $U = eI(e > 0)$ and $V = |e|I(e < 0)$. Therefore, the problem of finding the unknown parameters $\beta(\tau)$ in a linear quantile regression can be written as the following linear programming problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}(\tau)) \equiv \min_{\{\boldsymbol{\beta}(\tau),u,v\}} \left\{ \tau 1_n^\top U + (1-\tau) 1_n^\top V \right\}, \tag{1.11}$$

subject to

$$Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}(\tau) = U - V, \quad \boldsymbol{\beta}(\tau) \in R^p, U \geq 0, V \geq 0.$$

Equivalently, we can write (1.10) as follow

$$\min\{\tau 1_n^\top U + (1-\tau) 1_n^\top V \mid Y = \mathbf{X}\boldsymbol{\beta}(\tau) + U - V, (\boldsymbol{\beta}(\tau), U, V) \in R^p \times R_+^n \times R_+^n\}.$$

As shown in Koenker Roger (2005) the primal quantile regression problem has the corresponding dual problem as below

$$\max_d \{y^\top d \mid X^\top d, d \in [\tau - 1, \tau]^\top\}.$$

By linear programming, solutions are obtained on the vertices of the constraint set (Gill, P. et al., 1991). That is, subsets that provide exact fits to the p observations. Using the simplex method, the first step is to find an initial feasible vertex. For example, one can choose any subset h such that $X(h)$ is of full rank and $\boldsymbol{\beta}(h) = X(h)^{-1}Y$. Second step is to travel from one vertex to another until optimality is achieved (Barrodale, I. and Roberts, F. D. K., 1973).

This method is efficient for problems with modest size (several thousands). Speed is comparable to the least squares estimate (LSE) for sample sizes up to several hundreds. However, the method is slow relative to LSE for large problems (Koenker, R. and D'Orey, V., 1985).

A more efficient method than the simplex method for larger sample size is the Frisch-Newton interior point method. In contrast to the simplex method, the algorithm traverses the interior of the feasible region. More details can be found in Stephen, P. and Koenker, R. (1997). In this thesis, we use the simplex method to obtain the unknown parameters $\boldsymbol{\beta}$ in a linear quantile regression. Another approach that we use is based on the M-quantiles which we briefly explain in the next section.

1.5 M-Quantile Regression

In this section, we present another approach, known as M-Quantile regression, to estimate the unknown τ^{th} quantile regression coefficients which uses an iterative method to obtain its solution. M-Quantile regression introduced by Breckling, J. and Chambers, R. (1988) is a generalization of the M-Estimation (Huber, P. J., 1964) approach on robust regression methods. M-Quantile regression models the relationship between the response variable and the covariates for different quantiles of the distribution. The τ^{th} M-Quantile of a continuous random variable Y with CDF $F(y)$ is obtained as the minimizer of $\int \rho_\tau(y-a)\,dF(y)$. M-Quantile estimator reduces to the least absolute value estimator when the loss function $\rho_\tau(U) = |U|$ and to the least squares estimator when the loss function is $\rho_\tau(U) = U^2$.

M-Quantile regression has been widely studied in the literature using nonparametric and semiparametric approaches for count data (Pratesi, M. et al., 2009; Dresassi, E. et al., 2014). Chambers, R. and Tzavidis, N. (2006) applied the M-Quantile approach in small area estimation. Bianchi, A. and Salvati, N. (2015) derived the asymptotic properties and variance estimators of the M-Quantile regression coefficients estimators. In this thesis, we use the M-Quantile regression to estimate the regression coefficients and τ^{th} conditional quantile, $Q_\tau(Y \mid \mathbf{X})$, using maxima or minima nominated samples.

To this end, assuming a linear relationship between $Q_\tau(Y \mid \mathbf{X})$ and $\mathbf{X}$, the linear M-quantile regression model for a given conditional quantile τ can be formulated as follows

$$MQ_\tau(Y \mid \mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}(\tau), \tag{1.12}$$

where $0 < \tau < 1$, $MQ_\tau(\cdot \mid \cdot)$ denotes the conditional quantile function for the τ^{th} quantile and $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau))^\top$ is the quantile coefficient that depends on τ. $\boldsymbol{\beta}(\tau)$ is the solution of the following minimization problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E\left[\rho_\tau\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma_\tau}\right)\right],$$

where σ_τ is a scale parameter that characterizes the distribution of $\epsilon_i(\tau) = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}(\tau)$. Since ρ_τ is continuously differentiable and convex, the M-Quantile regression coefficient estimator $\widehat{\boldsymbol{\beta}(\tau)}$ can be obtained as the solution of the system of equations $\sum_{i=1}^n \psi_\tau\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\hat{\sigma}_\tau}\right)\mathbf{X}_i = 0$, where $\psi_\tau(U) = d\rho_\tau(U)/dU = |\tau - I(U < 0)|\psi(U)$, with $\psi(U) = d\rho(U)/dU$, and $\hat{\sigma}_\tau$ is an estimator of σ_τ. We need to solve this system of equations using an iterative method.

1.6 Properties of Quantile Regression

Some useful theoretical properties of the quantile regression are shown in Koenker, R. and Bassett, G. (1978). Here we mention a few of them that will be used in the thesis. To this end, let $\mathbb{H}(\tau, Y, \mathbf{X})$ be the parameter space, and suppose $\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X}) \in \mathbb{H}(\tau, Y, \mathbf{X})$. Then,

(1) $\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$ is scale invariant. In other words,

(a) $\hat{\boldsymbol{\beta}}^*(\tau, \lambda Y, \mathbf{X}) = \lambda\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$ for $\lambda \in [0, \infty)$,

(b) $\hat{\boldsymbol{\beta}}^*(\tau, -\lambda Y, \mathbf{X}) = \lambda\hat{\boldsymbol{\beta}}^*(1-\tau, Y, \mathbf{X})$ for $\lambda \in (-\infty, 0)$.

To see this, for example when $\lambda \in (0, \infty)$, one can easily see that

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^*(\tau, \lambda Y, X) &= \tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |\lambda Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |\lambda Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}|, \qquad \lambda > 0, \\
&= \tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |\lambda Y_i - \lambda\frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |\lambda Y_i - \lambda\frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}|, \\
&= \lambda\tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}| + \lambda(1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}|, \\
&= \lambda \left[\tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \frac{\mathbf{X}_i^\top}{\lambda} \boldsymbol{\beta}| \right], \\
&= \lambda \left[\tau \sum_{\{i:Y_i > {\mathbf{X}^*}_i^\top \boldsymbol{\beta}\}} |Y_i - {\mathbf{X}^*}_i^\top \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < {\mathbf{X}^*}_i^\top \boldsymbol{\beta}\}} |Y_i - {\mathbf{X}^*}_i^\top \boldsymbol{\beta}| \right], \qquad \mathbf{X}^{*\top} = \mathbf{X}^\top/\lambda, \\
&= \lambda\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})
\end{aligned}
$$

(2) $\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$ is invariant under regression shift. That is, for any $\gamma \in R^p$, $\hat{\boldsymbol{\beta}}^*(\tau, Y + \mathbf{X}_\gamma, \mathbf{X}) = \hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X}) + \gamma$

To show this, we have

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}(\tau, Y+\mathbf{X}_\gamma, \mathbf{X}) &= \tau \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i+\mathbf{X}_{\gamma_i}-\mathbf{X}_i^\top\boldsymbol{\beta}| + (1-\tau)\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i+\mathbf{X}_{\gamma_i}-\mathbf{X}_i^\top\boldsymbol{\beta}|,\ \lambda>0,\\
&= \tau \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}+\mathbf{X}_{\gamma_i}| + (1-\tau)\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}+\mathbf{X}_{\gamma_i}|,\\
&\le \tau \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}| + |\mathbf{X}_{\gamma_i}| + (1-\tau)\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}| + |\mathbf{X}_{\gamma_i}|,\\
&= \tau \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}| + (1-\tau)\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}} |Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}| + \sum_{i=1}^{n}|\mathbf{X}_{\gamma_i}|,\\
&= \hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X}) + \gamma, \quad \text{where,} \quad \gamma = \sum_{i=1}^{n}|\mathbf{X}_{\gamma_i}|.
\end{aligned}
$$

More generally, one can easily show that $\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$ is *equivariance w.r.t monotone transformations.* This is true, since for any non decreasing function h, $P(Y \le a) = P(h(Y) \le h(a))$ and so $Q_{h(Y)}(\tau) = h(Q_\tau(Y))$. The quantiles of the transformed random variable $h(Y)$ are simply the transformed quantities on the original scale. For example, if $\mathbf{X}^\top\boldsymbol{\beta}$ is the τ^{th} conditional quantile of $\ln Y$, then $\exp(\mathbf{X}^\top\boldsymbol{\beta})$ is the τ^{th} conditional quantile of Y. Expectation does not have this property because in general, $E(h(Y)) \neq h(E(Y))$ except when h is linear.

Finally, $\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$ is invariant with respect to the re-parameterization of the design matrix. In other words, for any $|A| \neq 0$, $\hat{\boldsymbol{\beta}}(\tau, Y, \mathbf{X}A) = A^{-1}\hat{\boldsymbol{\beta}}^*(\tau, Y, \mathbf{X})$. This can be easily verified as follows

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{*}(\tau, Y, \mathbf{X}A) &= \tau \sum_{\{i: Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top A\boldsymbol{\beta}| + (1-\tau) \sum_{\{i: Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top A\boldsymbol{\beta}|, \\
&= \tau \sum_{\{i: Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i A A^{-1} - \mathbf{X}_i^\top A\boldsymbol{\beta}| + (1-\tau) \sum_{\{i: Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i A A^{-1} - \mathbf{X}_i^\top A\boldsymbol{\beta}|, \\
&= A^{-1} \left[\tau \sum_{\{i: Y_i^* > \mathbf{X^*}_i^\top \boldsymbol{\beta}\}} |Y_i^* - \mathbf{X^*}_i^\top \boldsymbol{\beta}| + (1-\tau) \sum_{\{i: Y_i^* < \mathbf{X^*}_i^\top \boldsymbol{\beta}\}} |Y_i^* - \mathbf{X^*}_i^\top \boldsymbol{\beta}| \right], \\
&= A^{-1} \hat{\boldsymbol{\beta}}^{*}(\tau, Y, \mathbf{X}),
\end{aligned}
$$

where $Y^* = YA$ and $\mathbf{X}^* = \mathbf{X}^\top A^2$.

1.7 Inference on Quantile Regression

Statistical inference on quantile regression is very important to validate our theory, methods, and practice of forming judgments about the parameters of the population quantiles and the reliability of statistical relationships. In most applications, inferences are based on confidence intervals associated with the quantile regression. In this section, we show the consistency and the asymptotic normality of the linear quantile regression coefficient estimators which are necessary to construct confidence intervals and conduct hypothesis testing. In the context of rank-based set sampling, we rely on inference based on the bootstrap method as we explain in section 1.7. This is because theoretical results regarding statistical inference on quantile regression based on rank-based sampling design are very challenging and left as open problems for future studies in this direction.

1.7.1 Consistency

To show the consistency of the estimators of the coefficients of a linear quantile regression model, we need the following regularity conditions (Koenker, R. and Bassett, G., 1982).

(A_1) The distribution functions of Y given $\mathbf{X}_i$, $F_i(\cdot)$, are absolutely continuous with continuous densities $f_i(\cdot)$ that are uniformly bounded away from 0 and ∞ at $\eta_i(\tau) = Q_\tau(Y \mid \mathbf{X}_i)$.

(A_2) There exist positive definite matrices D_0 and D_1 such that $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^\top = D_0$, $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f_i(\eta_i(\tau))\mathbf{X}_i\mathbf{X}_i^\top = D_1$ and $\max_{i=1,\ldots,n} ||\mathbf{X}_i|| = o(n^{\frac{1}{2}})$.

Having these conditions, we show that $\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \boldsymbol{\beta}(\tau)$ as $n \to \infty$. To this end, using the uniform law of large numbers, we have

$$\sup_{b\in\mathbf{B}} \frac{1}{n}\sum_{i=1}^{n} \left[\rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{b}) - E[\rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{b})]\right],$$

where $\mathbf{B}$ is a compact subset of $\mathbf{R}^p$. If for any $\epsilon > 0$, $\bar{Q}(b) = \frac{1}{n}\sum_{i=1}^{n} E[\rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{b})]$ is bounded away from zero with probability approaching one for $||b - \boldsymbol{\beta}(\tau)|| > \epsilon$ then $\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \boldsymbol{\beta}(\tau)$. Under the regularity conditions, $\bar{Q}(b)$ has a unique minimizer $\boldsymbol{\beta}(\tau)$.

1.7.2 Asymptotic Normality

Under the regularity conditions (A_1) and (A_2) the linear quantile regression model coefficient estimators $\hat{\beta}(\tau)$ are asymptotically normal. In other words,

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\} \xrightarrow{d} N(0, \tau(1-\tau)D_1^{-1}D_0D_1), \tag{1.13}$$

where D_1 and D_0 are defined in (A_2).

For the independent and identically distributed (i.i.d) error models, i.e, $f_i(\eta_i(\tau)) = f_\epsilon(0)$, (1.13) can be simplified as

$$\sqrt{n}\{\hat{\beta}(\tau) - \beta(\tau)\} \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f_\epsilon^2(0)}D_0^{-1}\right).$$

Similar results hold for the non-i.i.d. case. In this situation, Koenker, R. and Machado, J.A.F. (1999) showed that

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\} \xrightarrow{d} N(0, \tau(1-\tau)D_1^{-1}D_0D_1(\tau)),$$

where,

$$D_1(\tau) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{X}_i^\top \boldsymbol{\beta}(\tau))\mathbf{X}_i\mathbf{X}_i^\top.$$

It is worth noting that, the usual derivation of the asymptotic normality for the non-i.i.d. case using the Taylor expansion will not directly work as $\sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$ is not differentiable everywhere.

Using the asymptotic distribution of $\hat{\boldsymbol{\beta}}(\tau)$, one can easily construct confidence intervals and perform testing hypothesis about $\boldsymbol{\beta}(\tau)$. See Koenker, R. and Bassett, G. (1982) for more details.

1.8 Resampling methods

Resampling method especially for small sample sizes, is a very good alternative to make inferences on parameters of interest as it allows estimation of quantities of interest, e.g., standard errors or confidence intervals, with minimal assumptions.

In this thesis, we use the bootstrap method as a reliable and highly recommended re-sampling technique in quantile regression analysis (Cristina, D. et al., 2013). The bootstrap methods are useful tools when the asymptotic approximation on the parameter of interest is difficult or hard to compute. However, the bootstrap method can be computationally expensive and may not be the best method to use for high-dimensional problems.

The two common methods for bootstrap estimation of the parameter of interest are; the residual and the paired bootstrap (Efron, B. and Tibshirani, R.J., 1998). The residual bootstrap works well when the errors are exchangeable in distribution, while the paired bootstrap is based on the assumption that the observations come from a multivariate distribution $F(X, Y)$ (Kocherginsky, M. and He, X., 2007).

Resampling the residuals or pairs of observations $(\mathbf{X}_i, \mathbf{Y}_i)$ B times, we obtain the bootstrap samples $(\mathbf{X}_i^*, \mathbf{Y}_i^*)$ which are used to produce estimates $\hat{\boldsymbol{\beta}}_1^*(\tau), \ldots, \hat{\boldsymbol{\beta}}_B^*(\tau)$ of $\boldsymbol{\beta}(\tau)$. The paired bootstrap makes no assumption about error variance homogeneity, and is therefore more robust (Kocherginsky, M. and He, X., 2007).

1.8.1 Residual Bootstrap

Consider (1.2) with i.i.d. errors. To make inference about $\hat{\beta}(\tau)$ using the residual bootstrap we follow the following steps

1. Compute $\hat{\boldsymbol{\beta}}(\tau)$ using the observed sample, and calculate residuals $\hat{\boldsymbol{\epsilon}}_i = \mathbf{Y}_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}(\tau)$.

2. Draw bootstrap samples $\boldsymbol{\epsilon}_i^*$ $i = 1, \ldots, n$ from $\hat{\boldsymbol{\epsilon}}_i$ with replacement, and define $\mathbf{Y}_i^* = \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}(\tau) + \boldsymbol{\epsilon}_i^*$.

3. Compute the quantile regression bootstrap estimator $\hat{\boldsymbol{\beta}}^*(\tau)$ using the bootstrap sample. Compute the steps B times to obtain $\hat{\boldsymbol{\beta}}_1^*(\tau), \ldots, \hat{\boldsymbol{\beta}}_B^*(\tau)$

4. Estimate the standard error of $\hat{\boldsymbol{\beta}}(\tau)$ with $\widehat{SE(\hat{\boldsymbol{\beta}}(\tau))} = \sqrt{\widehat{Var(\hat{\boldsymbol{\beta}}(\tau))}}$ where,

$$\widehat{Var(\hat{\boldsymbol{\beta}}(\tau))} = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\boldsymbol{\beta}}^*(\tau) - \bar{\hat{\boldsymbol{\beta}}}^*(\tau))^2,$$

   and $\bar{\hat{\boldsymbol{\beta}}}^*(\tau) = \frac{1}{B} \sum_{i=1}^{B} \hat{\boldsymbol{\beta}}_i^*(\tau)$.

5. Carry out inference on $\hat{\boldsymbol{\beta}}(\tau)$ based on the standard error of the bootstrap estimator obtained by constructing confidence interval using the approximate normal method or the percentile method.

1.8.2 Paired Bootstrap

Considering the same scenario as used in the residual bootstrap, the first two steps in the residual bootstrap will not be considered in the paired bootstrap. To perform the paired bootstrap algorithm we proceed as follows:

1. Generate bootstrap samples (Y^*_{ib}, X^*_{ib}) by drawing with replacement samples from the n pairs $(\mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n$ and $b = 1, \ldots, B$.
2. Compute the quantile regression bootstrap estimators using the bootstrap samples, to obtain $\hat{\boldsymbol{\beta}}^*_1(\tau), \ldots, \hat{\boldsymbol{\beta}}^*_B(\tau)$.
3. Repeat steps 4-5 as stated under the residual bootstrap.

1.9 Implementation in R

The quantile regression can be implemented in R following the steps below;

1. Install the "quantreg" package in R
2. Load the library using require("quantreg")
3. Use rq() to fit the necessary model.

```
require("quantreg")

quantile = c(0.10,0.25,0.35,0.50,0.75,0.95)

for(i in 1:length(quantile)){
fit = rq(y~x, tau = quantile[i],data = pop.data)

# Assuming iid errors use
summary.rq(fit,tau = quantile[i],alpha=0.05,iid=TRUE)

# Assuming non iid errors use
summary.rq(fit,tau=0.5,alpha=0.05,iid=FALSE)
}
```

1.10 Example

In this section, we present an example for the quantile regression using a dataset collected by Bachrach, L.K. et al. (1999) to study on bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian Youth. In this example, we consider the baseline data of participants who completed the follow up study. In estimating the BMD using the age, we showed in Figure 1.1 that the conditional mean regression model does not provide a complete picture of the relationship between BMD and age in the sense that the change in the shape of the distribution of BMD is not captured in the model. Also, we observed that the typical constant variance assumption for the conditional mean does not hold. In Figure 1.4, we use the quantile regression to show how the shape of the distribution of BMD changes with respect to age. The intercept and slope of different quantiles of the BMD (Y) as a function of age (X) are shown in the Table 1.1. We also present the bootstrap confidence intervals and schematic boxplots of the estimates of the quantile regression coefficients at different choices of τ in Figures 1.5 and 1.6.

Table 1.1: Intercept and slope obtained from linear quantile regression between BMD and age in bone mineral acquisition dataset.

τ	Intercept	Slope
0.10	0.017	−0.002
0.25	0.057	−0.003
0.35	0.091	−0.005
0.50	0.110	−0.005
0.75	0.169	−0.007
0.95	0.286	−0.011

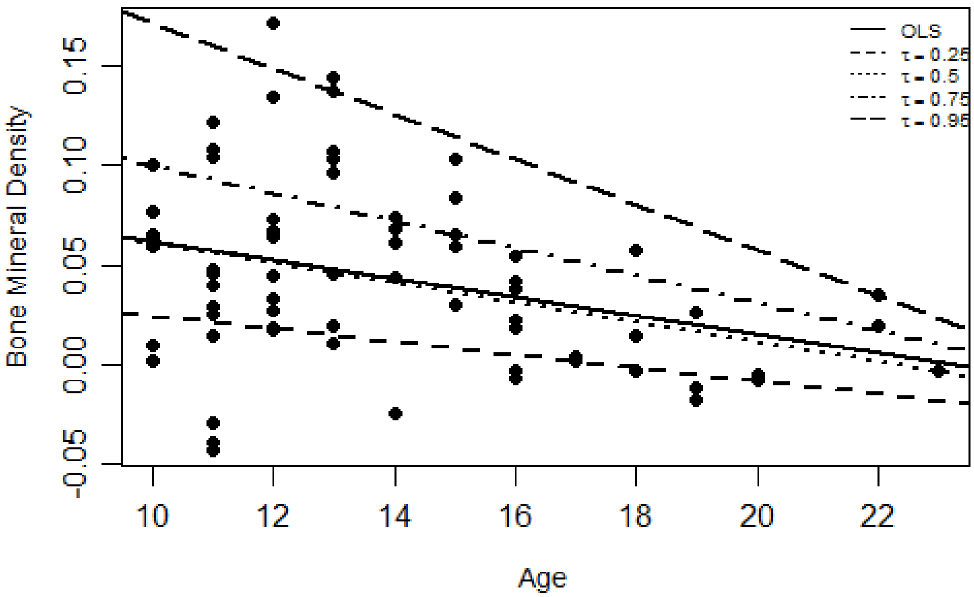

Figure 1.4: Bone mineral density as a function of the adolescent age for the baseline study with the least squares and quantile regression lines associated with different τ.

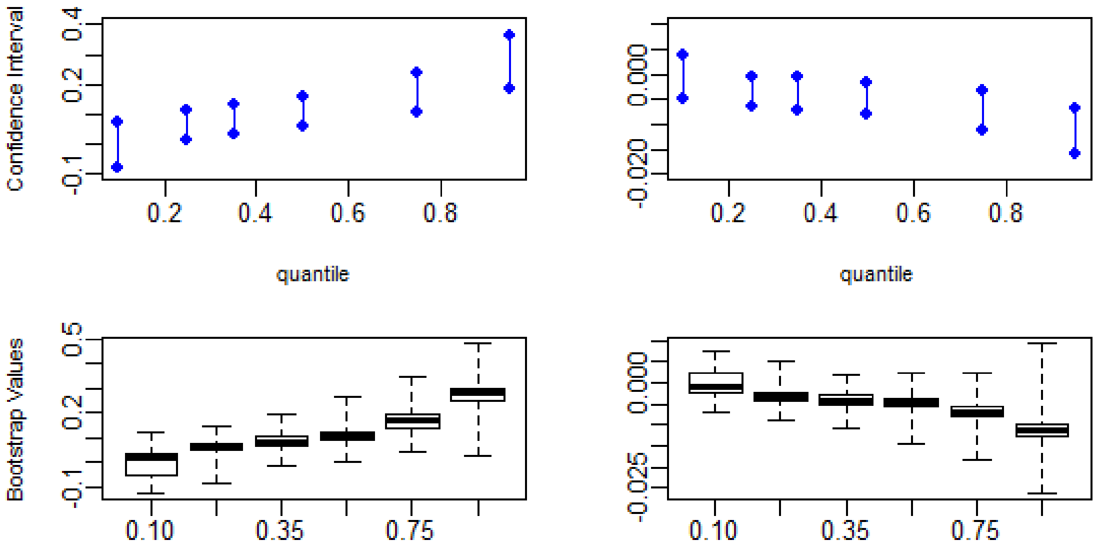

Figure 1.5: Bootstrap Confidence intervals and schematic boxplots of the estimates of the quantile regression coefficients at different quantiles $\tau = 0.10, 0.25, 0.35, 0.50, 0.75, 0.95$, Intercept(Left), slope(Right).

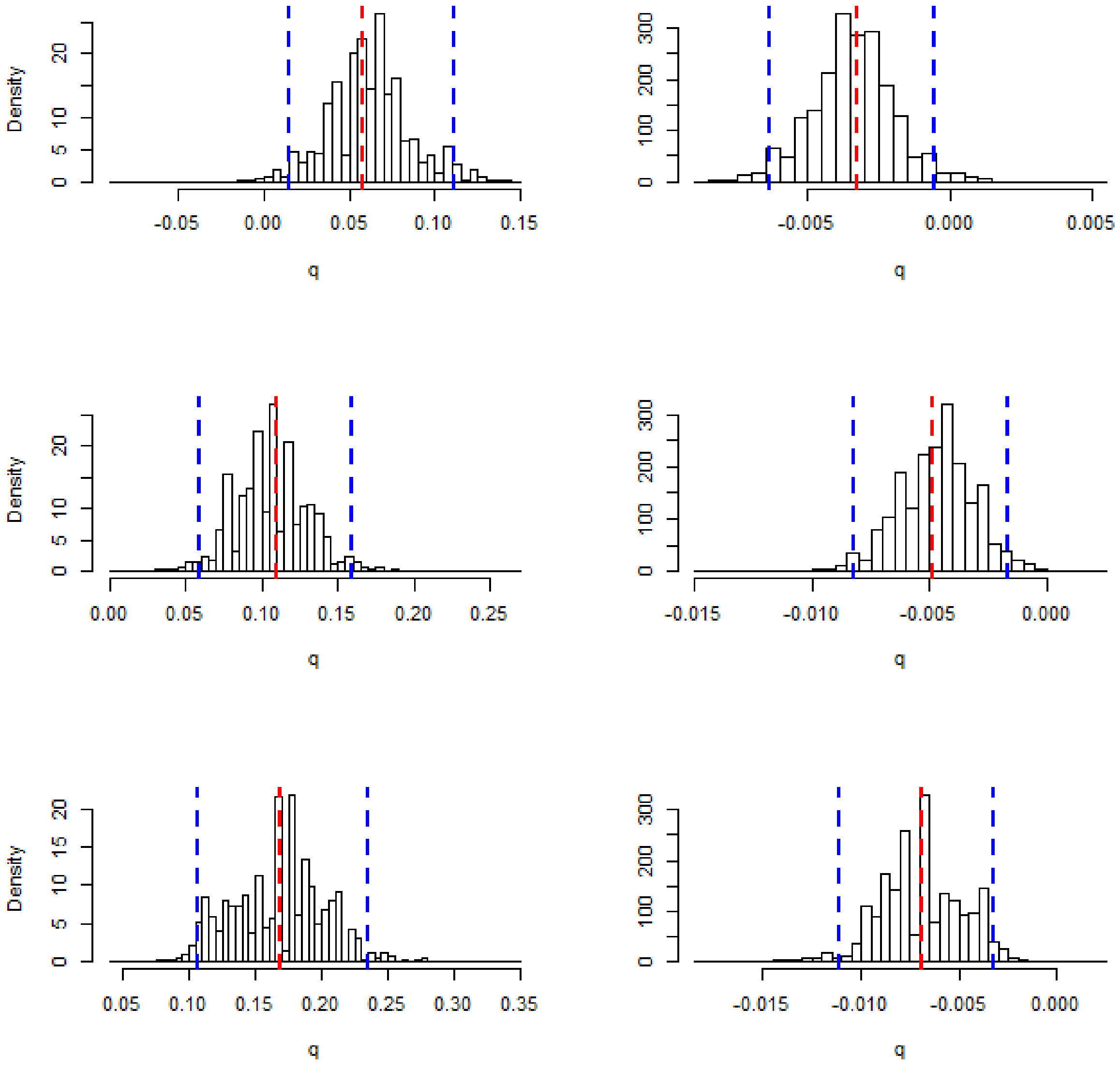

Figure 1.6: Distribution of the Bootstrap values of the regression coefficients and 95% C.I for quantiles at $\tau = 0.25, 0.50, 0.75$. Intercept(Left) and slope(Right)

1.11 Thesis Outline

The outline of the thesis is as follow. In chapter 2, we introduce the balanced ranked set sampling (RSS) and study some of its properties for estimating the population mean and quantiles. We also review some resampling methods to make inference about population parameters using the balanced RSS. In chapter 3 we propose a new objective function which takes into account the ranking information to estimate the unknown model parameters of quantile regression based on maxima and minima nominated samples. Also we conduct a simulation study to investigate the performance of the proposed objective function for minima and maxima-nominated samples for different quantiles. In Chapter 4, we present a real data application of our new method to a BMD dataset collected by the Manitoba Bone Density Program. Finally Chapter 5 provides a concluding remark and some future works.

Chapter 2

Ranked Set Sampling

In this chapter, we introduce the balanced and unbalanced ranked set sampling (RSS) designs and show how they can be used for the estimation of the population mean and quantiles. Also, we discuss resampling techniques for balanced and unbalanced ranked set samples using different algorithms.

2.1 Introduction

The idea of RSS was first proposed by McIntyre, G.A. (1952) in his effort to find a more efficient method to estimate the mean yield of pastures. As an alternative to simple random sampling (SRS), RSS has been proven to improve the efficiency of some statistical procedures in situations where the measurement of the variable of interest is destructive, difficult or expensive to obtain, but sampling units can be easily ordered by some means before taking the actual measurements on them. The improved efficiency of the statistical procedures based on RSS can be explained by its feature of assigning observations to homogeneous judgment order groups without complete measurement of the entire sample units. As a result, these judgment order groups act like strata in stratified sampling, and homogeneity in these strata leads to improvement in precision for many statistical procedures relative to their SRS counterparts.

In the literature, RSS has been widely studied for many purposes such as the one sample inference (McIntyre, G.A., 1952; Dell, T.R. and Clutter, J.L., 1972), estimation of the distribution function (Samawi,H. M. et al., 1996), one-way design layout (Muttlak, H. A., 1996), rank regression (Ozturk, O., 2002), confidence intervals for a population proportion (Terpstra and Wang, 2008), design-based estimation in finite populations (Jafari Jozani, M. and Johnson, B., 2011) and so on. In addition to these areas, Muttlak, H. A. (1995) studied the estimation of regression parameters in a simple linear regression model using RSS. Similarly, Philip, L.H. and Lam, K. (1997) studied regression type RSS estimators of the population mean for the response variable Y by utilizing the concomitant variable $\mathbf{X}$ in both the ranking process of the units and the estimation process. Barreto, M. C. M. and Barnett, V. (1999) studied the best linear unbiased estimators for the simple linear regression model using RSS. In this thesis, we study the estimation of the conditional quantile regression $Q_\tau(Y \mid \mathbf{X})$ using some variations of RSS with different set sizes and cycles.

The process of generating a ranked set sample involves randomly drawing k units (called a set of size k) from the underlying population and ranking them (ranking can be perfect or imperfect) using some means that do not require actual quantification of the units. For example, ranking can be done using concomitant variables or visual inspection. For this ranked set, the unit ranked the lowest is chosen for taking the actual measurement of the variable of interest. This initial measurement is called the first judgement order statistic and is denoted by $Y_{(1)}$. A second set of size k

(independent of the first set) is drawn and ranking is done as before. The unit in the second lowest position is then chosen and the variable of interest for this unit is measured and it is denoted by $Y_{(2)}$. This process continues until we select the unit ranked k from the k^{th} independent set and include its measurement, say $Y_{(k)}$, into our sample. This entire process is called a cycle and results in k measured observations $Y_{(1)}, \ldots, Y_{(k)}$. To complete one cycle of RSS, a total number of k^2 units from the underlying population is required. The measured observations $Y_{(1)}, \ldots, Y_{(k)}$ are called a balanced ranked set sample (BRSS) of size k which implies that one judgement order statistic has been collected for each of the ranks $1, 2, \ldots, k$. A desired BRSS of size $n = mk$ can be obtained by repeating the entire process for m independent cycles.

Another form of the RSS is the unbalanced ranked set sample (URSS). The structure of the URSS can be represented as follow;

$$\begin{array}{ccc}
\{Y_{(1)1}, \ldots, Y_{(1)n_1}\} & \overset{i.i.d.}{\sim} & F_{(1)}, \\
\{Y_{(2)1}, \ldots, Y_{(2)n_2}\} & \overset{i.i.d.}{\sim} & F_{(2)}, \\
\vdots & \vdots & \vdots \\
\{Y_{(k)1}, \ldots, Y_{(k)n_k}\} & \overset{i.i.d.}{\sim} & F_{(k)},
\end{array}$$

where all the $Y_{(i)j}$'s are independent and the $Y_{(i)j}$'s with the same i are identically distributed. Here, $Y_{(i)j}$ is the j^{th} observation $(j = 1, \ldots, n_i)$ from $F_{(i)}$, the distribution of the i^{th} order statistic in a set of size k, $i = 1, \ldots, k$. The sample can be considered as obtained from $n = \sum_{i=1}^{k} n_i$ sets of SRS units, each of size k, by first ranking the units of each set and then, for $i = 1, \ldots, k$, measuring the i^{th} ranked order statistic for n_i ranked sets. For example, consider an underlying distribution which is symmetric and unimodal about its median $Q_{0.5}$ and suppose our interest is to make inference about $Q_{0.5}$ using data collected by RSS design, where the set size k is supposed to be an odd number. Park, S. (1996) showed that among all the order statistics of a random sample of size k, the sample median $Y_{\frac{(k+1)}{2}}$ has the most information about η. Consequently, to estimate $Q_{0.5}$ in this case, it is natural to consider measuring the same order statistic, namely, the sample median $Y_{\frac{(k+1)}{2}}$, in each set, so that it is measured all k times in each of the m cycles. The resulting ranked set sample is an unbalanced ranked set sample of size mk measurements, each of which is a judgment median from a set of size k. For more details about the theory and application of URSS see Samawi,H. M. et al. (1996), Muttlak, H. A. (1998), Al-Omari, A.I. and Al-Nasser, A.D. (2012) and references there in.

RSS designs are efficient when the ranking process is free of error, that is, when we have perfect rankings. Perfect ranking with respect to the response Y is consistent, that is $F(y) = \frac{1}{k}\sum_{i=1}^{k} F_{(i)}(y)$ for all y. However, in many applications, it is difficult to have an error free ranking, thus it becomes essential to study the impact of imperfect ranking. We use $Y_{[i]}$ to show that the i^{th} judgment order statistic is obtained via imperfect ranking with corresponding CDF $F_{[i]}$. Imperfect ranking can arise as a result of ranking errors due to the use of auxiliary variables for the ranking of the variable of interest or the mis-match of the ranks and the real numerical orders when the ranks are induced by some other mechanism such as visual ranking and so on. Chen Zehua et al. (2004) showed that as long as the ranking in RSS is not done at random, RSS always provides more information about the underlying population than SRS even when the ranking is imperfect. Dell, T.R. and Clutter, J.L. (1972) reviewed the RSS concept with particular consideration of errors in judgment ordering. In their simulation study, it was assumed that observer ranks the elements on the basis of estimates that are equal to the true values plus random error component (an additive model). It was established that the RSS method yields an unbiased estimate provided that ranking errors are not related to the process of selecting elements for quantification.

For more details as well as several applications of RSS in reliability, environmental and medical studies see Chen Zehua et al. (2004).

2.2 Estimation of the Population Mean using RSS

Suppose $Y_1, \ldots, Y_n$ is random sample of size $n = mk$ collected using SRS design from a population with a continuous distribution function F, density function f and finite mean μ and variance σ^2. An unbiased estimator of the population mean μ is the sample mean $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ with $Var(\bar{Y}) = \frac{\sigma^2}{n}$.

Suppose $\{Y_{[1]i}, \ldots, Y_{[k]i}, i = 1, \ldots, m\ \}$ represents a BRSS of size $n = km$ from F where m is the cycle size and k is the set size. The mean of this sample is given by

$$\bar{Y}_{RSS} = \frac{1}{km}\sum_{j=1}^{m}\sum_{i=1}^{k} Y_{[i]j}. \tag{2.1}$$

Consider only the case of one cycle ($m = 1$) for simplicity, so that the total sample size n is equal to the set size k. Under the perfect ranking assumption, we can represent the RSS observations for this setting by $Y_{(1)}, \ldots, Y_{(k)}$ where these k variables are mutually independent and $Y_{(i)}$, $i = 1, \ldots, k$ is distributed as the i^{th} order statistic of a random sample of size k from F. It follows that

$$E[\bar{Y}_{RSS}] = \frac{1}{k} \sum_{i=1}^{k} E[Y_{(i)}], \tag{2.2}$$

where

$$E[Y_{(i)}] = \int_{-\infty}^{\infty} y \frac{k!}{(i-1)!(k-i)!} [F(y)]^{i-1} [1 - F(y)]^{k-i} f(y)\, dy, \tag{2.3}$$

for $i = 1, \ldots, k$. Combining (2.2) and (2.3), we obtain

$$\begin{aligned} E[\bar{Y}_{RSS}] &= \frac{1}{k} \sum_{i=1}^{k} \left[\int_{-\infty}^{\infty} ky \binom{k-1}{i-1} [F(y)]^{i-1} [1 - F(y)]^{k-i} f(y)\, dy \right] \\ &= \int_{-\infty}^{\infty} y f(y) \left[\sum_{i=1}^{k} \binom{k-1}{i-1} [F(y)]^{i-1} [1 - F(y)]^{k-i} \right] dy. \end{aligned} \tag{2.4}$$

setting $z = i - 1$ in (2.4), we get

$$\sum_{i=1}^{k} \binom{k-1}{i-1} [F(y)]^{i-1} [1 - F(y)]^{k-i} = \sum_{z=0}^{k-1} \binom{k-1}{z} [F(y)]^{z} [1 - F(y)]^{(k-1)-z} \quad = 1. \tag{2.5}$$

Substituting (2.5) in (2.4), (2.4) becomes

$$E[\bar{Y}_{RSS}] = \int_{-\infty}^{\infty} y f(y)\, dy = \mu, \tag{2.6}$$

showing that $\bar{Y}_{RSS}$ is an unbiased estimator of the population mean μ.

The mutual independence of the $Y_{(i)}$'s, $i = 1, \ldots, k$ enables us to write the variance of the RSS estimator as (Wolfe, D.A., 2012)

$$Var[\bar{Y}_{RSS}] = \frac{1}{k^2}\sum_{i=1}^{k} Var(Y_{(i)}). \tag{2.7}$$

Letting $\mu_{(i)} = E[Y_{(i)}]$, for $i = 1, \ldots, k$, we get

$$E[(Y_{(i)} - \mu)^2] = E[(Y_{(i)} - \mu_{(i)} + \mu_{(i)} - \mu)^2] = E[(Y_{(i)} - \mu_{(i)})^2] + (\mu_{(i)} - \mu)^2 \tag{2.8}$$

$$= Var(Y_{(i)}) + (\mu_{(i)} - \mu)^2.$$

Combining (2.7) and (2.8) yields the expression

$$Var(\bar{Y}_{RSS}) = \frac{1}{k^2}\sum_{i=1}^{k} E[Y_{(i)} - \mu]^2 - \frac{1}{k^2}\sum_{i=1}^{k}(\mu_{(i)} - \mu)^2. \tag{2.9}$$

$$= \frac{\sigma^2}{k} - \frac{1}{k^2}\sum_{i=1}^{k}(\mu_{(i)} - \mu)^2,$$

$$= Var(\bar{Y}) - \frac{1}{k^2}\sum_{i=1}^{k}(\mu_{(i)} - \mu)^2$$

$$\leq Var(\bar{Y}),$$

since $\sum_{i=1}^{k}(\mu_{(i)} - \mu)^2 \geq 0$. In the case of perfect ranking not only is $\bar{Y}_{RSS}$ an unbiased estimator, but also its variance is always less than the variance of the SRS estimator $\bar{Y}$ considering equal sample sizes. For similar results under imperfect ranking see Chen Zehua et al. (2004).

2.3 Quantile Estimation using RSS

In this section, we define the sample quantiles for balanced and unbalanced ranked set samples analogous to the simple random sample quantiles and investigate their properties. Let $Y_1, \ldots, Y_n$ be a simple random sample of size n from F, with an empirical distribution function given by $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq t)$. The τ^{th} quantile, Q_τ, is then estimated by $\hat{Q}_{n\tau,SRS} = \inf\{t : \hat{F}_n(t) \geq \tau\}$.

Now, consider a BRSS of size $n = mk$ denoted by $Y_{RSS} = \{Y_{[i]j}, i = 1, \ldots, k, j = 1, \ldots, m\}$. Let the empirical distribution function of Y_{RSS} be defined as

$$\widehat{F_B}(t) = \frac{1}{km}\sum_{j=1}^{m}\sum_{i=1}^{k} I(Y_{[i]j} \leq t).$$

For $0 < p < 1$, the τ^{th} ranked set sample quantile, denoted by $\hat{Q}_{n\tau,BRSS}$, is defined as the τ^{th} quantile of $\widehat{F}_B$, i.e.,

$$\hat{Q}_{n\tau,BRSS} = \inf\{t : \widehat{F}_B(t) \geq \tau\}.$$

Some useful theoretical results in estimating population quantiles using ranked set samples are shown in Chen Zehua et al. (2004). Below we provide some of these properties while proofs can be found in Chen Zehua et al. (2004).

Theorem 2.3.1. *Suppose that the ranking mechanism in RSS is consistent. Then, with probability 1,*

$$|\hat{Q}_{n\tau,BRSS} - Q_\tau| \leq \frac{2(log\, n)^2}{f(Q_\tau)n^{1/2}},$$

for sufficiently large n, when $n = mk$ and k is fixed.

Theorem 2.3.2. *Suppose that the ranking mechanism in RSS is consistent and that the density function f is continuous at Q_τ and positive in a neighborhood of Q_τ.*

Then, for large n, with $n = mk$ and fixed k,

$$\hat{Q}_{n\tau,BRSS} = Q_\tau + \frac{\tau - \hat{F}_B(Q_\tau)}{f(Q_\tau)} + R_n,$$

where, with probability one,

$$R_n = O(n^{-3/4}(logn)^{3/4}).$$

Using theorem 2.3.2 the asymptotic normality of the ranked set sample quantile follows from the following result.

Theorem 2.3.3. *Under the conditions of theorem 2.3.2, as $m \to \infty$,*

$$\sqrt{n}(\hat{Q}_{n\tau,BRSS} - Q_\tau) \xrightarrow{D} N\left(0, \frac{\sigma^2_{k,\tau}}{f^2(Q_\tau)}\right),$$

where,

$$\sigma^2_{k,\tau} = \frac{1}{k}\sum_{i=1}^{k} F_{[i]}(Q_\tau)[1 - F_{[i]}(Q_\tau)].$$

In particular, if ranking is perfect, noting that $F_{[i]}(Q_\tau) = B(i, k+i-1, \tau)$, we have

$$\sigma^2_{k,\tau} = \frac{1}{k}\sum_{i=1}^{k} B(i, k+i-1, \tau)[1 - B(i, k+i-1, \tau)],$$

where $B(i, s, x) = \int_0^x t^{i-1}(1-t)^{s-1}\,dt$, $i > 0, s > 0$ denotes the distribution function of a beta distribution with parameters i and s.

The quantity $\sigma^2_{k,\tau}$ does not depend on any unknowns when ranking is perfect. In general, $F_{[i]}(Q_\tau)$ depends on both the ranking mechanism and the unknown F, that are needed to be estimated from the data.

Theorem 2.3.4. *Let $0 < \tau_1 < \ldots < \tau_j < \tau_l < 1$ be l probabilities. Let $\psi = (Q_{\tau_1}, \ldots, Q_{\tau_l})^\top$ and $\hat{\psi} = (\hat{Q}_{\tau_1,BRSS}, \ldots, \hat{Q}_{\tau_l,BRSS})^\top$. Then, as $n \to \infty$, with $n = mk$ and fixed k,*

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{D} N_l(0, \Sigma),$$

where, for $r \le j$, the $(i, j)^{th}$ entry of Σ is given by

$$\sigma_{rj} = \frac{1}{k} \sum_{i=1}^{k} F_{[i]}(Q_{\tau_r})[1 - F_{[i]}(Q_{\tau_j})]/[f(Q_{\tau_r})f(Q_{\tau_j})].$$

Theorem 2.3.5. *(i) If $\frac{k_n}{n} = \tau + o(n^{-1/2})$ then*

$$Z_{(k_n:n)} = Q_\tau + \frac{\frac{k_n}{n} - \hat{F}_B(Q_\tau)}{f(Q_\tau)} + R_n,$$

where, with probability 1,

$$R_n = O(n^{-3/4}(log\, n)^{3/4}),$$

as $n \to \infty$.

(ii) If $\frac{k_n}{n} = \tau + \frac{c}{n^{1/2}} + o(n^{-1/2})$ then,

$$\sqrt{n}(Z_{(k_n:n)} - \hat{Q}_\tau) \xrightarrow{D} N\left(\frac{c}{f(Q_\tau)}, \frac{\sigma^2_{k,\tau}}{f^2(Q_\tau)}\right).$$

The asymptotic relative efficiency (ARE) of $\hat{Q}_{n\tau,BRSS}$ with respect to $\hat{Q}_{n\tau,SRS}$ is given by

$$ARE(\hat{Q}_{n\tau,BRSS}, \hat{Q}_{n\tau,SRS}) = \frac{\tau(1-\tau)}{\frac{1}{k}\sum_{i=1}^{k} F_{[i]}(Q_\tau)[1 - F_{[i]}(Q_\tau)]},$$

where $\hat{Q}_{n\tau,SRS}$ is the τ^{th} sample quantile in SRS. Chen Zehua et al. (2004) showed that $ARE(\hat{Q}_{n\tau,BRSS}, \hat{Q}_{n\tau,SRS})$ is always greater than 1 for any $\tau \in (0,1)$.

For an unbalanced RSS $Y_{URSS} = \{Y_{[i]j}, i = 1, \ldots, k,\ j = 1, \ldots, n_i\}$ with $n = \sum_{j=1}^{k} n_j$, the empirical distribution function of Y_{URSS} is defined as

$$\hat{F}_U(t) = \frac{1}{n}\sum_{j=1}^{n_i}\sum_{i=1}^{k} I(Y_{(i)j} \le t) = \sum_{i=1}^{k} q_{n_i}\hat{F}_{(i)}(t),$$

where $q_{n_i} = \frac{n_i}{n}$, $q_{n_i} = (q_{n1}, \ldots, q_{nk})^\top$ and $\hat{F}_{(i)}(t) = \frac{1}{n}\sum_{j=1}^{n_i} I(Y_{(i)j} \le t)$. For $0 < \tau < 1$, the τ^{th} unbalanced ranked set sample quantile, denoted by $\hat{Q}_{n\tau,URSS}$, is defined as the τ^{th} quantile of $\hat{F}_U$, i.e.,

$$\hat{Q}_{n\tau,URSS} = \inf\{t : \hat{F}_U(t) \ge \tau\}.$$

Suppose that, as $n \to \infty, q_{n_i} \to q_i, i = 1, \ldots, k$. Let $F_U = \sum_{i=1}^{k} q_i F_{(i)}$, where F_U is the distribution function.

Chen Zehua et al. (2004) showed the asymptotic properties of the unbalanced ranked set sample quantile using the following theorems.

Theorem 2.3.6. *(i) (The strong consistency). With probability 1,* $\hat{Q}_{n\tau,URSS}$ *converges to* Q_τ. *(ii) Suppose that* $q_{n_i} = q_i + O(n^{-1})$. *If* f_u *is continuous at* Q_τ *and* *positive in a neighborhood of* Q_τ*, then*

$$\hat{Q}_{n\tau,URSS} = Q_\tau + \frac{\tau - \hat{F}_U(Q_\tau)}{f_u(Q_\tau)} + R_n,$$

where, with probability one,

$$R_n = O(n^{-3/4}(log\, n)^{3/4}),$$

as $n \to \infty$*. (iii) (The Asymptotic Normality). Suppose that* $q_{n_i} = q_i + O(n^{-1})$*. If* f_u *is continuous at* Q_τ *and positive in a neighborhood of* Q_τ*, then,*

$$\sqrt{n}(\hat{Q}_{n\tau,URSS} - Q_\tau) \xrightarrow{D} N\left(0, \frac{\sigma^2_{u_\tau}}{f_u^2(Q_\tau)}\right),$$

where

$$\sigma^2_{u_\tau} = \sum_{i=1}^{k} q_i F_{(i)}(Q_\tau)[1 - F_{(i)}(Q_\tau)].$$

The above results hold in both perfect and imperfect ranking cases.

2.4 Resampling Ranked Set Samples

So far in this chapter, we have presented the asymptotic results for the balanced and unbalanced ranked set samples with focus on inferences about the population characteristics under specific parametric assumptions. However, in applications with small samples, which often happens when dealing with expensive measurements, asymptotic inference for complex statistics, say, $\hat{\psi}_n$ may not be valid with sampling distribution $G_{n,F}(t)$; their standard errors may not be known which will make it difficult or impossible to construct confidence intervals. Therefore, other means such as resampling techniques are used to make inference in such situations. Resampling is a viable approach to obtain sampling distribution of a statistic of interest. In particular, the bootstrap approach is used for the estimation of the standard error of any well-defined statistic and allows one to draw inferences when the exact or asymptotic distribution of the statistic of interest is unavailable or difficult to obtain. In this section, we describe methods of bootstrapping BRSS using three different algorithms: BRSSR (bootstrap RSS row-wise), BoRSS (bootstrap RSS), and MRBRSS (mixed row bootstrap RSS). Also we discuss the methods of bootstrapping URSS as presented in Amiri, S. et al. (2013).

2.4.1 BRSSR: Bootstrap RSS by Rows

Chen Zehua et al. (2004) introduced the method of bootstrapping ranked set samples row-wise (BRSSR) by considering a ranked set sample, with each row being i.i.d. samples from the distribution function $F_{(i)}$ of the i^{th} order statistic; i.e., $F_{(i)}(y) = P(Z_{(i)} \leq y)$ where $Z_{(i)}$ is the i^{th} order statistic from F. Chen Zehua et al. (2004) showed that as the distribution function F can be represented by $F(y) = \frac{1}{k}\sum_{i=1}^{k} F_{(i)}(y)$ for each i, each $F_{(i)}$ can be estimated using the observed data on the i^{th} order statistic by $F_{(i),m}(t) = \frac{1}{m}\sum_{j=1}^{m} I(Y_{(i)j} \leq t)$. The BRSSR method resamples from each $F_{(i),m}(y)$ independently and then combines the obtained samples to form a bootstrap sample. The algorithm is as follow;

1. Assign to each element of the i^{th} row a probability of $\frac{1}{m}$ and select m elements randomly from $F_{(i),m}(y)$ with replacement to obtain $Y^*_{(i)1}, \ldots, Y^*_{(i)m}$.

2. Repeat step 1 for $i = 1, \ldots, k$ to obtain a bootstrap ranked set sample $Y^*_{(i)j}$.

3. For each $i = 1, \ldots, k$, define

$$F^*{}_{(i),m}(t) = \frac{1}{m}\sum_{j=1}^{m} I(Y^*_{(i)j} \leq t) \quad \text{and} \quad F^*{}_n(y) = \frac{1}{k}\sum_{i=1}^{k} F^*{}_{(i),m}(t).$$

Note that this method cannot be applied when $m = 1$ as the bootstrap empirical distribution is degenerate.

2.4.2 BoRSS : Bootstrap RSS

The bootstrap RSS as described by Modarres, R. et al. (2006) generally performs better than SRS of the same size. It draws ranked set samples from F_n. The algorithm is as follow;

1. Assign to each element of the ranked set sample a probability of $1/mk$.

2. Randomly draw k elements $Y_1, \ldots, Y_k \overset{i.i.d.}{\sim} F_n$, sort them in ascending order $Y_{(1)} \leq \ldots \leq Y_{(k)}$, retain $Y^*_{(i)} = Y_{(i)}$.

3. Perform step 2 for $i = 1, \ldots k$.

4. Repeat steps 2 and 3, m times, to obtain $Y^*_{(i)j}$.

5. Define the bootstrap empirical distribution by

$$F^*{}_n(t) = \frac{1}{km}\sum_{j=1}^{m}\sum_{i=1}^{k} I(Y^*_{(i)j} \leq t).$$

Note, when $k = 1$, this method reduces to the usual SRS bootstrap and this method is valid for all $m \geq 1$.

2.4.3 MRBRSS : mixed row bootstrap RSS

The MRBRSS as proposed by Modarres, R. et al. (2006) differs from other resampling approach because it draws more stratified units than BRSSR by taking into account some partial ordering information in RSS which are not considered in the row-wise resampling. The MRBRSS method resamples from the entire F_n. The resampling algorithm is as follow;

1. Assign to each element of the i^{th} row a probability of $\frac{1}{m}$ for $i = 1, \ldots, k$ and randomly select one element from each row to obtain $y_1^*, \ldots, y_k^*$.

2. Sort $Y_1^*, \ldots, Y_k^*$ in ascending order to get $Y_{(1)} \leq \ldots \leq Y_{(k)}$ and retain $Y_{(i)}^* = Y_{(i)}$.

3. Perform steps 1 and 2 for $i = 1, \ldots, k$ to obtain $Y_{(1)1}^*, \ldots, Y_{(k)1}^*$.

4. Repeat steps 1-3, m times, to obtain $Y_{(i)j}^*$.

5. Define the bootstrap empirical distribution by

$$F^*{}_n(t) = \frac{1}{km} \sum_{j=1}^{m} \sum_{i=1}^{k} I(Y_{(i)j}^* \leq t).$$

The statistical procedure of SRS that are extended to balanced RSS cannot directly be applied to URSS due to the complication in the design, as a result of the absence of the balanced structure. Amiri, S. et al. (2013) described one possible approach to avoid this difficulty by transforming the URSS to a BRSS using a transformation (BTR) that involves an initial step of resampling within strata to create a balanced RSS. This BTR is included in each bootstrap run as part of new bootstrap algorithms that are designed to allow for URSS. Once the bootstrap transformation of the URSS to BRSS is achieved, one can use the bootstrap techniques discussed above.

Chapter 3

Quantile Regression with Nominated Samples

In this chapter, we estimate the quantile regression using the nomination sampling technique, which is a variation of the ranked set sampling. To this end, we propose a new objective function which takes into account the ranking information to estimate the unknown model parameters based on data that are obtained from a maxima nomination sampling (MNS). Our method is built upon the distribution of the maximum order statistic from an asymmetric Laplace distribution, which serves as the link between the minimization of the sum of deviations and the maximum likelihood method. We show how the results can be extended to a minima nomination sample. Finally, we perform simulation studies to compare the performance of our proposed method with that of SRS for estimating the upper and lower quantiles of the variable of interest.

3.1 Introduction

The concept of nomination sampling (NS) design was first proposed by Willemain, T.R. (1980) in his effort to study new ways to pay for nursing home services. Willemain, T.R. (1980) presented an application where is practically impossible to select samples using SRS that can serve as a representative sample from the underlying population. In this application, the federal officials were interested in relating a facility's rate of reimbursement to the typical level of debility of its residents so that payment might better match the costs of care. Willemain, T.R. (1980) introduced NS design and showed that his proposed sampling design not only was advantageous for political reasons but also provided an improvement in the distribution of estimation error. NS is a sampling process in which every observation is the maximum or the minimum (depending on the research interest) of a random sample set from the population of interest. Let $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{ik_i}, i = 1, \ldots, n)$, with $Y_{i1}, \ldots, Y_{ik_i}$ be a random sample of size k_i from a distribution with continuous distribution function. Define the map $\Psi_i : \mathrm{R}^{k_i} \rightarrow \mathrm{R}$ such that Ψ_i maps $\mathbf{Y_i}$ into a particular element in $\mathbf{Y_i}$, say Y_i. Then, Y_i is called the nominee of $\mathbf{Y_i}$ and the collection of $\{Y_i, i = 1, \ldots, n\}$ is called the nomination sample. For example, consider an experiment in which only one observation is available from each of the N populations. Assume that observation from the i^{th} population is in fact the maximum or the minimum of a random sample of size k from that population, this experiment is clearly a NS.

In several applications such as economics, environmental studies and medical research, focus is often on the description of the upper or lower quantiles as well as extremes of the study population. For instance, in medical studies, exposure to infectious diseases varies from one patient to the other, patients with high exposure to the disease are most likely the first set of patients to receive treatments, therefore the nominee could be those who have the maximum exposure to the disease. Another example is in consumer behavior studies, where consumers have a known number of options from which a single decision is required. Consumers will usually choose the option that costs the least and hence the nominee will be the option with the minimal cost.

Over the past few years, NS has gained a reasonable attention in the literature. For example, NS has been used in the estimation of the population median (Willemain, T.R., 1980), the distribution function (Boyles, R.A. and Samaniego, F.J., 1986), and in quality control charts for attributes (Jafari Jozani, M. and Mirkamali, S.J., 2011). In all these application, $\Psi_i(\mathbf{Y_i}) = \max_{1\leq j\leq k_i} Y_{ij}$ was used and the population was also assumed to be infinite or very large. Wells, M.T. and Tiwari, R.C. (1990), estimated the distribution function in the case where $\Psi_i(\mathbf{Y_i}) = \min_{1\leq j\leq k_i} Y_{ij}$ for an infinite population. Jafari Jozani, M. and Johnson, B. (2012) introduced a randomized minima-maxima nomination sampling design, which is a random combination of the minima and maxima NS in a finite population and results in both minima and maxima NS as special cases.

In the next section, we present some preliminary results of quantile regression and a parametric link between the minimization of the sum of absolute deviations and the maximum likelihood approach. Then, we present different methods of estimating the τ^{th} regression quantile parameter using MNS data. Also we present comprehensive simulation study to evaluate the performance of our proposed methods in the subsequent section. We show how the results can be extended to a minimum nomination sampling. Similar to Nourmohammadi, M. et al. (2015) we observe that one can better model the upper or lower quantiles of the study variable using the maxima or minima NS compared with SRS. This will help to reduce the number of required observations using NS designs to achieve the same efficiency as in SRS. Hence, when there is a budget constraints NS could be used as a more economical rival to SRS.

3.2 Preliminary Results

In chapter 1, where the distributional form of the response variable Y is unknown, we defined the τ^{th} regression quantile as a solution to the equation (1.10) given as;

$$\min_{\boldsymbol{\beta} \in R^p} \left[\tau \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| + (1-\tau) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| \right]. \tag{3.1}$$

which is an optimization problem. We presented the solution to this problem using the simplex method. The method described in Chapter 1 for estimating the conditional mean regression is called least squares approach. Another method to estimate the unknown parameter in the conditional mean regression model is the maximum likelihood estimation (MLE) approach, where the error term (ϵ) is assumed to follow a normal distribution, $\epsilon \sim N(0, \sigma^2)$. The τ^{th} regression quantile can also be considered as the MLE by assuming a distribution for the response variable.

A possible distribution which connects the minimization of (3.1), the weighted sum of the absolute deviations and the maximum likelihood approach is given by the asymmetric Laplace distribution (ALD). ALD is a skewed distribution which has been extensively used to study quantile regression, see for example Koenker, R. and Machado, J.A.F. (1999) and Yu, K. and Moyeed, R.A. (2001). A continuous random variable Y is said to follow an ALD with parameters (μ, σ, τ) and is denoted by $Y \sim AL(\mu, \sigma, \tau)$, if its probability density function (pdf) can be expressed as:

$$f(y \mid \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{3.2}$$

where $\rho_\tau(U) = (\tau - I(U \leq 0))U$, $0 < \tau < 1$ is the skewness parameter, $\sigma > 0$ is the scale parameter and $\mu \in R$ is the location parameter. $AL(\mu, \sigma, \tau)$ is right skewed when $\tau < 0.5$ and left skewed when $\tau > 0.5$. When $\tau = 0.5$, we have the Laplace (double exponential) distribution which has the pdf of the form

$$f(y \mid \mu, \sigma, 0.5) = \frac{1}{4\sigma} \exp\left\{-\frac{|y-\mu|}{2\sigma}\right\}, y \in \mathbb{R}, y \in \mathbb{R}, \sigma > 0.$$

The CDF and the quantile functions of $Y \sim AL(\mu, \sigma, \tau)$ are given respectively as;

$$F(y) = \begin{cases} \tau \exp(-(\tau-1)(\frac{y-\mu}{\sigma})), & \text{if } y < \mu, \\ \\ 1-(1-\tau)\exp(-(\frac{y-\mu}{\sigma})\tau), & \text{if } y > \mu, \end{cases} \tag{3.3}$$

and

$$F^{-1}(y) = \begin{cases} \mu + \frac{\sigma}{1-\tau}\log(\frac{y}{\tau}), & \text{if } 0 \leq y < \tau, \\ \\ \mu - \frac{\sigma}{\tau}\log(\frac{1-y}{1-\tau}), & \text{if } \tau < y \leq 1. \end{cases} \tag{3.4}$$

The mean and the variance of Y are respectively given by $E(Y) = \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)}$ and $Var(Y) = \frac{\sigma^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}$ (Yu, K. and Zhang, J., 2005).

From equation (1.1), let $\mu_i = \mathbf{X}_i^\top \boldsymbol{\beta}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. Assume that $Y_i \sim AL(\mathbf{X}_i^\top \boldsymbol{\beta}, \sigma, \tau)$, then the likelihood for n independent observations can be written as

$$L_\tau(\boldsymbol{\beta}, \sigma) = \left(\frac{\tau(1-\tau)}{\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n} \rho_\tau \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right\}. \tag{3.5}$$

The maximization of equation (3.5) w.r.t $\boldsymbol{\beta}$ is equivalent to the minimization of equation (3.1) when σ is considered as a nuisance parameter (normally taken to be $\sigma = 1$). In the literature, QR methods based on the ALD has been widely used in longitudinal data analysis (Geraci, M. and Bottai, M., 2006; Farcomeni, A, 2012), and in Bayesian analysis (Yu, K. and Moyeed, R.A., 2001; Yu, K. and Stander, J., 2007). Bera, A. K. et al. (2016) studied the connections between the asymmetric Laplace pdf, maximum likelihood, maximum entropy and quantile regression. In

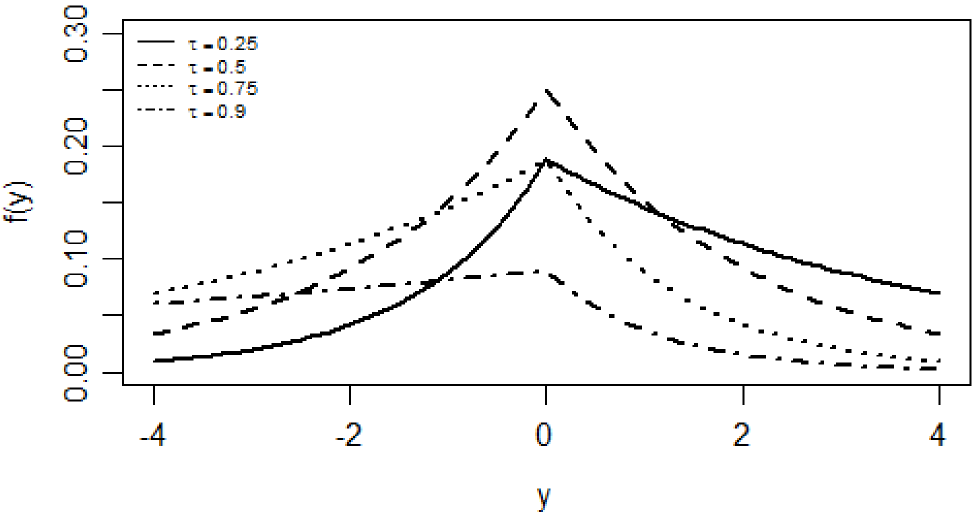

Figure 3.1: Asymmetric Laplace densities with $\tau \in (0.25, 0.5, 0.75, 0.90)$, $\sigma = 1$ and $\mu = 0$

this thesis, we use the ALD to propose an objective function for QR using maxima or minima nominated samples. The proposed method can be easily extended to any rank-based sampling design. However, the difficulty resides in theoretical derivation of the properties of the resulting estimators. The proposed objective function is derived and explained in the next section.

3.3 Estimation

In this section, we present two different methods for estimating the τ^{th} regression quantile parameters using MNS data. The first method is based on the distribution of the maximum order statistic from an ALD. The second method extends Nourmohammadi, M. et al. (2015) and is based on the relationship between SRS and MNS data in estimating τ^{th} regression quantile parameters. This method is implemented by selecting samples from the population using MNS approach. The maxima nominated samples are then treated as simple random samples using the relationship between MNS and SRS as shown in Nourmohammadi, M. et al. (2015).

3.3.1 First Method

Without loss of generality, we assume that set sizes k_i are all the same, Let Y_i be the maximum of the sample $\mathbf{Y_i}$ of size k, $i = 1, \ldots, n$. Then Y_i has a distribution function of $G_Y(y) = [F_Y(y)]^k$ and density $g_Y(y) = kf(y)[F_Y(y)]^{k-1}$ for all $y \in R$. The joint density for n independent copies of Y_i can be written as

$$\prod_{i=1}^{n} kf(y_i)[F_Y(y_i)]^{k-1}. \qquad (3.6)$$

Suppose it is known that $Y \sim AL(\mathbf{X}^\top \boldsymbol{\beta}, \sigma, \tau)$, using equation (3.3) we write the distribution function as

$$F(y; \boldsymbol{\beta}, \sigma, \tau) = \begin{cases} \tau \exp\{-(\frac{y - \mathbf{X}^\top \boldsymbol{\beta}}{\sigma})(\tau - 1)\}, & \text{if } y < \mathbf{X}^\top \boldsymbol{\beta}, \\ \\ 1 - (1-\tau) \exp\{-(\frac{y - \mathbf{X}^\top \boldsymbol{\beta}}{\sigma})\tau\}, & \text{if } y > \mathbf{X}^\top \boldsymbol{\beta}. \end{cases} \tag{3.7}$$

Using (3.7) and (3.6) the likelihood function is given by

$$\begin{aligned} &L_\tau(\boldsymbol{\beta}, \sigma) \\ &= \prod_{i=1}^{n} \left[\frac{k\tau(1-\tau)}{\sigma} \exp\left(-(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma})(\tau - 1) \right) \left[\tau \exp\left(-(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma})(\tau - 1) \right) \right]^{k-1} \right]^{I(Y_i < \mathbf{X}_i^\top \boldsymbol{\beta})} \\ &\times \prod_{i=1}^{n} \left[\frac{k\tau(1-\tau)}{\sigma} \exp\left(-(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma})\tau \right) \left[1 - (1-\tau) \exp\left(-(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma})\tau \right) \right]^{k-1} \right]^{I(Y_i > \mathbf{X}_i^\top \boldsymbol{\beta})}. \end{aligned} \tag{3.8}$$

Now, the log-likelihood function is given by;

$$\begin{aligned} &l_\tau(\boldsymbol{\beta}) \\ &= n \ln\left(\frac{k\tau(1-\tau)}{\sigma} \right) - \sum_{\{i: Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} \left[(\tau - 1) \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) - (k-1) \left(\ln \tau - (\tau - 1) \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right) \right] \\ &- \sum_{\{i: Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} \left[\tau \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) - (k-1) \ln\left(1 - (1-\tau) \exp\left(-\tau(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma}) \right) \right) \right]. \end{aligned} \tag{3.9}$$

From equation (3.9), we define the τ^{th} regression quantile using MNS as

$Q_\tau(Y \mid \mathbf{X}) = \mathbf{X}^\top \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is obtained as the solution of

$$\begin{aligned} &\max_{\boldsymbol{\beta} \in R^p} \left[n \ln\left(\frac{k\tau(1-\tau)}{\sigma} \right) - \sum_{\{i: Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} \left[(\tau - 1) \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) - (k-1) \left(\ln \tau - (\tau - 1) \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right) \right] \right. \\ &\left. - \sum_{\{i: Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} \left[\tau \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) - (k-1) \ln\left(1 - (1-\tau) \exp\left(-\tau \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right) \right) \right] \right], \end{aligned} \tag{3.10}$$

or equivalently,

$$\min_{\boldsymbol{\beta}\in R^p}\left[\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)-(k-1)\left(\ln\tau-(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right)\right]\right.$$
$$\left.+\sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[\tau\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)-(k-1)\ln\left(1-(1-\tau)\exp\left(-\tau\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right)\right)\right]\right], \tag{3.11}$$

which again is an optimization problem. Equation (3.10) can therefore be solved using any appropriate optimization method. However, optimizing equation (3.10) using the simplex method is difficult because it is unclear how to define the constraints with the set size k. To obtain the maximum likelihood estimate of the unknown parameter for the τ^{th} regression quantile using maxima nominated samples we adopt the idea of M-Quantiles approach introduced by Breckling, J. and Chambers, R. (1988). The M-quantile approach is an extension of the M-estimation of Huber, P. J. (1964) to a wide range of the location parameters of the population of interest. To solve (3.11), we use the M-estimation approach. First, note that minimizing (3.11) in $\boldsymbol{\beta}^p$ does not depend on the first term $n\ln\left(\frac{k\tau(1-\tau)}{\sigma}\right)$. Let $\epsilon_i = Y_i - \mathbf{X}_i^\top\hat{\boldsymbol{\beta}}$ be the residual, where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. Using the M-estimation approach, $\hat{\boldsymbol{\beta}}$ is defined by minimizing the following objective function over all $\boldsymbol{\beta}^p$

$$\sum_{i=1}^{n}\rho\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right),$$

with $\rho(U) = -[(\tau-1)U-(k-1)\ln\tau-(\tau-1)U]I(U<0)+[\tau U-(k-1)\ln(1-(1-\tau)\exp(-\tau U)]I(U>0)$, where the function $\rho(.)$ gives the contribution of each residual to the objective function. Let $\psi=\rho'$ be the derivative of ρ, which is also known as the influence curve. Differentiating the objective function w.r.t $\boldsymbol{\beta}$ and setting the partial derivatives to 0 produces a system of equations for the coefficients

$$\sum_{i=1}^{n} \psi \left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \mathbf{X}_i^\top = 0. \tag{3.12}$$

Let $w(t) = \frac{\psi(t)}{t}$ and $w_i = w(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$. To estimate $\boldsymbol{\beta}$, we first write (3.12) as below

$$\sum_{i=1}^{n} w_i \left(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta} \right) \mathbf{X}_i^\top = 0.. \tag{3.13}$$

Solving this problem will be done using an iterative method

1. Define an initial estimate for $\boldsymbol{\beta}$ at any τ, and denote it by $\widehat{\boldsymbol{\beta}}^{(0)}$. This can be done using quantile regression when the MNS data is treated as SRS.

2. Calculate $\epsilon_i^{(j-1)} = Y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}^{(j-1)}$ at each iteration j.

3. Calculate $w_i^{(j-1)} = w(\epsilon_i^{(j-1)})$ where we take $\sigma = 1$ without loss of generality.

4. Solve $\sum_{i=1}^{n} w_i^{(j-1)} (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}) \mathbf{X}_i^\top = 0$, in $\boldsymbol{\beta}$ to obtain $\boldsymbol{\beta}^{(j)}$.

5. Repeat steps 2-4 until convergence. Convergence is achieved when $|\widehat{\boldsymbol{\beta}}^{(j+1)} - \widehat{\boldsymbol{\beta}}^{(j)}| < \epsilon$, for small predefined tolerance level $\epsilon > 0$.

3.3.2 Second Method

The second method we present is built upon the result established in Nourmohammadi, M. et al. (2015). This approach is carried out by selecting samples from the population using MNS. The maxima nominated samples are then treated as simple random samples to estimate the τ^{th} regression coefficients. To this end, we first present the following result.

Lemma 3.3.1. *Suppose F is a continuous cdf and let R and S denote observations obtained from F using SRS and MNS designs respectively. Then, $Q_\tau(R) = Q_{\tau^k}(S)$, where $\tau \in (0, 1)$ and k is the set size.*

Proof: By definition, the distribution function of S is given as $G(s) = [F(s)]^k$ which is a non-decreasing function. Then,

$$\begin{aligned} Q_\tau(R) &= \inf\{r : F(r) \geq \tau\} \\ &= \inf\{r : [F(r)]^k \geq \tau^k\} \\ &= \inf\{r : G(r) \geq \tau^k\} \\ &= Q_{\tau^k}(S). \end{aligned}$$

Lemma 3.3.1 allows us to define the τ^{th} regression quantile using MNS as the solution of

$$\min_{\boldsymbol{\beta} \in R^p} \left[\tau^k \sum_{\{i:Y_i > \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| + (1 - \tau^k) \sum_{\{i:Y_i < \mathbf{X}_i^\top \boldsymbol{\beta}\}} |Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}| \right]. \tag{3.14}$$

Remark. *The methods discussed in this section can be easily extended to minima nomination samples. To see this, let Y_i be the minimum of the sample $\mathbf{Y_i}$ of size k, $i = 1, \ldots, n$. Then Y_i has a distribution function of $G_Y(y) = 1 - [1 - F_Y(y)]^k$ and density $g_Y(y) = kf(y)[1 - F_Y(y)]^{k-1}$ for all $y \in R$. Now the τ^{th} regression quantile using minima nominated samples can be easily derived by simple modifications of previously derived methods for MNS data.*

3.4 Simulation Study

In this section, we conduct simulation studies to investigate the performance of the proposed objective function for maxima-nominated samples for different quantiles and set sizes. From the result established in Nourmohammadi, M. et al. (2015), it has been shown that maxima-nominated samples performs well at upper quantiles. Hence, we set $\tau = (0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ and work with set sizes $k = (2, 3, 4, 5, 6, 7, 8)$ to obtain maxima-nominated samples from a population of size $N = 10000$. To this end, we generated datasets from a classical linear regression model given as; $\mathbf{Y} = \mathbf{X}^\top\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the response variable, $\mathbf{X} = (X_1, \ldots, X_n)^\top$ are the covariates which can be categorical or continuous, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of errors with ϵ_i's being generated from a normal distribution. The regression coefficients of the classical regression model were randomly generated between 1 to 10 which resulted in the parameter values $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (7, 10, 6, 9)$ for four different categories which we assumed to have unequal sample sizes. The categories are denoted as X_1, X_2, X_3, X_4. For the continuous case, we generated our continuous variable X_5 as, $X_5 = \rho Y + (\sqrt{1-\rho^2})Z$, where $Z \sim N(0, 5)$, such that the correlation coefficient between X_5 and Y is ρ. We conducted our simulation studies in two stages. In both stages, additive perceptual error model of Dell, T.R. and Clutter, J.L. (1972) discussed in Chapter 2 of this thesis, was used to generate the population and associated concomitants with correlation coefficient $\rho = 0.6, 0.7, 0.8$ and 0.9 that are used for the purpose of ranking of the variable of interest. It is important to note that throughout this study our ranking process is imperfect (see Chapter 2 for details). The histogram of the underlying population is given in Fig. 3.2. Also, in

Fig. 3.3 we provide the population quantiles for different values of τ corresponding to the four subpopulations in our data set.

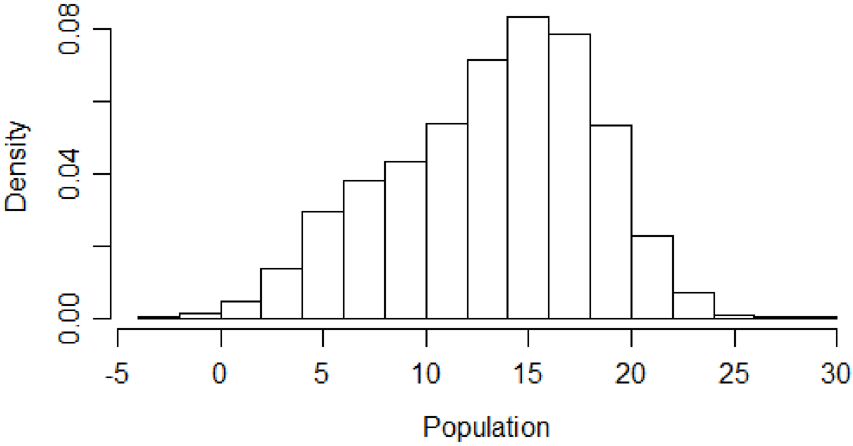

Figure 3.2: Histogram of the underlying population.

3.4.1 Categorical Covariates

In many applications in the social sciences, medical sciences and public health sciences, investigating the effect of covariates which split the population into several subgroups, is of interest. For example, in a study to predict the change in BMD which is important to determine the risk of osteoporosis, factors to be considered include: smoking history, family history of bone fracture, age. It is evident that most of the factors that will help us explain the response variable of interest are categorical variables. To this end, we considered in this section purely categorical covariates to explain the response variable. We select a random sample of size 200 from the population of size 10000, using the proportional to size sampling approach. We estimate the τ^{th} regression quantile for specified τ using simple random samples and maxima-nominated samples with the “quantreg” a built-in function in R and our proposed objective function respectively. We examined the effect of different set sizes for given parameter values of perceptual error model with $\rho = 0.6, 0.7, 0.8, 0.9$, where ρ is the correlation coefficient between the ranking variable and response variable. We considered three scenarios based on (i) our proposed objective function, (ii) approximation of the objective function and (iii) treating MNS as SRS, that is, instead of estimating $Q_\tau(Y \mid \mathbf{X})$ we estimate $Q_{\tau^k}(Y \mid \mathbf{X})$. This process is repeated 5000 times to see the behaviour of our proposed method in the long run. One can also try the simulation study with different number of replications other than 5000, however, as our simulation studies (that are not presented here) show, the results will remain almost the same as long as the number of replication is high enough, say more than a few thousand times. The estimated bias, averaged over the number of simulations, is given as;

$$\widehat{Bias}(\hat{Q}_\tau(Y \mid \mathbf{X})) = \frac{1}{5000}\sum_{i=1}^{5000} \hat{Q}_\tau^{(i)}(Y \mid \mathbf{X}) - Q_\tau(Y),$$

and the estimated mean squared error (MSE) is calculated as,

$$\widehat{MSE}(\hat{Q}_\tau(Y \mid \mathbf{X})) = \widehat{Bias}^2(\hat{Q}_\tau(Y \mid \mathbf{X})) + Var(\hat{Q}_\tau(Y \mid \mathbf{X})),$$

where $\hat{Q}_\tau(Y \mid \mathbf{X})$ is the estimated τ^{th} regression quantile of $Y \mid \mathbf{X}$. We presented in figures and tables the relative efficiency of MNS with respect to SRS with different set sizes and different correlation coefficients. The relative efficiency is computed as;

$$\widehat{RE}(\hat{Q}_\tau(Y \mid \mathbf{X})) = \frac{\widehat{MSE}_{SRS}(\hat{Q}_\tau(Y \mid \mathbf{X}))}{\widehat{MSE}_{MNS}(\hat{Q}_\tau(Y \mid \mathbf{X}))},$$

where $\widehat{RE} > 1$ indicates the superiority of MNS over SRS method. Also, we presented relative efficiency of SRS compared with MNS using the approximated objective function of τ^{th} quantile regression, in order to check how much efficiency we loose if we exclude the last terms from each line in the objective function 3.10 because of the difficulty in its maximization. The approximated objective function of the τ^{th} quantile regression is given as;

$$\min_{\boldsymbol{\beta}\in R^p}\left[\sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)-(k-1)\left(\ln\tau-(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right)\right]+\sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}}\tau\left(\frac{Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right]. \tag{3.15}$$

In each simulation, using the method of Dell, T.R. and Clutter, J.L. (1972), we create a concomitant variable which has a correlation coefficient ρ with the study variable and use it for the ranking purpose.

In each simulation, using the method of Dell, T.R. and Clutter, J.L. (1972), we create a concomitant variable which has a correlation coefficient ρ with the study variable and use it for the ranking purpose. One can easily observe that our proposed methods (quantile regression using MNS) outperform the quantile regression using SRS at upper quantiles even if the ranking is not very good. One can also observe that if the ranking is relatively good the MNS outperforms SRS for most quantiles. This is because the ranking information is considered in deriving the objective function. The relative efficiency of the approximated objective function of MNS over SRS showed that we will be losing efficiency by approximating the objective function when compared with the full objective function. This could be significant when ρ is small. On the other, the results obtained using lemma 3.3.1, that is, treating MNS data as SRS showed that we gained efficiency mostly at the upper quantile when compared with quantile regression using SRS which makes it similar to what we obtained using the proposed method.

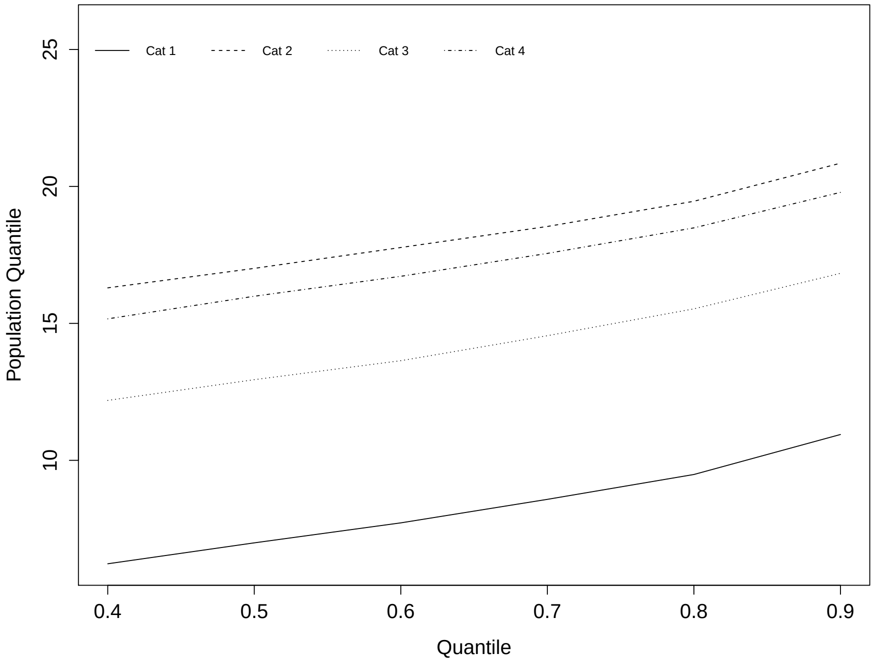

Figure 3.3: Population Quantiles for $\tau = (0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ corresponding to four subpopulations created by a categorical variable with 4 levels (Cat1, Cat2, Cat3, Cat4)

.

Table 3.1: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples for categorical covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.6$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.059	0.272	0.072	0.276	0.160	0.428	0.158	0.652
	2	-0.046	0.240	-0.094	0.253	-0.119	0.348	-0.247	0.487
	3	-0.018	0.282	0.039	0.289	-0.082	0.382	-0.117	0.568
	4	-0.101	0.300	-0.079	0.285	-0.118	0.342	-0.194	0.473
$k = 2$	1	-0.379	0.378	-0.304	0.312	-0.171	0.294	-0.201	0.376
	2	-0.359	0.328	-0.349	0.318	-0.257	0.289	-0.257	0.362
	3	-0.364	0.366	-0.260	0.286	-0.264	0.328	-0.191	0.383
	4	-0.378	0.365	-0.301	0.285	-0.250	0.278	-0.237	0.362
$k = 3$	1	-0.614	0.598	-0.515	0.458	-0.359	0.342	-0.318	0.361
	2	-0.627	0.577	-0.594	0.535	-0.454	0.377	-0.388	0.389
	3	-0.614	0.587	-0.478	0.416	-0.472	0.426	-0.325	0.371
	4	-0.600	0.562	-0.512	0.449	-0.459	0.383	-0.406	0.402
$k = 4$	1	-0.791	0.847	-0.681	0.659	-0.512	0.476	-0.425	0.418
	2	-0.836	0.894	-0.791	0.814	-0.612	0.541	-0.548	0.521
	3	-0.777	0.832	-0.632	0.606	-0.618	0.582	-0.434	0.439
	4	-0.784	0.813	-0.688	0.658	-0.622	0.537	-0.548	0.507
$k = 5$	1	-0.897	1.014	-0.780	0.791	-0.618	0.574	-0.512	0.471
	2	-1.009	1.209	-0.952	1.091	-0.739	0.697	-0.667	0.645
	3	-0.900	1.027	-0.746	0.749	-0.748	0.734	-0.550	0.509
	4	-0.933	1.060	-0.827	0.859	-0.759	0.712	-0.690	0.661
$k = 6$	1	-1.010	1.250	-0.892	0.997	-0.735	0.741	-0.617	0.580
	2	-1.151	1.519	-1.087	1.377	-0.861	0.899	-0.794	0.825
	3	-1.035	1.301	-0.869	0.960	-0.875	0.939	-0.675	0.652
	4	-1.043	1.291	-0.932	1.052	-0.871	0.894	-0.804	0.828
$k = 7$	1	-1.088	1.406	-0.960	1.120	-0.812	0.857	-0.680	0.658
	2	-1.260	1.787	-1.181	1.593	-0.946	1.052	-0.891	0.971
	3	-1.124	1.491	-0.949	1.101	-0.958	1.080	-0.754	0.751
	4	-1.153	1.540	-1.031	1.253	-0.980	1.094	-0.956	1.026
$k = 8$	1	-1.157	1.579	-1.024	1.260	-0.885	0.988	-0.747	0.754
	2	-1.358	2.044	-1.279	1.836	-1.041	1.245	-0.989	1.149
	3	-1.203	1.686	-1.020	1.254	-1.041	1.246	-0.843	0.885
	4	-1.250	1.769	-1.119	1.440	-1.076	1.287	-1.038	1.233

Table 3.2: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples for categorical covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.7$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.059	0.272	0.072	0.276	0.160	0.428	0.158	0.652
	2	-0.046	0.240	-0.094	0.253	-0.119	0.348	-0.247	0.487
	3	-0.018	0.282	0.039	0.289	-0.082	0.382	-0.117	0.568
	4	-0.101	0.300	-0.079	0.285	-0.118	0.342	-0.194	0.473
$k = 2$	1	-0.215	0.262	-0.170	0.234	-0.069	0.260	-0.136	0.338
	2	-0.212	0.233	-0.215	0.232	-0.156	0.239	-0.184	0.305
	3	-0.208	0.260	-0.124	0.222	-0.166	0.270	-0.121	0.353
	4	-0.220	0.256	-0.170	0.214	-0.153	0.231	-0.161	0.326
$k = 3$	1	-0.341	0.315	-0.283	0.254	-0.190	0.226	-0.196	0.274
	2	-0.378	0.310	-0.365	0.299	-0.288	0.244	-0.252	0.268
	3	-0.345	0.307	-0.243	0.227	-0.294	0.273	-0.193	0.292
	4	-0.345	0.308	-0.285	0.249	-0.294	0.247	-0.265	0.297
$k = 4$	1	-0.447	0.396	-0.378	0.315	-0.288	0.270	-0.273	0.289
	2	-0.517	0.445	-0.494	0.413	-0.394	0.310	-0.369	0.328
	3	-0.430	0.384	-0.318	0.277	-0.379	0.326	-0.255	0.304
	4	-0.454	0.387	-0.384	0.308	-0.397	0.292	-0.352	0.313
$k = 5$	1	-0.502	0.436	-0.431	0.346	-0.343	0.286	-0.316	0.288
	2	-0.626	0.559	-0.584	0.500	-0.468	0.348	-0.436	0.356
	3	-0.498	0.426	-0.379	0.303	-0.461	0.373	-0.328	0.305
	4	-0.532	0.451	-0.457	0.357	-0.484	0.350	-0.440	0.364
$k = 6$	1	-0.574	0.525	-0.495	0.413	-0.416	0.340	-0.389	0.319
	2	-0.718	0.697	-0.665	0.615	-0.543	0.424	-0.517	0.432
	3	-0.577	0.525	-0.450	0.370	-0.551	0.464	-0.414	0.358
	4	-0.597	0.533	-0.515	0.418	-0.561	0.432	-0.520	0.437
$k = 7$	1	-0.617	0.574	-0.532	0.449	-0.459	0.377	-0.426	0.348
	2	-0.783	0.796	-0.715	0.685	-0.594	0.479	-0.580	0.489
	3	-0.628	0.586	-0.491	0.408	-0.599	0.511	-0.460	0.383
	4	-0.661	0.612	-0.573	0.479	-0.632	0.509	-0.598	0.518
$k = 8$	1	-0.657	0.631	-0.567	0.494	-0.494	0.414	-0.462	0.376
	2	-0.857	0.920	-0.780	0.784	-0.653	0.548	-0.647	0.568
	3	-0.670	0.636	-0.522	0.439	-0.650	0.571	-0.517	0.434
	4	-0.733	0.713	-0.632	0.549	-0.704	0.602	-0.683	0.622

Table 3.3: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples for categorical covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.8$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.059	0.272	0.072	0.276	0.160	0.428	0.158	0.652
	2	-0.046	0.240	-0.094	0.253	-0.119	0.348	-0.247	0.487
	3	-0.018	0.282	0.039	0.289	-0.082	0.382	-0.117	0.568
	4	-0.101	0.300	-0.079	0.285	-0.118	0.342	-0.194	0.473
$k = 2$	1	-0.081	0.208	-0.057	0.197	0.008	0.245	-0.088	0.318
	2	-0.076	0.184	-0.096	0.185	-0.073	0.216	-0.123	0.268
	3	-0.077	0.206	-0.013	0.197	-0.092	0.238	-0.067	0.335
	4	-0.084	0.201	-0.058	0.181	-0.074	0.208	-0.106	0.304
$k = 3$	1	-0.111	0.184	-0.084	0.166	-0.051	0.177	-0.101	0.229
	2	-0.150	0.176	-0.161	0.175	-0.149	0.176	-0.149	0.202
	3	-0.111	0.180	-0.041	0.159	-0.154	0.197	-0.093	0.254
	4	-0.121	0.184	-0.091	0.161	-0.154	0.174	-0.154	0.235
$k = 4$	1	-0.140	0.188	-0.114	0.169	-0.098	0.174	-0.146	0.215
	2	-0.212	0.203	-0.212	0.192	-0.199	0.184	-0.219	0.207
	3	-0.126	0.186	-0.050	0.160	-0.180	0.194	-0.116	0.240
	4	-0.160	0.190	-0.124	0.158	-0.207	0.171	-0.197	0.212
$k = 5$	1	-0.151	0.176	-0.117	0.153	-0.107	0.156	-0.153	0.189
	2	-0.259	0.214	-0.248	0.195	-0.235	0.174	-0.252	0.198
	3	-0.142	0.173	-0.065	0.141	-0.219	0.189	-0.150	0.209
	4	-0.187	0.181	-0.147	0.147	-0.255	0.173	-0.244	0.210
$k = 6$	1	-0.182	0.190	-0.142	0.165	-0.144	0.162	-0.202	0.188
	2	-0.307	0.246	-0.208	0.216	-0.274	0.189	-0.296	0.214
	3	-0.173	0.185	-0.088	0.144	-0.264	0.207	-0.202	0.216
	4	-0.213	0.195	-0.169	0.156	-0.298	0.195	-0.282	0.223
$k = 7$	1	-0.191	0.187	-0.144	0.159	-0.154	0.157	-0.216	0.189
	2	-0.333	0.267	-0.295	0.228	-0.298	0.198	-0.333	0.229
	3	-0.196	0.193	-0.100	0.145	-0.284	0.210	-0.222	0.206
	4	-0.238	0.198	-0.185	0.155	-0.333	0.209	-0.322	0.240
$k = 8$	1	-0.211	0.199	-0.154	0.170	-0.167	0.165	-0.235	0.192
	2	-0.369	0.293	-0.319	0.246	-0.326	0.210	-0.373	0.252
	3	-0.204	0.196	-0.099	0.143	-0.301	0.221	-0.240	0.214
	4	-0.276	0.212	-0.212	0.159	-0.378	0.236	-0.370	0.267

Table 3.4: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples for categorical covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.9$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.059	0.272	0.072	0.276	0.160	0.428	0.158	0.652
	2	-0.046	0.240	-0.094	0.253	-0.119	0.348	-0.247	0.487
	3	-0.018	0.282	0.039	0.289	-0.082	0.382	-0.117	0.568
	4	-0.101	0.300	-0.079	0.285	-0.118	0.342	-0.194	0.473
$k = 2$	1	0.040	0.190	0.039	0.186	0.067	0.239	-0.049	0.302
	2	0.045	0.172	0.009	0.165	-0.004	0.205	-0.081	0.246
	3	0.039	0.184	0.086	0.198	-0.027	0.220	-0.029	0.321
	4	0.032	0.184	0.036	0.174	-0.011	0.197	-0.065	0.287
$k = 3$	1	0.091	0.159	0.086	0.152	0.061	0.172	-0.030	0.209
	2	0.063	0.147	0.023	0.133	-0.027	0.150	-0.070	0.162
	3	0.085	0.154	0.127	0.163	-0.039	0.159	-0.017	0.235
	4	0.074	0.158	0.074	0.146	-0.037	0.145	-0.074	0.203
$k = 4$	1	0.122	0.158	0.112	0.152	0.058	0.151	-0.051	0.182
	2	0.074	0.141	0.038	0.126	-0.032	0.137	-0.107	0.146
	3	0.135	0.165	0.176	0.174	-0.020	0.140	-0.014	0.210
	4	0.095	0.146	0.097	0.136	-0.047	0.124	-0.088	0.165
$k = 5$	1	0.154	0.152	0.161	0.148	0.081	0.135	-0.037	0.150
	2	0.077	0.128	0.048	0.114	-0.033	0.116	-0.110	0.119
	3	0.166	0.153	0.211	0.164	-0.018	0.120	-0.013	0.168
	4	0.117	0.136	0.118	0.125	-0.055	0.105	-0.089	0.141
$k = 6$	1	0.166	0.157	0.177	0.155	0.082	0.129	-0.058	0.135
	2	0.077	0.125	0.060	0.113	-0.040	0.107	-0.128	0.113
	3	0.175	0.156	0.220	0.165	-0.029	0.114	-0.040	0.160
	4	0.126	0.139	0.136	0.130	-0.060	0.103	-0.093	0.132
$k = 7$	1	0.185	0.154	0.207	0.161	0.103	0.122	-0.046	0.124
	2	0.080	0.129	0.074	0.118	-0.035	0.103	-0.140	0.107
	3	0.185	0.157	0.244	0.172	-0.017	0.104	-0.029	0.142
	4	0.132	0.131	0.151	0.124	-0.063	0.095	-0.104	0.124
$k = 8$	1	0.196	0.161	0.225	0.172	0.111	0.124	-0.053	0.118
	2	0.069	0.119	0.081	0.113	0.039	0.098	-0.155	0.107
	3	0.204	0.157	0.274	0.183	-0.008	0.102	-0.025	0.136
	4	0.113	0.122	0.148	0.121	-0.080	0.096	-0.124	0.127

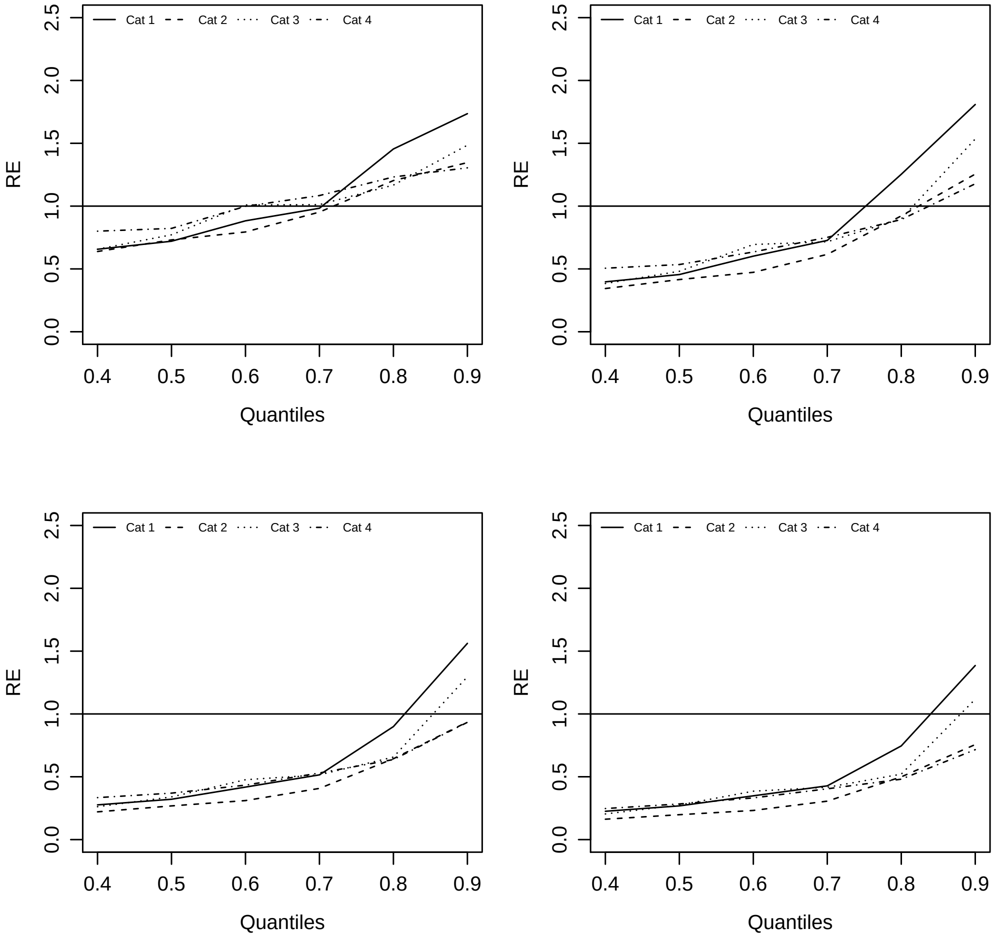

Figure 3.4: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.6$

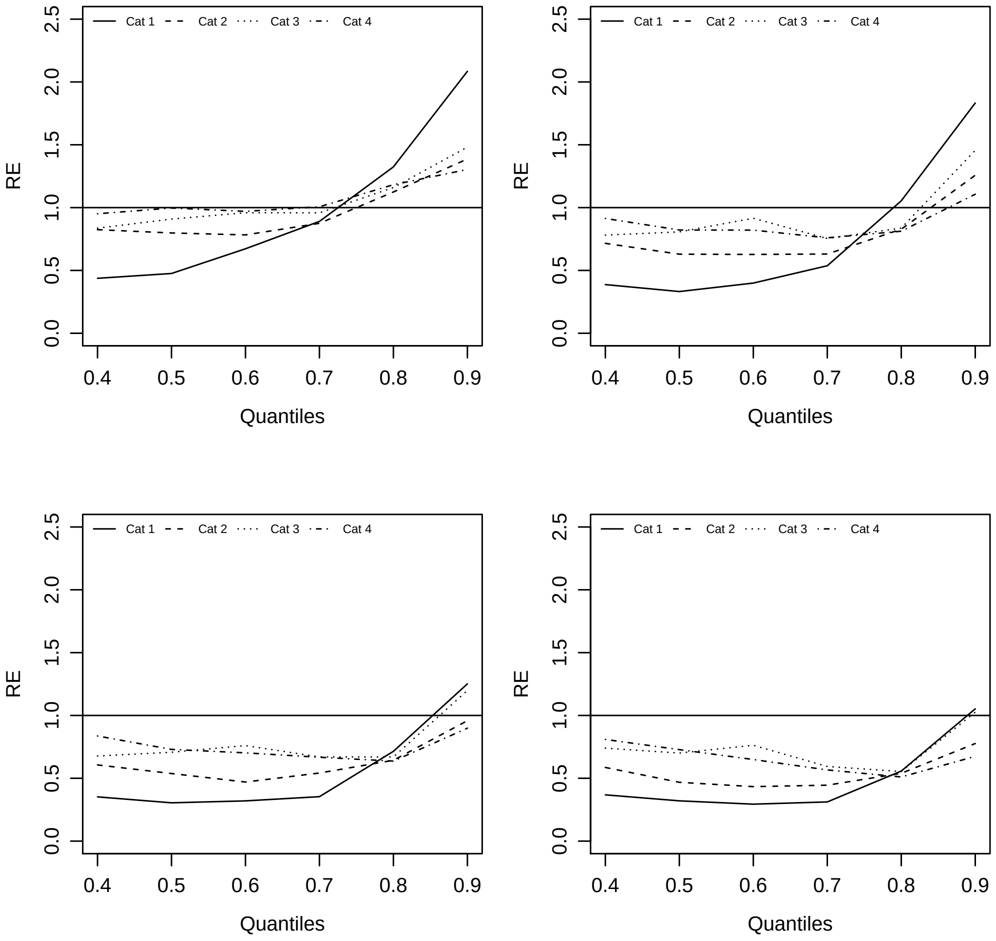

Figure 3.5: Relative Efficiency of SRS vs Approximated objective function of MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.6$

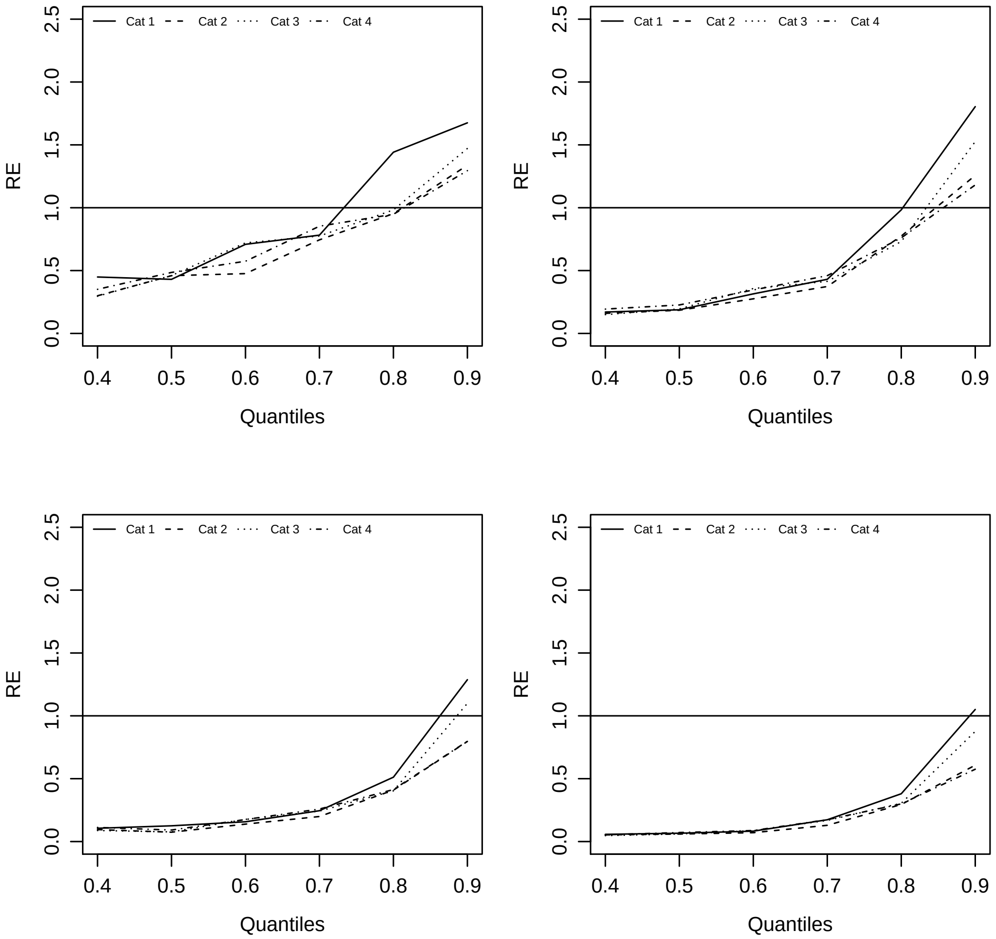

Figure 3.6: Relative Efficiency of SRS vs MNS, where MNS data is treated as SRS when $k = (2, 3, 4, 5)$ and $\rho = 0.6$

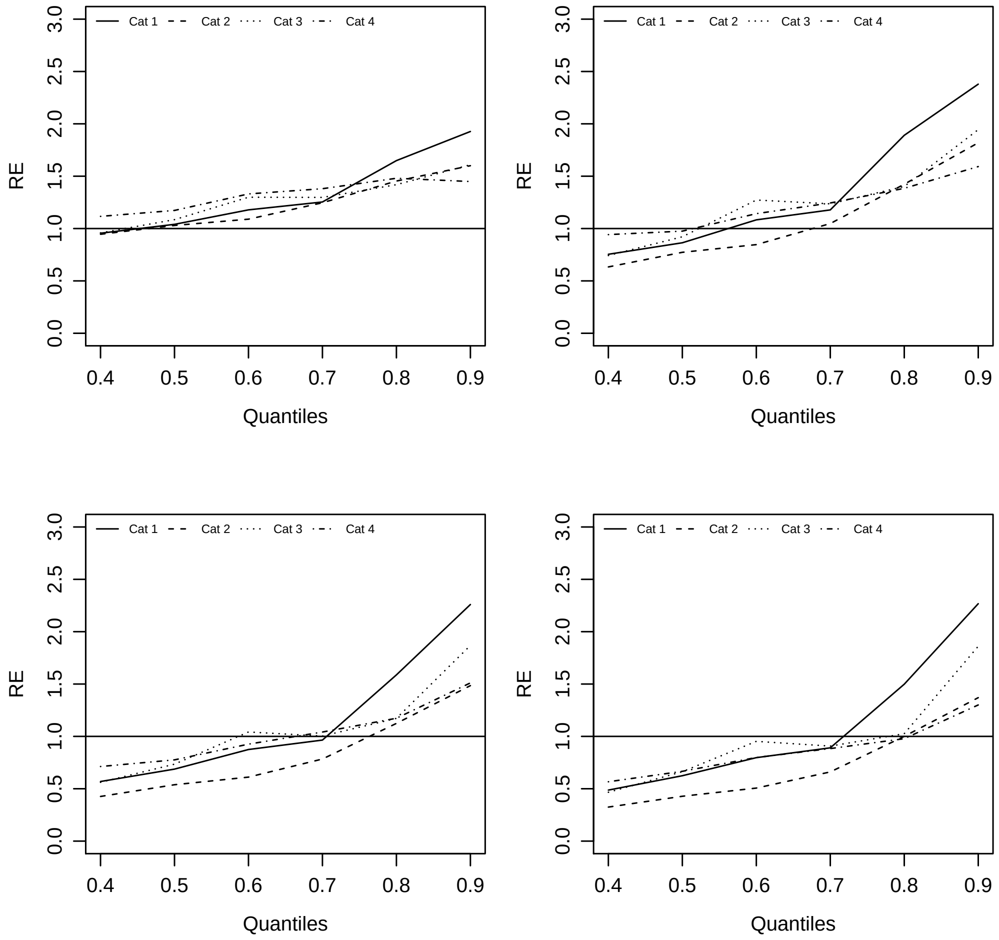

Figure 3.7: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.7$

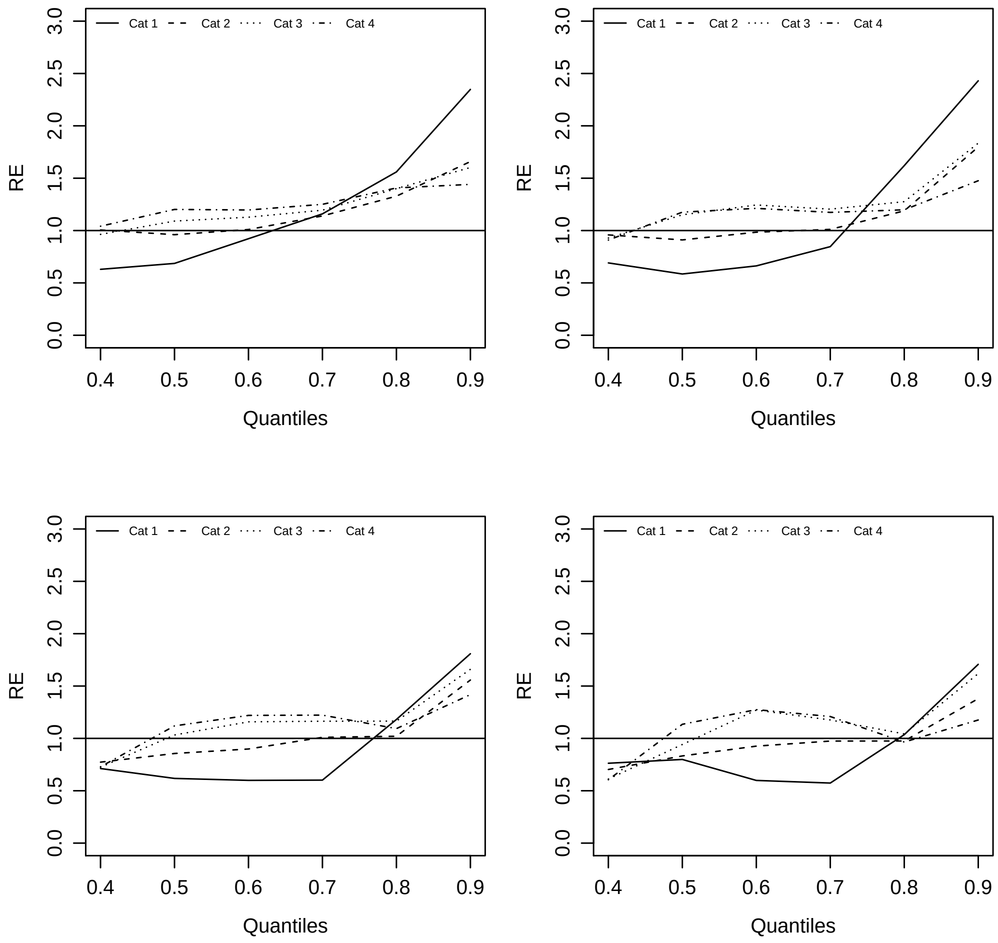

Figure 3.8: Relative Efficiency of SRS vs Approximated objective function of MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.7$

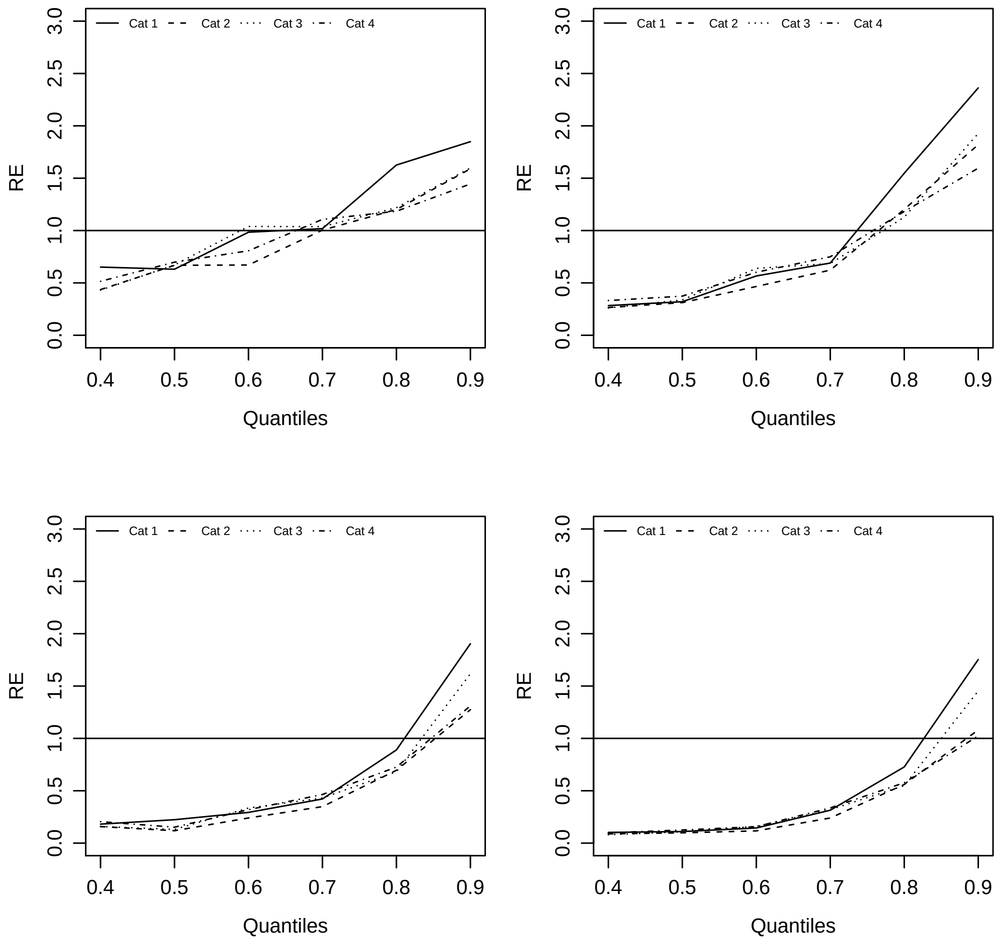

Figure 3.9: Relative Efficiency of SRS vs MNS, where MNS data is treated as SRS when $k = (2, 3, 4, 5)$ and $\rho = 0.7$

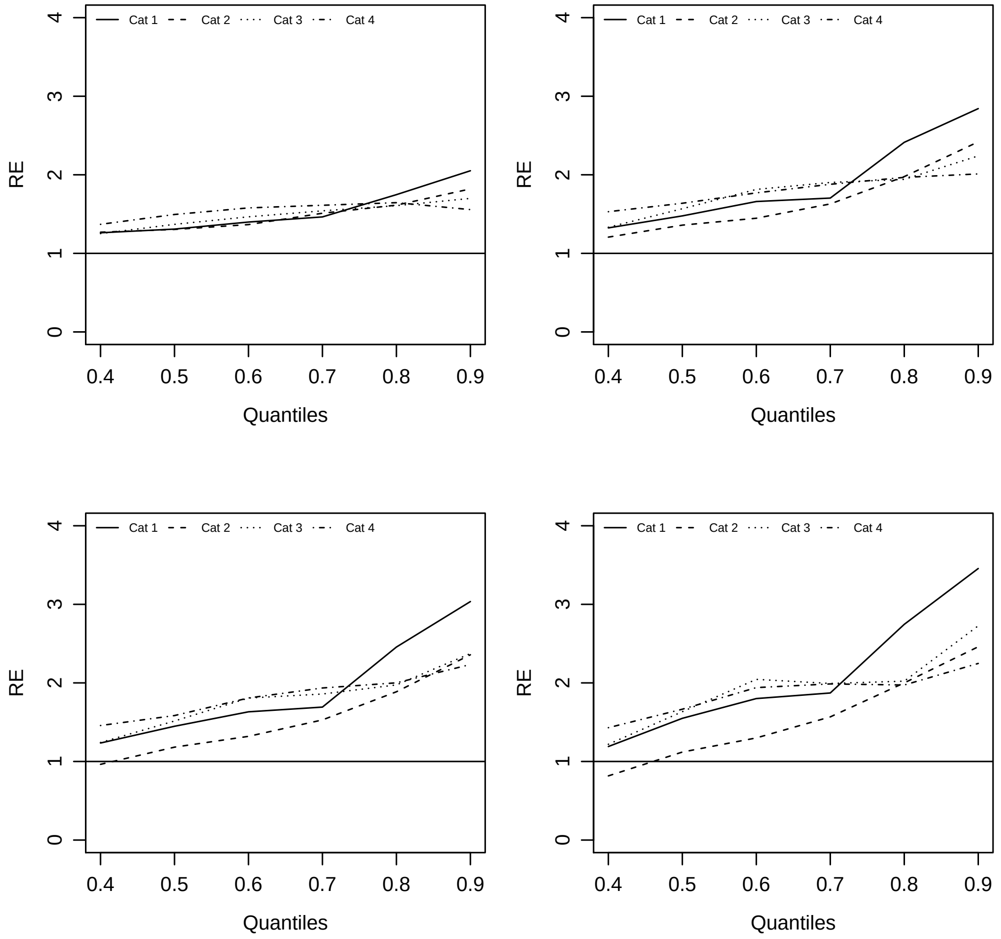

Figure 3.10: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.8$

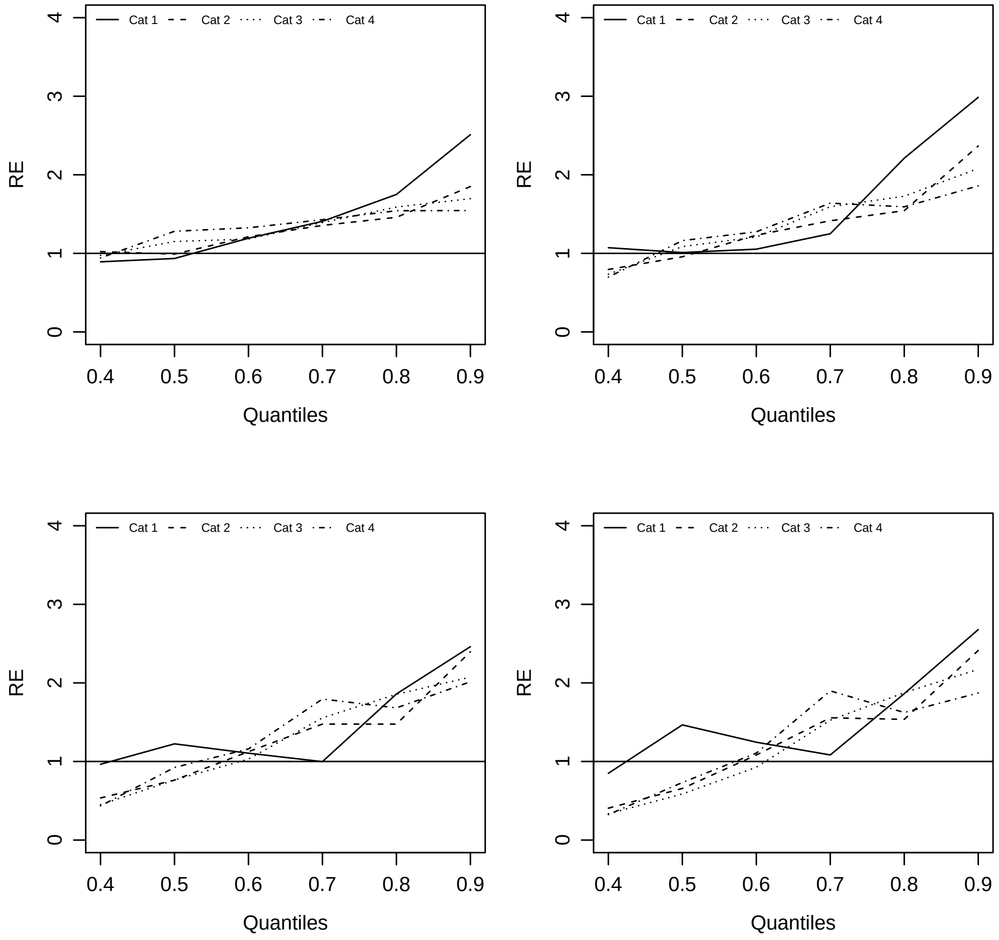

Figure 3.11: Relative Efficiency of SRS vs Approximated objective function of MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.8$

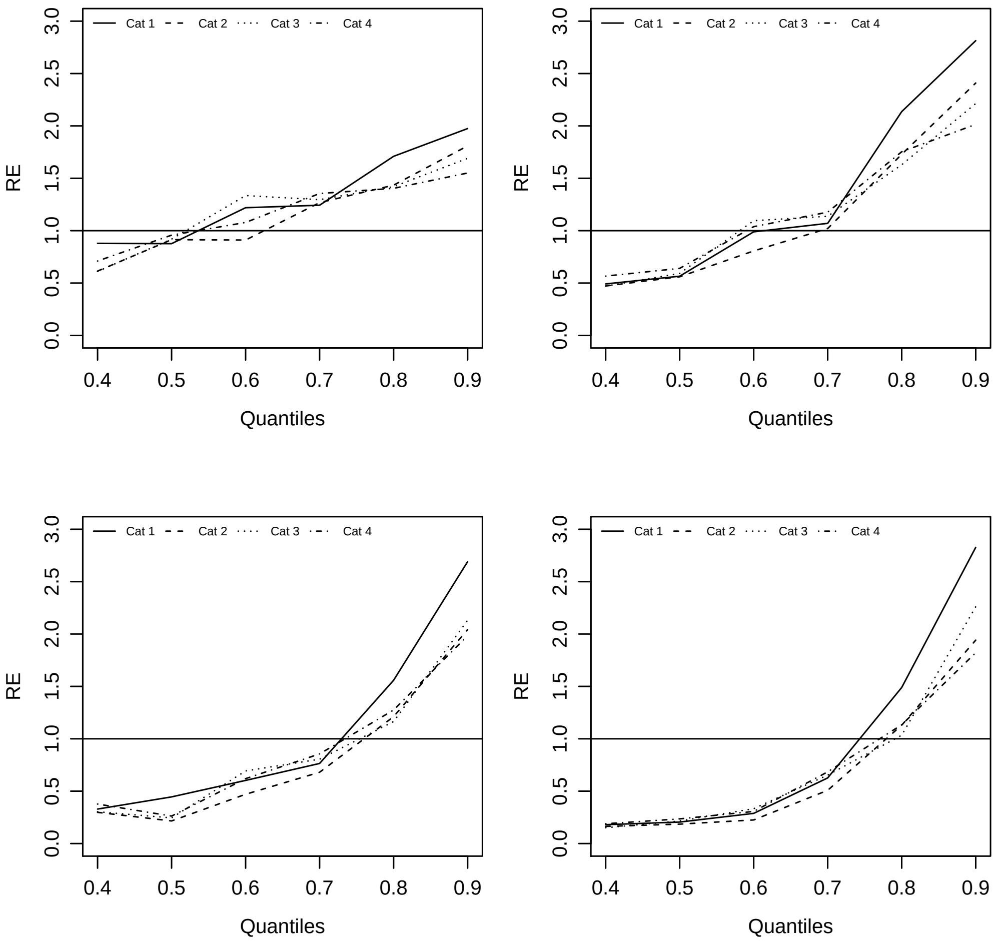

Figure 3.12: Relative Efficiency of SRS vs MNS, where MNS data treated as SRS when $k = (2, 3, 4, 5)$ and $\rho = 0.8$

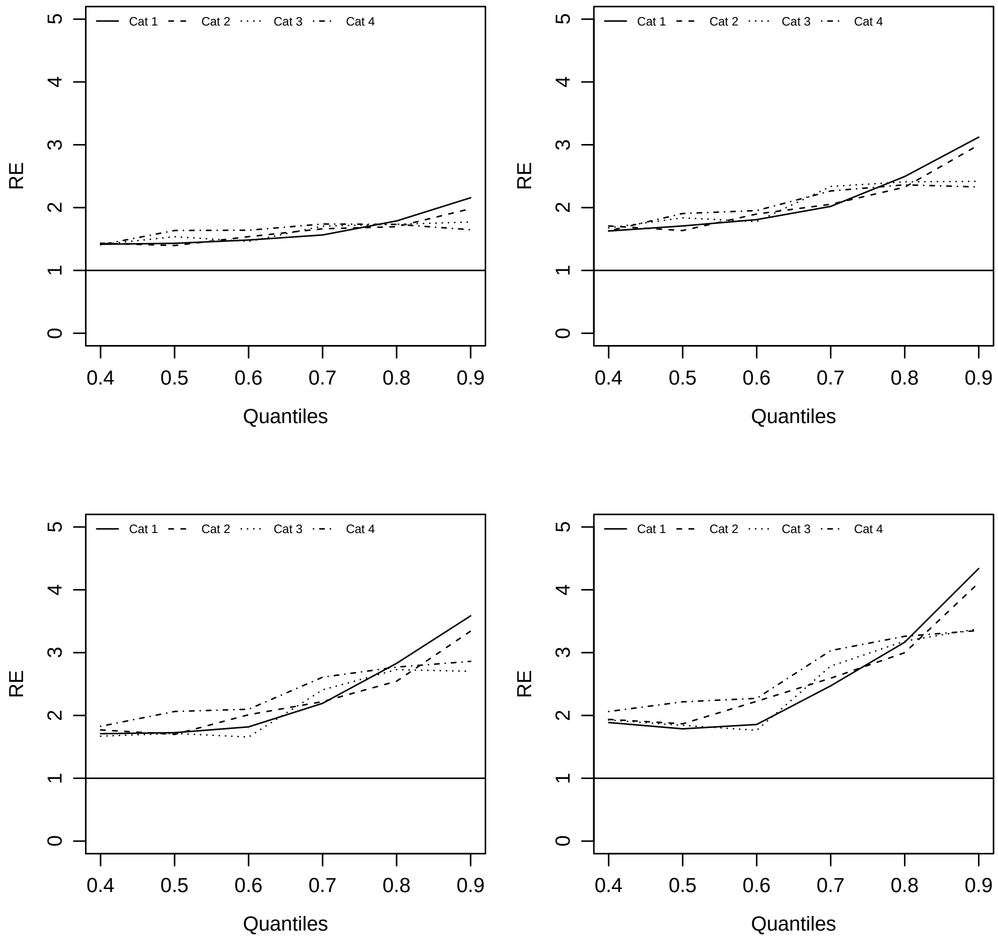

Figure 3.13: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

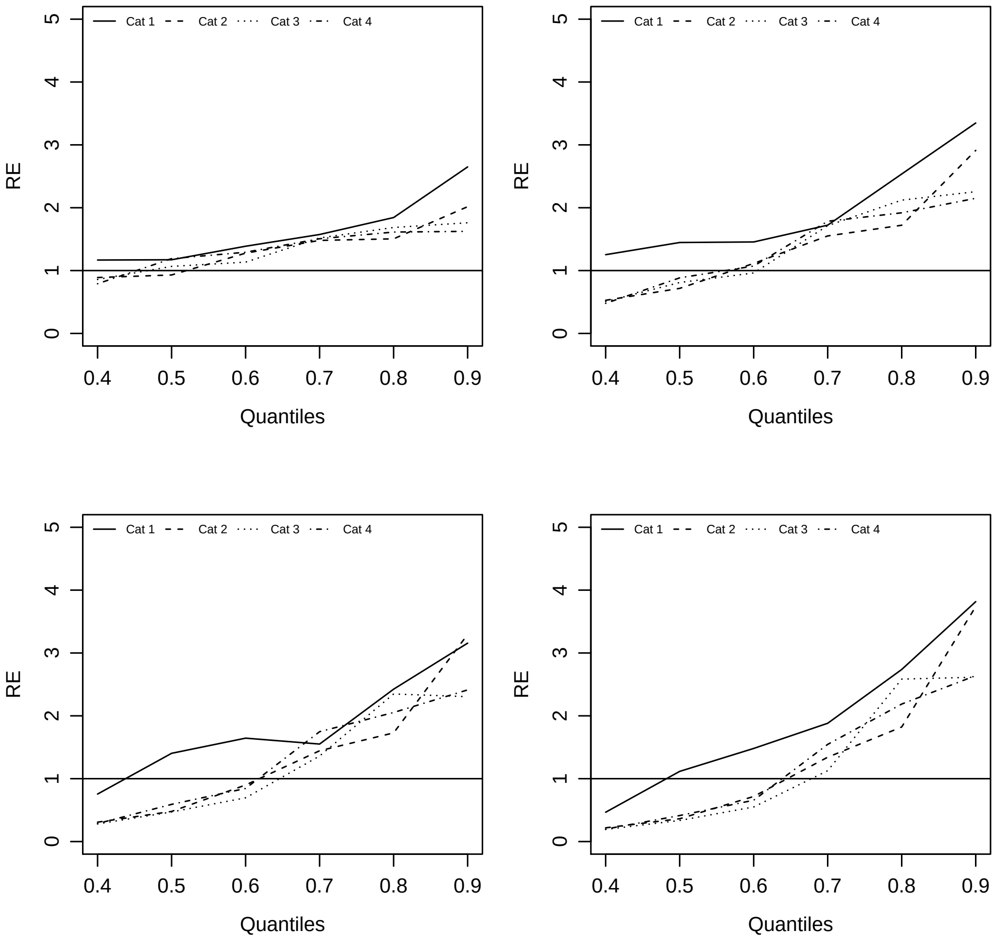

Figure 3.14: Relative Efficiency of SRS vs Approximated objective function of MNS when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

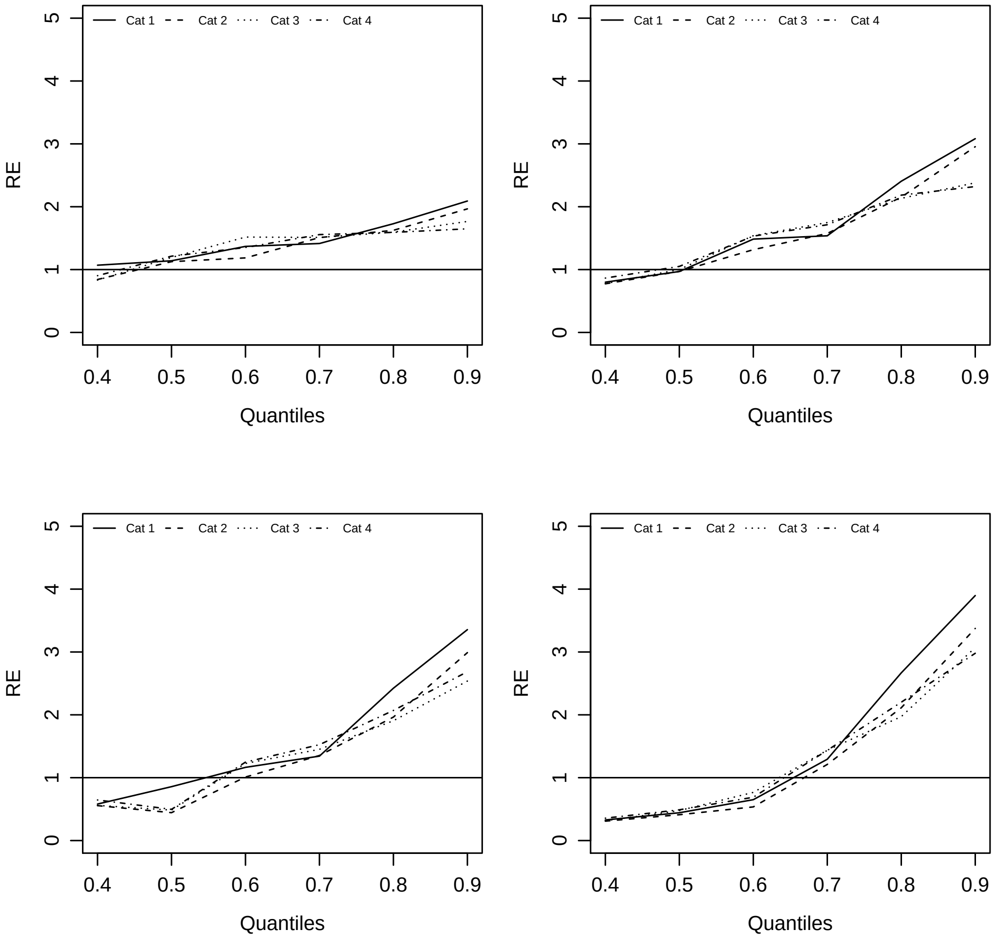

Figure 3.15: Relative Efficiency of SRS vs MNS, where MNS data is treated as SRS when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

3.4.2 Categorical and Continuous Covariates

In this section, we consider a case where we have both categorical and continuous variables as our explanatory variables. Applications of such a case is common in medical science and public health studies. For example, we might be interested in estimating or predicting the change in BMD given the age group category, weight, height and body mass index (BMI) as continuous variables. We select a random sample of size 200 from the population of size 10000. We estimate the τ^{th} regression quantile for specified τ using simple random and maxima-nominated samples with the "quantreg" a built-in function in R and our proposed objective function respectively.

From the graphs and tables presented, we observe that the MSE and the relative efficiency of the τ^{th} quantile regression using MNS is better than the case where we use the SRS technique for our estimation. The effects of set size k and the correlation coefficient between the response variable and concomitant (ρ) based on the relative efficiency are examined. The relative efficiency increases as ρ increases. Also, increase in the set size increases our efficiency at the upper quantiles and reduces the efficiency at the lower quantile. This implies that there is a trade off between the set size and the efficiency in estimating the quantile of interest. If our interest is to gain efficiency across all quantiles, we require a small set size, say $k = 2$. On the other hand, if our goal is to estimate the τ^{th} quantile regression of the upper quantile with high efficiency, it is required to increase our set size, say $k = 4, 5$.

In general, the first method which is built upon the distribution of maximum order statistics from an ALD outperforms the second method which allows us to treat the MNS as SRS, especially when the ranking is relatively good. However, the two methods have almost the same relative efficiency at the upper quantiles, say $\tau \geq 0.7$. Also, we observe that MNS works better for analyzing the upper quantiles of the population. Similar behavior is expected for lower quantiles if we work with minima NS. This is explored in our real data application in chapter 4.

Table 3.5: Estimated Bias ($\widehat{Bias}$) and $\widehat{MSE}$ for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples categorical and continuous covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.6$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.039	0.255	-0.014	0.278	-0.193	0.384	-0.416	0.699
	2	0.044	0.237	-0.037	0.243	-0.196	0.333	-0.486	0.612
	3	-0.030	0.252	-0.024	0.279	-0.253	0.412	-0.322	0.608
	4	-0.074	0.269	-0.092	0.286	-0.292	0.402	-0.389	0.640
$k = 2$	1	-0.262	0.267	-0.255	0.269	-0.273	0.296	-0.424	0.522
	2	-0.241	0.256	-0.294	0.263	-0.296	0.307	-0.498	0.492
	3	-0.267	0.264	-0.241	0.263	-0.361	0.372	-0.381	0.459
	4	-0.327	0.319	-0.312	0.310	-0.369	0.347	-0.412	0.479
$k = 3$	1	-0.445	0.387	-0.429	0.361	-0.392	0.333	-0.476	0.472
	2	-0.497	0.441	-0.507	0.419	-0.462	0.403	-0.560	0.496
	3	-0.446	0.387	-0.392	0.332	-0.472	0.430	-0.425	0.414
	4	-0.553	0.507	-0.505	0.441	-0.514	0.428	-0.488	0.461
$k = 4$	1	-0.562	0.511	-0.547	0.487	-0.487	0.422	-0.550	0.536
	2	-0.689	0.673	-0.661	0.616	-0.606	0.562	-0.649	0.596
	3	-0.582	0.528	-0.507	0.438	-0.587	0.543	-0.502	0.466
	4	-0.751	0.759	-0.682	0.658	-0.647	0.579	-0.605	0.564
$k = 5$	1	-0.749	0.781	-0.623	0.558	-0.585	0.512	-0.613	0.575
	2	-0.801	0.845	-0.800	0.809	-0.725	0.704	-0.714	0.663
	3	-0.652	0.629	-0.598	0.526	-0.675	0.629	-0.580	0.512
	4	-0.856	0.942	-0.826	0.865	-0.765	0.730	-0.697	0.657
$k = 6$	1	-0.799	0.845	-0.709	0.675	-0.677	0.629	-0.694	0.668
	2	-0.921	1.052	-0.894	0.968	-0.821	0.851	-0.788	0.778
	3	-0.750	0.773	-0.666	0.613	-0.745	0.732	-0.668	0.616
	4	-1.001	1.210	-0.944	1.066	-0.868	0.898	-0.793	0.787
$k = 7$	1	-0.845	0.903	-0.785	0.797	-0.733	0.701	-0.741	0.718
	2	-1.035	1.259	-0.985	1.138	-0.937	1.050	-0.879	0.922
	3	-0.850	0.921	-0.727	0.692	-0.849	0.886	-0.749	0.711
	4	-1.114	1.445	-1.025	1.243	-0.966	1.084	-0.881	0.929
$k = 8$	1	-0.915	1.032	-0.900	1.026	-0.819	0.842	-0.806	0.816
	2	-1.114	1.435	-1.039	1.260	-1.014	1.198	-0.948	1.051
	3	-0.900	1.016	-0.748	0.741	-0.899	0.968	-0.812	0.804
	4	-1.208	1.665	-1.092	1.389	-1.058	1.268	-0.968	1.079

Table 3.6: Estimated Bias ($\widehat{Bias}$) and $\widehat{MSE}$ for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples categorical and continuous covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.7$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.039	0.255	-0.014	0.278	-0.193	0.384	-0.416	0.699
	2	0.044	0.237	-0.037	0.243	-0.196	0.333	-0.486	0.612
	3	-0.030	0.252	-0.024	0.279	-0.253	0.412	-0.322	0.608
	4	-0.074	0.269	-0.092	0.286	-0.292	0.402	-0.389	0.640
$k = 2$	1	-0.122	0.201	-0.125	0.210	-0.181	0.246	-0.349	0.449
	2	-0.089	0.196	-0.158	0.192	-0.192	0.246	-0.410	0.403
	3	-0.121	0.202	-0.100	0.206	-0.248	0.289	-0.292	0.380
	4	-0.165	0.228	-0.167	0.231	-0.263	0.274	-0.315	0.399
$k = 3$	1	-0.223	0.227	-0.217	0.217	-0.229	0.223	-0.343	0.353
	2	-0.241	0.237	-0.280	0.225	-0.277	0.254	-0.416	0.348
	3	-0.205	0.217	-0.174	0.194	-0.288	0.273	-0.287	0.303
	4	-0.286	0.274	-0.261	0.240	-0.330	0.264	-0.333	0.327
$k = 4$	1	-0.278	0.251	-0.286	0.259	-0.273	0.244	-0.371	0.358
	2	-0.363	0.308	-0.364	0.285	-0.359	0.303	-0.464	0.376
	3	-0.279	0.250	-0.217	0.215	-0.350	0.303	-0.325	0.310
	4	-0.414	0.355	-0.360	0.306	-0.405	0.308	-0.404	0.350
$k = 5$	1	-0.389	0.349	-0.319	0.262	-0.323	0.261	-0.395	0.341
	2	-0.435	0.358	-0.450	0.345	-0.422	0.334	-0.494	0.377
	3	-0.314	0.267	-0.270	0.225	-0.398	0.314	-0.365	0.299
	4	-0.479	0.406	-0.441	0.355	-0.475	0.353	-0.460	0.366
$k = 6$	1	-0.454	0.399	-0.364	0.294	-0.385	0.308	-0.448	0.379
	2	-0.499	0.423	-0.507	0.394	-0.482	0.384	-0.538	0.419
	3	-0.357	0.307	-0.303	0.240	-0.436	0.346	-0.424	0.339
	4	-0.536	0.472	-0.506	0.408	-0.525	0.401	-0.515	0.409
$k = 7$	1	-0.467	0.389	-0.398	0.321	-0.407	0.321	-0.468	0.385
	2	-0.586	0.503	-0.574	0.468	-0.565	0.466	-0.592	0.474
	3	-0.418	0.337	-0.343	0.260	-0.506	0.398	-0.480	0.368
	4	-0.611	0.561	-0.555	0.475	-0.588	0.473	-0.568	0.458
$k = 8$	1	-0.499	0.415	-0.458	0.396	-0.442	0.356	-0.509	0.416
	2	-0.642	0.571	-0.609	0.516	-0.621	0.537	-0.639	0.530
	3	-0.448	0.369	-0.349	0.271	-0.538	0.429	-0.520	0.406
	4	-0.686	0.637	-0.602	0.520	-0.649	0.540	-0.629	0.523

Table 3.7: Estimated Bias ($\widehat{Bias}$) and $\widehat{MSE}$ for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples categorical and continuous covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.8$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.039	0.255	-0.014	0.278	-0.193	0.384	-0.416	0.699
	2	0.044	0.237	-0.037	0.243	-0.196	0.333	-0.486	0.612
	3	-0.030	0.252	-0.024	0.279	-0.253	0.412	-0.322	0.608
	4	-0.074	0.269	-0.092	0.286	-0.292	0.402	-0.389	0.640
$k = 2$	1	0.006	0.180	-0.013	0.188	-0.099	0.218	-0.288	0.398
	2	0.045	0.176	-0.040	0.161	-0.102	0.209	-0.339	0.340
	3	0.007	0.180	0.018	0.190	-0.154	0.241	-0.222	0.337
	4	-0.029	0.188	-0.042	0.196	-0.173	0.228	-0.236	0.348
$k = 3$	1	-0.023	0.165	-0.029	0.158	-0.087	0.163	-0.240	0.282
	2	-0.015	0.160	-0.086	0.143	-0.124	0.181	-0.295	0.254
	3	0.002	0.159	0.021	0.150	-0.138	0.192	-0.177	0.242
	4	-0.049	0.171	-0.050	0.163	-0.175	0.175	-0.202	0.247
$k = 4$	1	-0.018	0.155	-0.052	0.167	-0.086	0.161	-0.225	0.258
	2	-0.075	0.162	-0.106	0.147	-0.147	0.178	-0.307	0.239
	3	-0.014	0.154	-0.034	0.156	-0.152	0.181	-0.179	0.227
	4	-0.106	0.173	-0.074	0.166	-0.201	0.174	-0.234	0.231
$k = 5$	1	-0.053	0.160	-0.040	0.147	-0.095	0.146	-0.223	0.221
	2	-0.111	0.152	-0.147	0.142	-0.165	0.160	-0.311	0.215
	3	-0.019	0.142	0.023	0.141	-0.159	0.164	-0.192	0.196
	4	-0.130	0.165	-0.103	0.152	-0.224	0.165	-0.257	0.211
$k = 6$	1	-0.125	0.192	-0.039	0.142	-0.126	0.153	-0.245	0.223
	2	-0.127	0.162	-0.171	0.148	-0.192	0.168	-0.327	0.218
	3	-0.004	0.154	0.018	0.135	-0.177	0.167	-0.222	0.204
	4	-0.141	0.176	-0.134	0.150	-0.243	0.169	-0.288	0.217
$k = 7$	1	-0.114	0.175	-0.041	0.148	-0.118	0.140	-0.238	0.205
	2	-0.184	0.175	-0.205	0.161	-0.241	0.182	-0.358	0.232
	3	-0.044	0.143	-0.001	0.124	-0.209	0.168	-0.257	0.200
	4	-0.169	0.189	-0.146	0.161	-0.269	0.183	-0.309	0.219
$k = 8$	1	-0.133	0.178	-0.052	0.157	-0.123	0.146	-0.262	0.212
	2	-0.209	0.182	-0.216	0.170	-0.262	0.193	-0.375	0.241
	3	-0.054	0.145	-0.009	0.126	-0.225	0.172	-0.272	0.206
	4	-0.211	0.190	-0.170	0.157	-0.302	0.195	-0.345	0.236

Table 3.8: Estimated Bias ($\widehat{Bias}$) and $\widehat{MSE}$ for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples categorical and continuous covariates with set size, $k = (2, 3, 4, 5, 6, 7, 8)$ and $\rho = 0.9$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.039	0.255	-0.014	0.278	-0.193	0.384	-0.416	0.699
	2	0.044	0.237	-0.037	0.243	-0.196	0.333	-0.486	0.612
	3	-0.030	0.252	-0.024	0.279	-0.253	0.412	-0.322	0.608
	4	-0.074	0.269	-0.092	0.286	-0.292	0.402	-0.389	0.640
$k = 2$	1	0.112	0.182	0.084	0.185	-0.039	0.203	-0.243	0.360
	2	0.165	0.187	0.062	0.154	-0.028	0.190	-0.278	0.296
	3	0.114	0.186	0.119	0.194	-0.083	0.211	-0.168	0.305
	4	0.087	0.182	0.063	0.191	-0.102	0.200	-0.177	0.311
$k = 3$	1	0.154	0.175	0.136	0.164	0.027	0.148	-0.158	0.237
	2	0.187	0.176	0.086	0.134	0.002	0.155	-0.198	0.197
	3	0.192	0.184	0.191	0.176	-0.017	0.156	-0.091	0.210
	4	0.153	0.172	0.129	0.166	-0.050	0.139	-0.107	0.211
$k = 4$	1	0.221	0.191	0.163	0.177	0.071	0.142	-0.110	0.203
	2	0.187	0.165	0.122	0.136	0.028	0.139	-0.179	0.165
	3	0.226	0.191	0.253	0.203	0.012	0.140	-0.064	0.189
	4	0.156	0.162	0.165	0.173	-0.032	0.124	-0.107	0.174
$k = 5$	1	0.243	0.192	0.223	0.185	0.097	0.131	-0.081	0.165
	2	0.189	0.153	0.126	0.122	0.047	0.118	-0.157	0.130
	3	0.251	0.184	0.278	0.198	0.032	0.119	-0.052	0.153
	4	0.173	0.153	0.178	0.152	-0.024	0.105	-0.104	0.145
$k = 6$	1	0.216	0.198	0.263	0.197	0.105	0.129	-0.074	0.152
	2	0.203	0.162	0.135	0.126	0.057	0.112	-0.151	0.120
	3	0.292	0.213	0.307	0.215	0.045	0.116	-0.057	0.147
	4	0.199	0.166	0.191	0.155	-0.009	0.098	-0.105	0.136
$k = 7$	1	0.225	0.202	0.296	0.214	0.128	0.122	-0.047	0.138
	2	0.185	0.151	0.134	0.117	0.046	0.107	-0.157	0.113
	3	0.283	0.201	0.308	0.204	0.039	0.105	-0.066	0.131
	4	0.201	0.170	0.202	0.155	-0.003	0.095	-0.102	0.126
$k = 8$	1	0.226	0.205	0.319	0.227	0.149	0.127	-0.053	0.130
	2	0.178	0.147	0.145	0.125	0.050	0.104	-0.156	0.106
	3	0.297	0.211	0.339	0.224	0.045	0.105	-0.058	0.127
	4	0.191	0.156	0.207	0.151	-0.008	0.089	-0.116	0.118

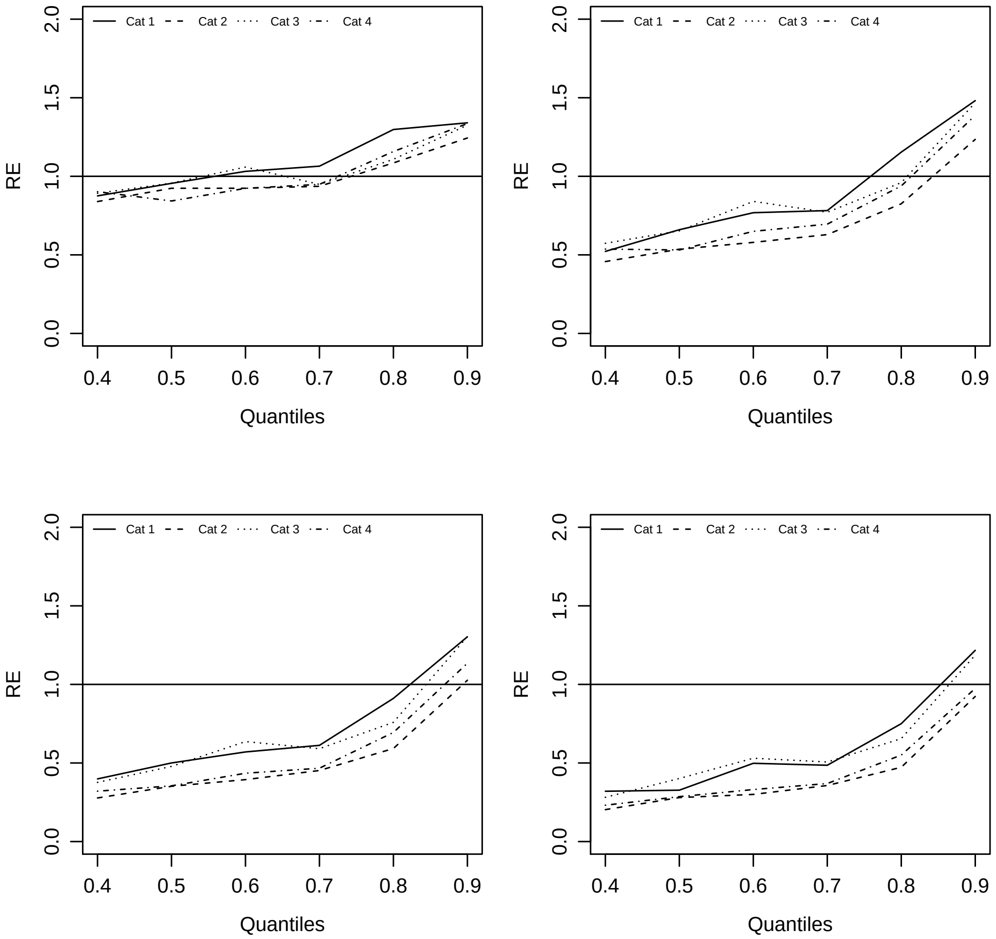

Figure 3.16: Relative Efficiency of SRS vs MNS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.6$

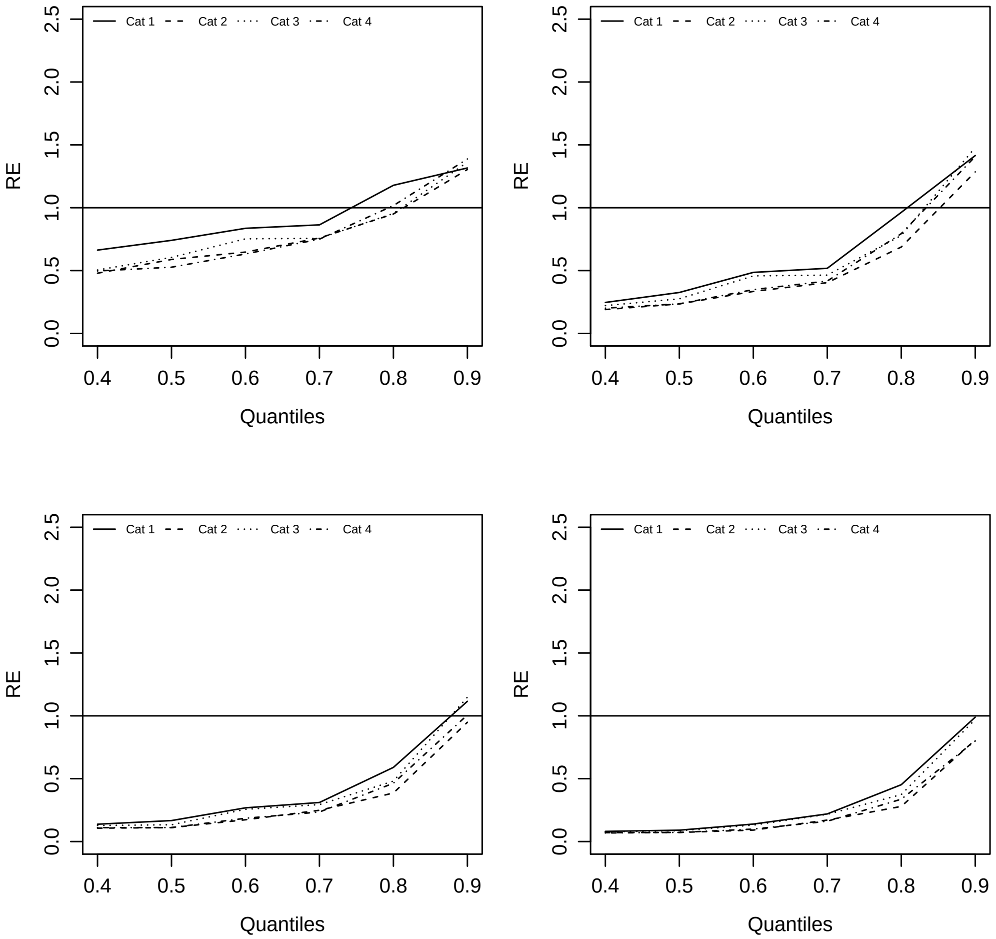

Figure 3.17: Relative Efficiency of SRS vs MNS treated as SRS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.6$

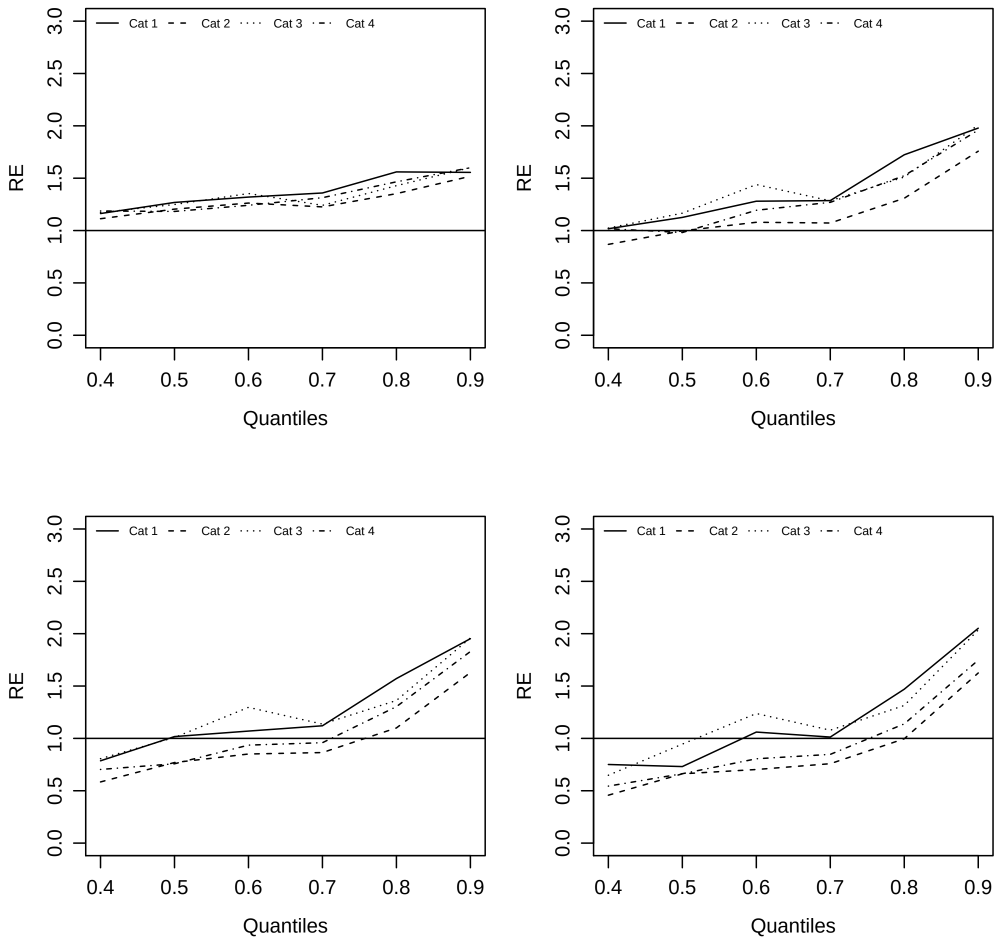

Figure 3.18: Relative Efficiency of SRS vs MNS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.7$

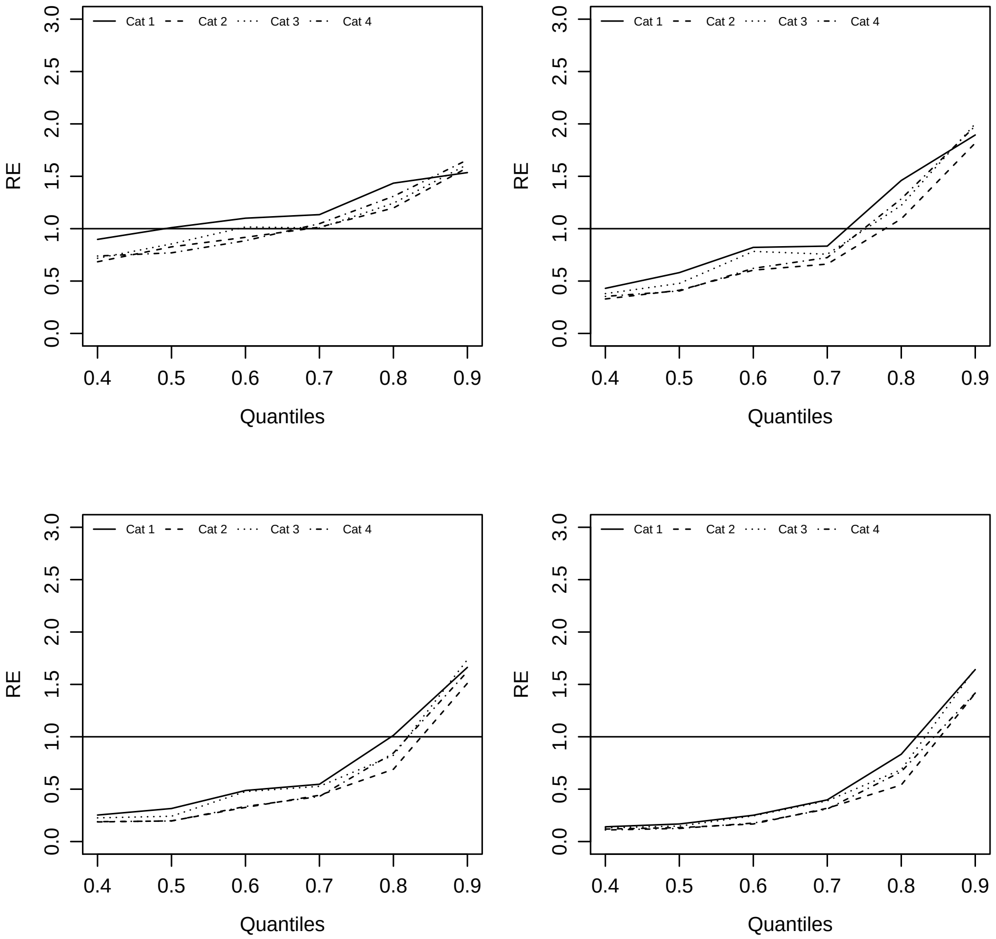

Figure 3.19: Relative Efficiency of SRS vs MNS treated as SRS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.7$

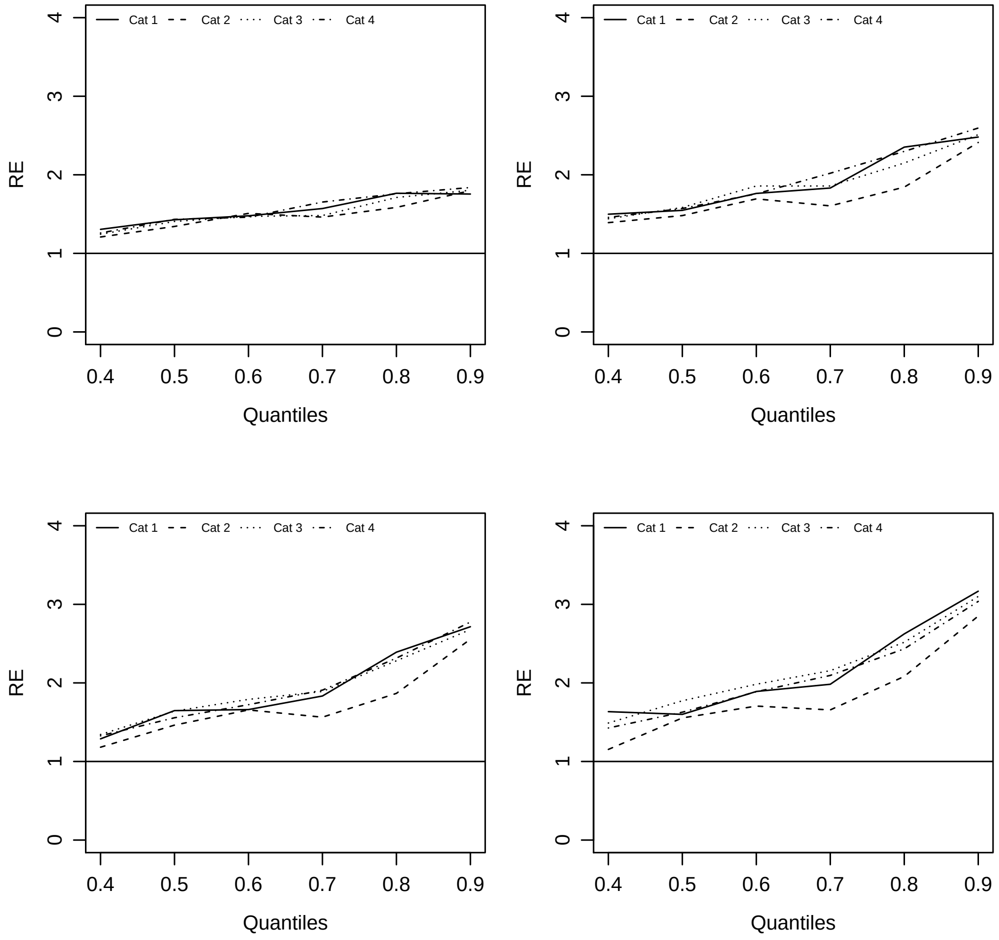

Figure 3.20: Relative Efficiency of SRS vs MNS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.8$

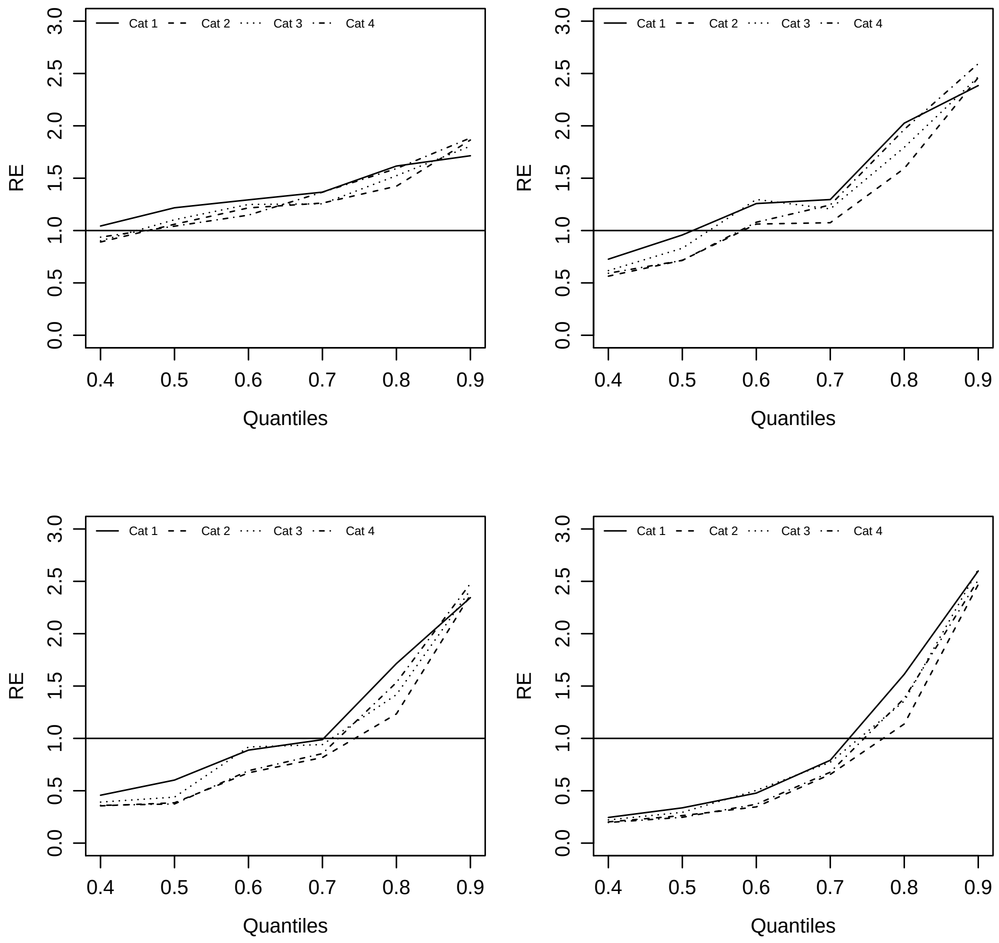

Figure 3.21: Relative Efficiency of SRS vs MNS treated as SRS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.8$

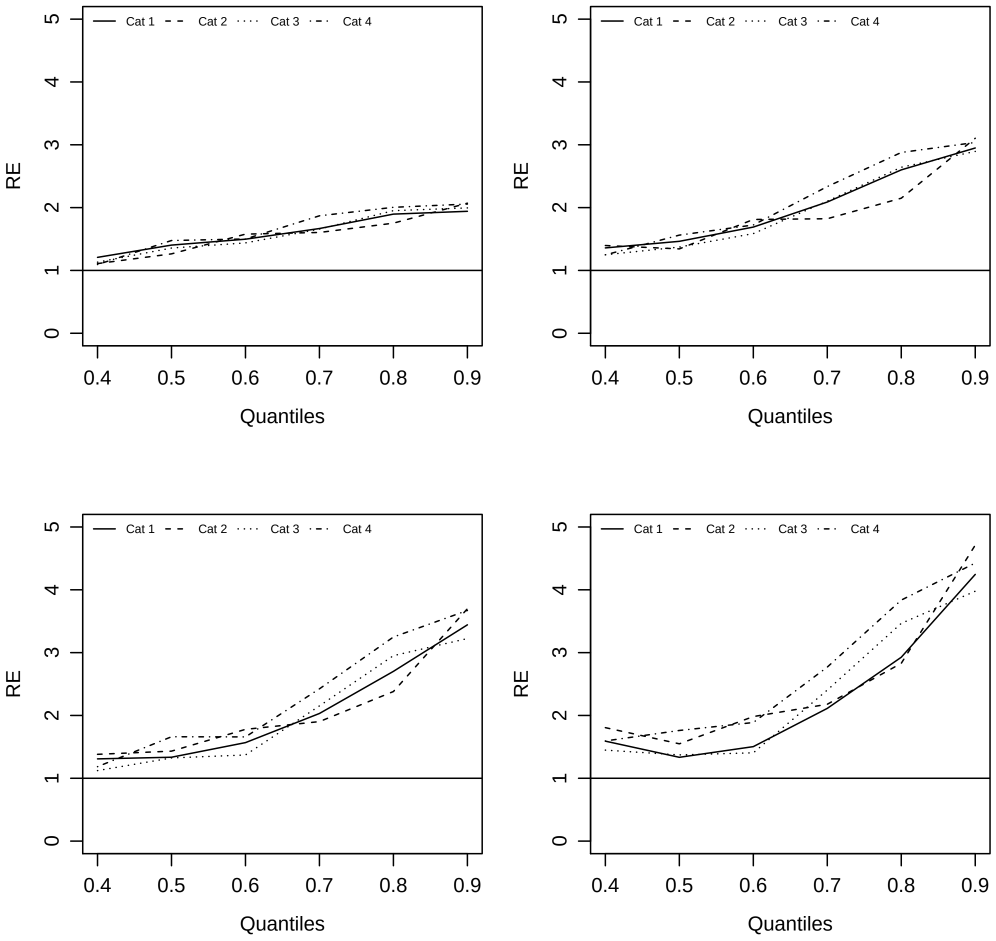

Figure 3.22: Relative Efficiency of SRS vs MNS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

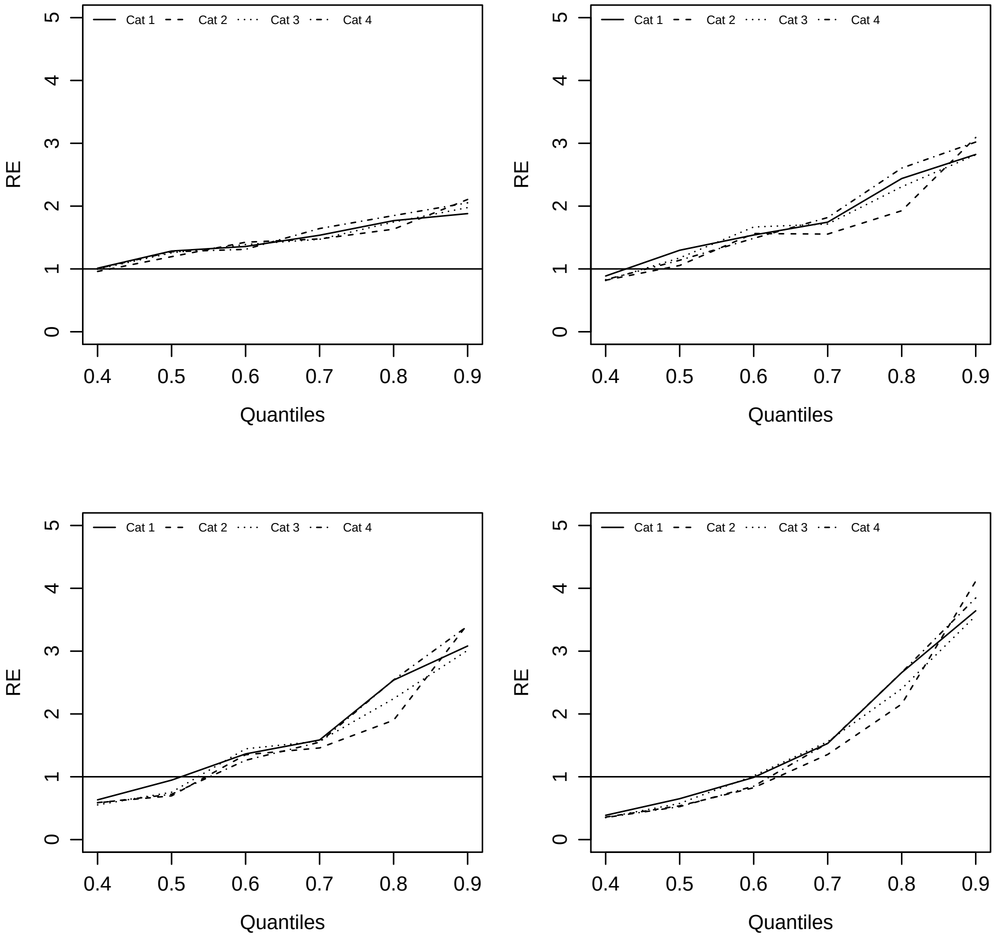

Figure 3.23: Relative Efficiency of SRS vs MNS treated as SRS of categorical and continuous covariates when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

3.4.3 Robustness

In previous sections, we generated ϵ's from a normal distribution implying that Y_i's are normally distributed in the underlying populations. In this section, we perform simulation studies to explore the effect of different distribution of ϵ on the performance of our proposed method. To this end, we repeat the simulation studies when ϵ's are assumed to follow a double exponential(0,2), chisq(3) and Exp(0.5). Fig.3.24 shows the histogram and the distribution of the Y values associated with four population generation using different distribution for ϵ. We only consider the case with categorical explanatory variable and work with the first method when $\rho = 0.9$. Results of the relative efficiency are presented accordingly. From the results obtained by assuming ϵ follows different distribution, it is evident that our proposed method is robust to different distribution of the underlying population.

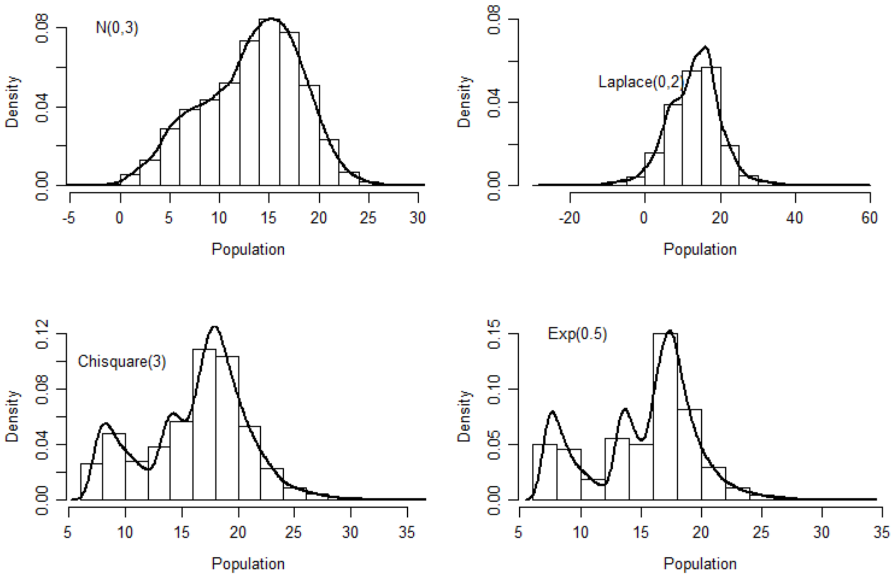

Figure 3.24: Histogram of the four distributions used in our simulation studies.

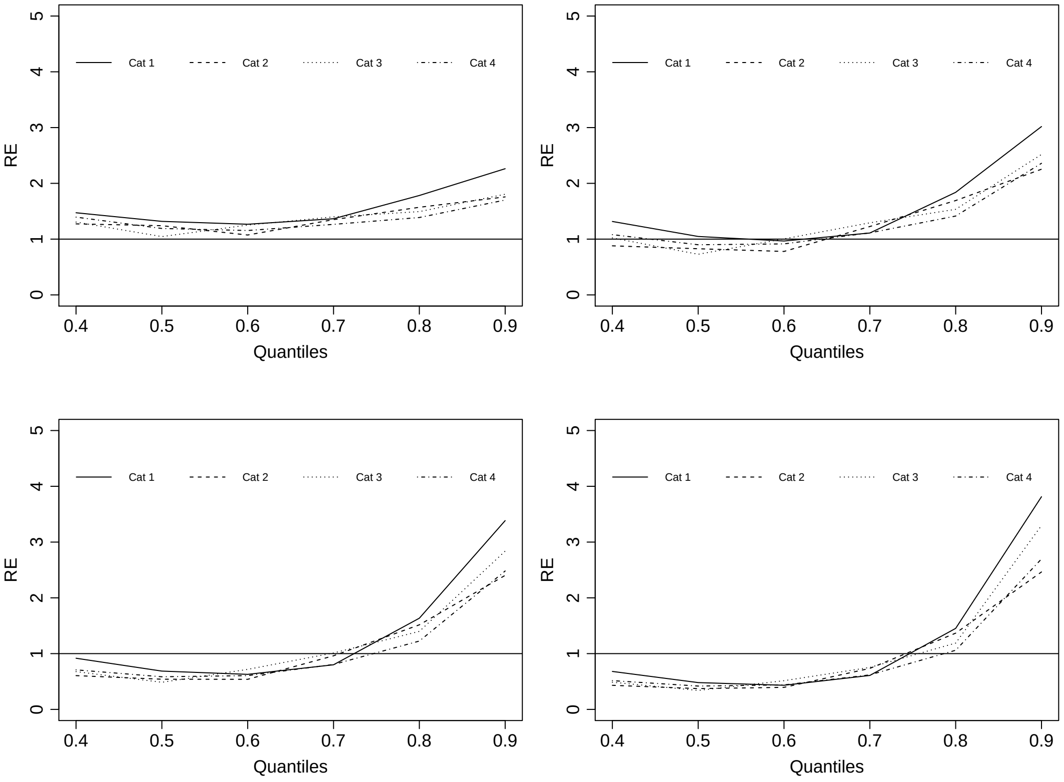

Figure 3.25: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$, $\rho = 0.9$ and we assume ϵ follows a Laplace distribution.

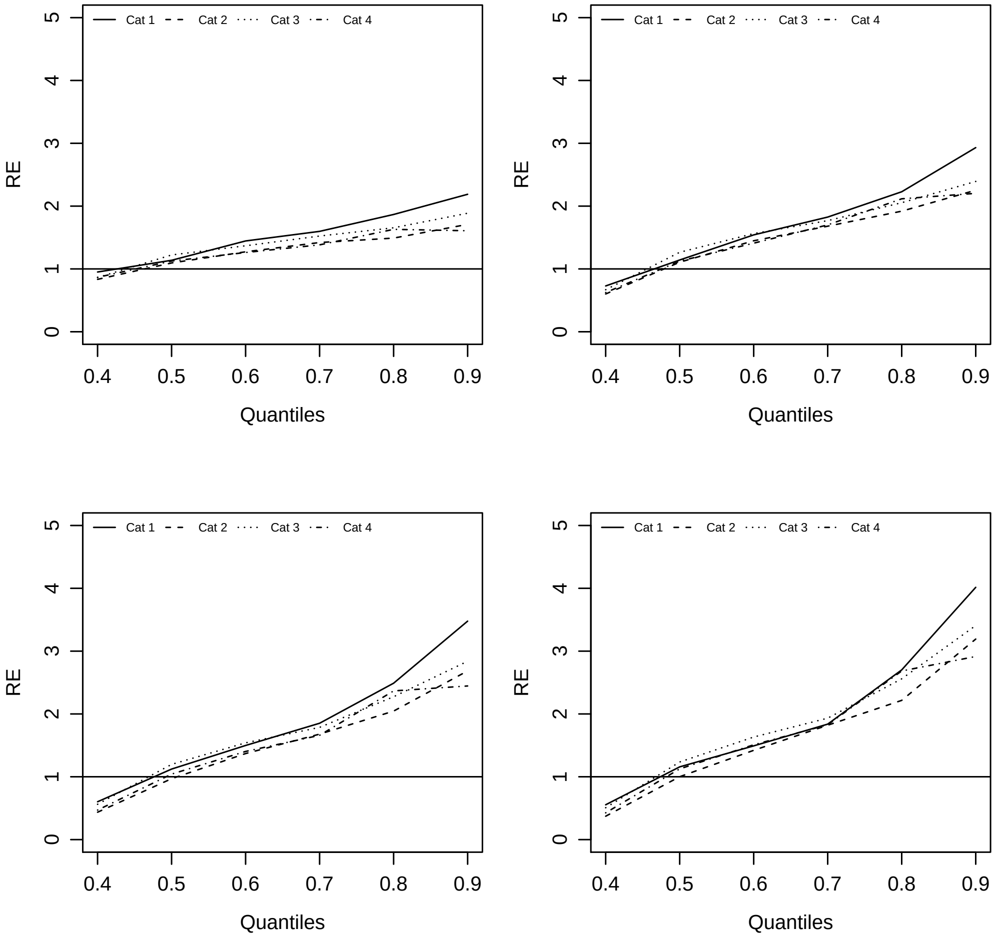

Figure 3.26: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$, $\rho = 0.9$ and we assume ϵ follows a Chi-square distribution.

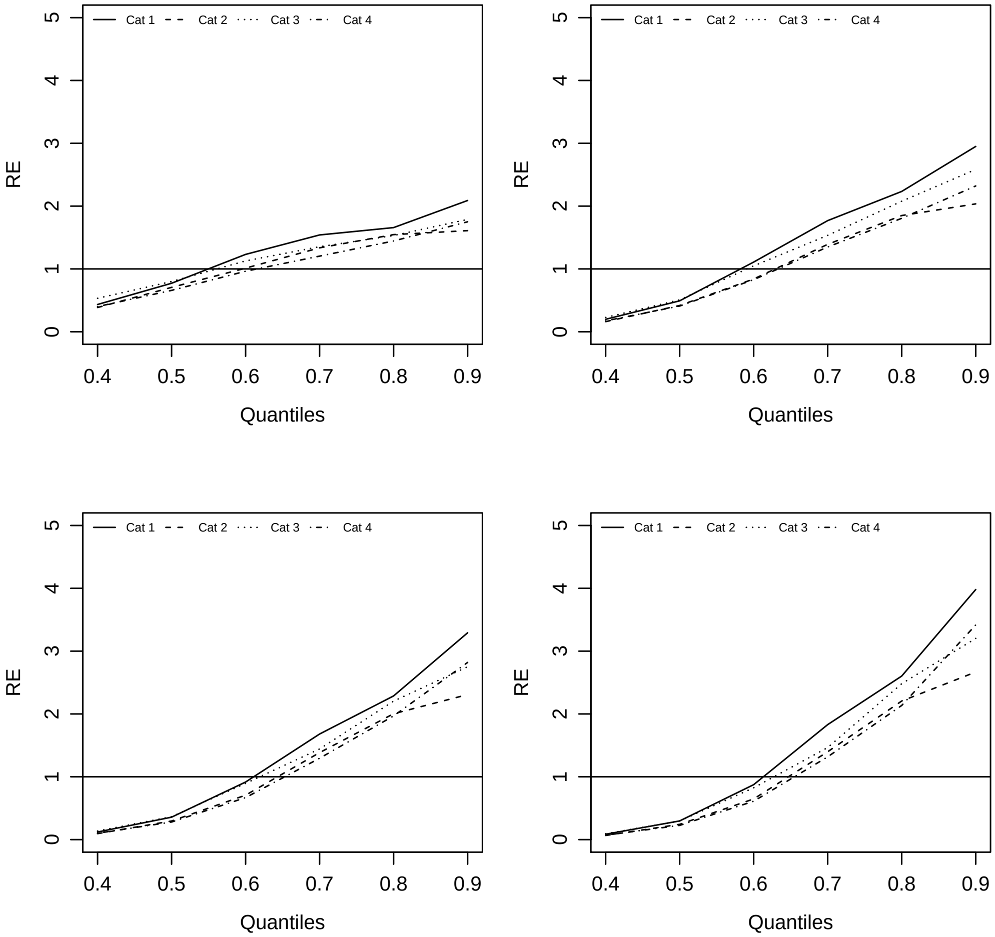

Figure 3.27: Relative Efficiency of SRS vs MNS when $k = (2, 3, 4, 5)$, $\rho = 0.9$ and we assume ϵ follows an Exponential distribution.

Chapter 4

Quantile Regression with rank based samples: An application to Bone Mineral Density

In this chapter, we apply our proposed method to a Bone mineral density (BMD) dataset collected in Manitoba since 2000 till present. Manitoba with a population of approximately 1.2 million provides health services to virtually all its residents through a public health care system. The BMD is measured using dual-energy x-ray absorptiometry (DXA). Measurement of the BMD is important for the diagnosis of osteoporosis which is often times based on finding low BMD from the DXA (Raisz, L., 2005). The cost of measuring the BMD of each patient is relatively high which requires a huge amount of money especially in a large cohort study. Our goal is to predict the lower and upper tail of the distribution of the follow-up BMD measurement as well as the BMD change of women using the baseline information as the covariates (note that the follow-up BMD measurements might change alot due to treatments received by patients). The objective here is to reduce the size of the follow-up in the cohort study due to budget constraints and still maintain the precision of our prediction. In the next section we present the summary of the data set.

4.1 Data Description

The underlying data set which we consider as our population consists of lumbar spine, femoral neck and total hip measurements of 10394, 10451 and 11069 women respectively. The summary statistics of the underlying population are reported in Table 4.1. Average weights and ages at baseline were $67.8kg$ and 61.9 respectively. The BMD testing interval is the difference between the first measurement (baseline) and next measurement (follow-up). This interval is categorized into three classes as; 1-3yrs, 4-5yrs and 6yrs above. The first, second and the third categories consists of 32.3%, 32.6% and 35.1% of the population respectively. The first quartile, median and the third quartile of the baseline lumbar spine BMD measurement are 0.91, 1.02 and 1.14 respectively. Figure 4.1 provides the distribution of the lumbar spine BMD for the baseline and the follow-up in three classes denoted by Cat1, Cat2 and Cat3. Figure 4.2 shows the scatter plots between the lumbar spine BMD in the baseline and the follow-up in three classes and their corresponding correlation coefficients. We also present the quantile of the follow-up population for the lumbar spine BMD, femoral neck BMD as well as total hip BMD for $\tau = (0.5, 0.6, 0.7, 0.8, 0.9, 0.95)$.

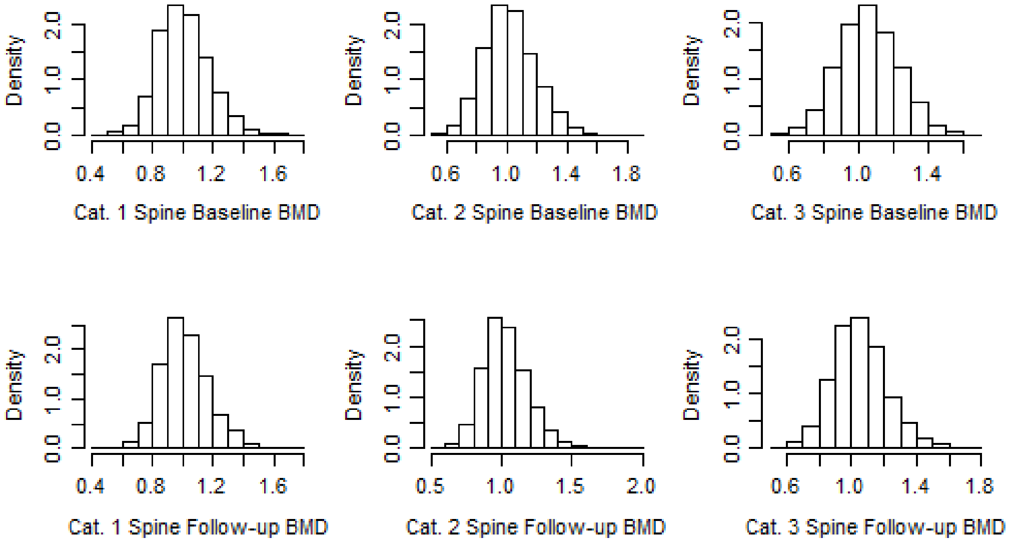

Figure 4.1: Histograms of the baseline and follow-up BMD measurements of the three categories respectively.

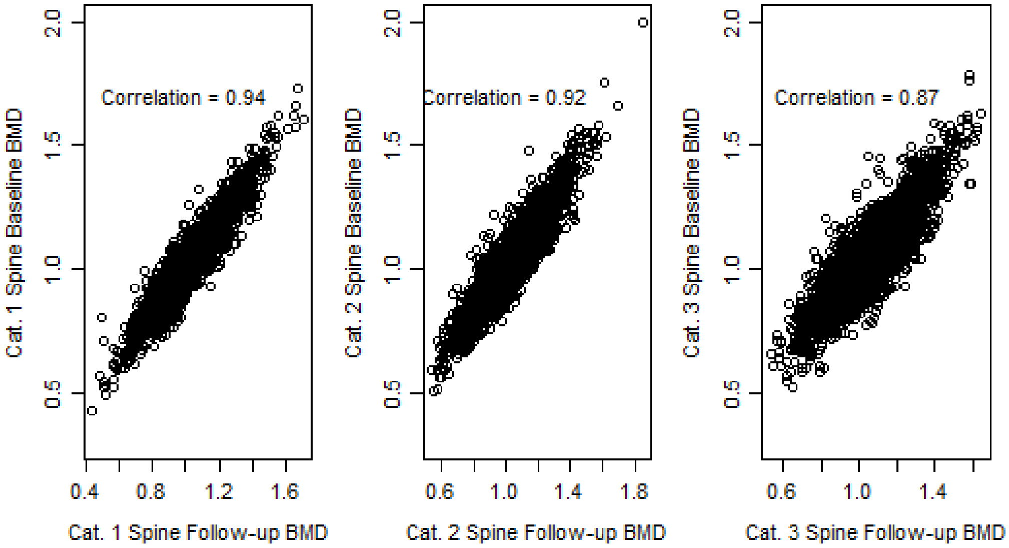

Figure 4.2: Scatter plots of the baseline and follow-up BMD measurements of the three categories respectively.

Table 4.1: Baseline and follow-up characteristics of the study population

Variables	Baseline $N = 10637$	First Follow-up $N = 10637$
Mean Age (SD)	61.9 (10.7)	66.9 (10.7)
Mean Weight (SD)	67.8 (13.3)	67.4 (13.8)
Hip Avg. fat (SD)	31.0 (5.8)	32.0 (6.1)
Spine Avg. fat (SD)	31.0 (10.1)	32.0 (10.3)
Mean BMD (SD)		
Lumbar Spine	1.03 (0.17)	1.03 (0.16)
Femoral Neck	0.84 (0.13)	0.82 (0.12)
Total Hip	0.88 (0.14)	0.86 (0.13)
BMD Interval, n (%)		
0-3yrs	3440 (32.3)	
4-5yrs	3472 (32.6)	
> 5yrs	3725 (35.1)	

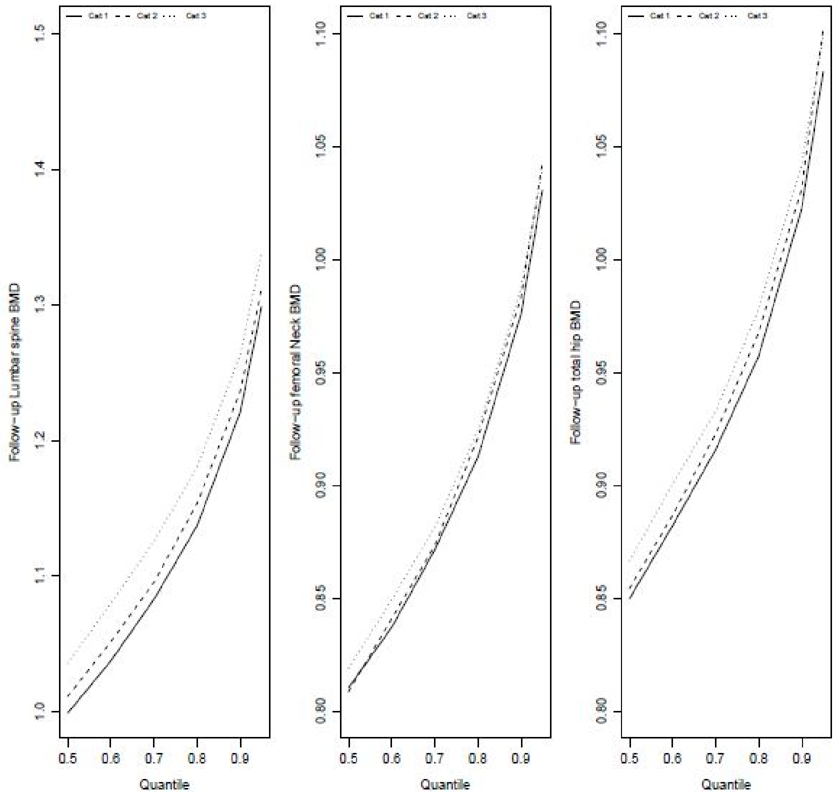

Figure 4.3: Quantiles of the follow-up population for the lumbar spine BMD, femoral neck BMD and Total hip BMD for $\tau = (0.5, 0.6, 0.7, 0.8, 0.9, 0.95)$

4.2 Results

In this section we present the results obtained using the minima as well as the maxima nominated samples to estimate the τ^{th} lower and upper quantiles of the variable of interest and study their behavior as a function of some auxillary information available in our data set.

We predict the follow-up lumbar spine BMD measurement using both methods as proposed in previous chapters. Significant covariates in our model for prediction are baseline age, baseline average hip fat, baseline spine BMD, baseline weight,BMD testing interval (difference between the baseline and first follow-up), baseline average spine fat, change in average hip fat, change in average spine fat and change in average weight.

The BMD testing interval is used as a categorical variable with three categories in both model. Category one includes patients with BMD testing interval less than three years. Category two includes patients with BMD testing interval greater than three years but less than or equal to five years and the third category includes patients with BMD testing interval greater than five years.

4.2.1 Quantile regression for lower quantiles of lumbar spine BMD

In clinical studies, BMD measurements are used to assess fracture risk, therefore it is important to understand the tail behavior of the BMD distribution. Estimating the τ^{th} quantile regression of the lower quantile of the lumbar spine BMD allows us to effectively study patients with low BMD, that is, patients that are likely to have bone fracture. We predict the follow-up lumbar spine BMD measurements for lower quantiles using minima nominated samples. The baseline lumbar spine BMD measurement was used for ranking because of its strong linear relationship with the follow-up lumbar spine BMD measurement. This approach is an extension of the quantile regression with maxima nominated samples discussed in Chapter 3.

To this end, let Y_i be the minimum of the sample $\mathbf{Y_i}$ of size k, $i = 1, \ldots, n$. Then Y_i has a distribution function of $G_Y(y) = 1 - [1 - F_Y(y)]^k$ and density $g_Y(y) = kf(y)[1 - F_Y(y)]^{k-1}$ for all $y \in R$. Now the τ^{th} quantile regression using minima nominated samples can be easily derived. The joint density for n independent copies of Y_i can be written as

$$\prod_{i=1}^{n} kf(y_i)[1 - F_Y(y_i)]^{k-1}. \tag{4.1}$$

Suppose it is known that $Y \sim AL(\mathbf{X}^\top \boldsymbol{\beta}, \sigma, \tau)$, using equation (3.3) and (3.7) we write the distribution function as substituting $F(y)$ and $f(y)$ in equation (4.1) we have

$L_\tau(\boldsymbol{\beta}, \sigma)$

$$= \prod_{i=1}^{n} \left[\frac{k\tau(1-\tau)}{\sigma} \exp\left(-\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) (\tau - 1) \right) \left[1 - \tau \exp\left(-\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) (\tau - 1) \right) \right]^{k-1} \right]^{I(Y_i < \mathbf{X}_i^\top \boldsymbol{\beta})}$$

$$\times \prod_{i=1}^{n} \left[\frac{k\tau(1-\tau)}{\sigma} \exp\left(-\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \tau \right) \left[-(1-\tau) \exp\left(-\left(\frac{Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\sigma} \right) \tau \right) \right]^{k-1} \right]^{I(Y_i > \mathbf{X}_i^\top \boldsymbol{\beta})}. \tag{4.2}$$

Now, the log-likelihood function is given by

$$
\begin{aligned}
& l_\tau(\boldsymbol{\beta},\sigma) \\
& \quad = n\ln\left(\frac{k\tau(1-\tau)}{\sigma}\right) - \sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)-(k-1)\left(\ln\left(1-\tau\exp\left(-(\tau-1)\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right)\right)\right] \\
& - \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[\tau\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)+(k-1)\left(\ln(1-\tau)-\left(\tau(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma})\right)\right)\right].
\end{aligned}
\tag{4.3}
$$

From equation (4.3), the τ^{th} regression quantile using minima nominated samples is obtained as the solution of

$$
\begin{aligned}
& \max_{\boldsymbol{\beta}\in R^p}\left[n\ln\left(\frac{k\tau(1-\tau)}{\sigma}\right) - \sum_{\{i:Y_i<\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[(\tau-1)\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)-(k-1)\left(\ln\left(1-\tau\exp\left(-(\tau-1)\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right)\right)\right]\right. \\
& \left. - \sum_{\{i:Y_i>\mathbf{X}_i^\top\boldsymbol{\beta}\}}\left[\tau\left(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma}\right)+(k-1)\left(\ln(1-\tau)-\left(\tau(\frac{Y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\sigma})\right)\right)\right]\right],
\end{aligned}
\tag{4.4}
$$

which can be completely optimized using similar approach discussed earlier.

We randomly select 200 samples from the BMD dataset using SRS and minima-nominated sampling approach, and estimate the τ^{th} quantile regression at $\tau = (0.1, 0.2, 0.3, 0.4, 0.5)$. Table 4.1 shows the estimated bias and MSE associated with each quantile. Also, we present the relative efficiency of our prediction using SRS compared with the proposed method in Figure 4.4. The MSE of the τ^{th} quantile regression at say, $\tau \leq 0.4$ of the proposed method is smaller than the MSE using SRS most especially when small set size is selected, say, $k = 2, 3$. This implies that at the lower quantiles even with an imperfect ranking, the proposed method outperforms the quantile regression using SRS.

The financial implication of this is simple. Suppose the DXA scan costs \$100 per person, and we take a simple random sample of 200 patients, and we estimate the MSE at different τ^{th} quantile regression with a reasonable precision. We require about one tenth of the sample used in SRS to estimate the MSE of the τ^{th} quantile regression with the same precision using our proposed method. We save about 90 percent of the total amount spent using the SRS approach.

Table 4.1: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.10, 0.20, 0.30, 0.50)$ quantile regression using simple random samples ($k = 1$) and minima-nominated samples with set size, $k = (2, 3, 4, 5)$ and $\rho = 0.9$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.10$		$\tau = 0.20$		$\tau = 0.30$		$\tau = 0.50$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.1232	0.0156	0.0878	0.0081	0.0603	0.0040	0.0112	0.0005
	2	0.1217	0.0153	0.0858	0.0078	0.0599	0.0040	0.0122	0.0006
	3	0.1106	0.0127	0.0777	0.0065	0.0511	0.0030	0.0113	0.0006
$k = 2$	1	0.0648	0.0045	0.0337	0.0014	0.0131	0.0004	0.0202	0.0008
	2	0.0644	0.0044	0.0346	0.0016	0.0191	0.0006	0.0237	0.0010
	3	0.0610	0.0040	0.0369	0.0017	0.0215	0.0008	0.0464	0.0026
$k = 3$	1	0.0409	0.0019	0.0172	0.0005	0.0192	0.0006	0.7102	0.5048
	2	0.0431	0.0021	0.0239	0.0008	0.0279	0.0010	0.6990	0.4890
	3	0.0444	0.0022	0.0316	0.0013	0.0430	0.0022	0.6920	0.4793
$k = 4$	1	0.0289	0.0010	0.0195	0.0006	0.2075	0.0432	1.2645	1.5998
	2	0.0335	0.0013	0.0286	0.0010	0.2047	0.0421	1.2565	1.5797
	3	0.0377	0.0016	0.0427	0.0021	0.2088	0.0438	1.2298	1.5131
$k = 5$	1	0.0227	0.0007	0.0445	0.0022	0.5062	0.2566	1.6950	2.8747
	2	0.0302	0.0011	0.0536	0.0031	0.5084	0.2588	1.6838	2.8369
	3	0.0363	0.0015	0.0783	0.0065	0.5018	0.2521	1.6596	2.7560

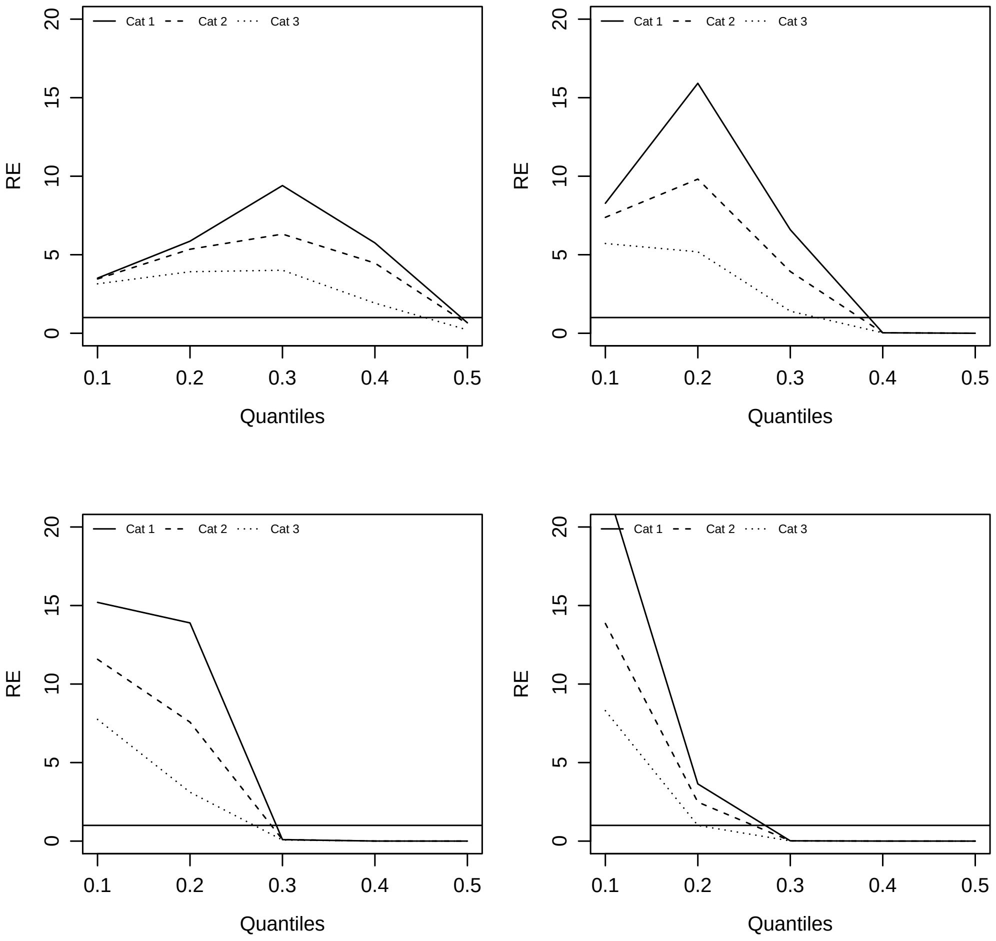

Figure 4.4: Relative Efficiency of SRS vs Minima-nominated sampling for lumbar spine BMD follow-up prediction, when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

4.2.2 Quantile regression for upper quantiles of lumbar spine BMD

In an effort to further understand the upper tail behavior of the lumbar spine BMD distribution, we predict the τ^{th} quantile regression of the upper quantiles of the follow-up lumbar spine BMD using maxima-nominated sample. Patients in these quantiles are less likely to fracture risk because of their high BMD measurement.

We randomly select 200 samples from the BMD dataset using SRS and MNS, and estimated the τ^{th} quantile regression at $\tau = (0.5, 0.6, 0.7, 0.8, 0.9)$. Table 4.2 provides the estimated bias and MSE associated with each quantile. Also, figure 4.5 shows the relative efficiency of the prediction of our method compared with the commonly used SRS approach.

The results obtained using the BMD dataset is closely related to what we established using our simulation study. The MSE of the τ^{th} quantile regression for $\tau \geq 0.7$ of the proposed method is smaller than the MSE associated with the τ^{th} quantile regression using SRS. This implies that at the upper quantiles even with an imperfect ranking, the proposed method outperforms the quantile regression using SRS. However, in this real data application, we observed that the relative efficiency nose dived at the upper quantile. This is as a result of the high variation in our dataset.

Table 4.2: Estimated Bias ($\widehat{Bias}$) and Estimated Mean Squared Error ($\widehat{MSE}$) for $Q_\tau(Y \mid \mathbf{X})$ when $\tau = (0.50, 0.60, 0.80, 0.90)$ quantile regression using simple random samples ($k = 1$) and maxima-nominated samples with set size, $k = (2, 3, 4, 5)$ and $\rho = 0.9$. The results are based on 5000 Monte Carlo replications for each sampling methods.

		$\tau = 0.50$		$\tau = 0.60$		$\tau = 0.80$		$\tau = 0.90$	
Set Size	Cat.	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{MSE}$
$k = 1$	1	0.0117	0.0005	-0.0140	0.0006	-0.0838	0.0075	-0.1445	0.0214
	2	0.0118	0.0006	-0.0139	0.0006	-0.0823	0.0073	-0.1382	0.0197
	3	0.0117	0.0006	-0.0146	0.0007	-0.0724	0.0057	-0.1211	0.0152
$k = 2$	1	0.0074	0.0005	0.0186	0.0007	-0.0294	0.0012	-0.0830	0.0073
	2	-0.0016	0.0004	0.0113	0.0005	-0.0311	0.0014	-0.0801	0.0069
	3	-0.0245	0.0011	-0.0060	0.0004	-0.0308	0.0014	-0.0719	0.0056
$k = 3$	1	-0.6645	0.4428	-0.1942	0.0382	-0.0089	0.0004	-0.0531	0.0032
	2	-0.6632	0.4411	-0.2045	0.0422	-0.0147	0.0006	-0.0529	0.0032
	3	-0.6712	0.4516	-0.2139	0.0462	-0.0219	0.0009	-0.0491	0.0028
$k = 4$	1	-1.2360	1.5301	-0.6470	0.4199	-0.0060	0.0004	-0.0364	0.0017
	2	-1.2113	1.4703	-0.6436	0.4155	-0.0132	0.0005	-0.0363	0.0017
	3	-1.1841	1.4047	-0.6581	0.4342	-0.0319	0.0014	-0.0377	0.0018
$k = 5$	1	-1.6526	2.7365	-0.9992	1.000	-0.0264	0.0011	-0.0255	0.0010
	2	-1.6099	2.5962	-0.9850	0.9721	-0.0370	0.0018	-0.0279	0.0011
	3	-1.6341	2.6743	-1.0047	1.0110	-0.0642	0.0045	-0.0328	0.0014

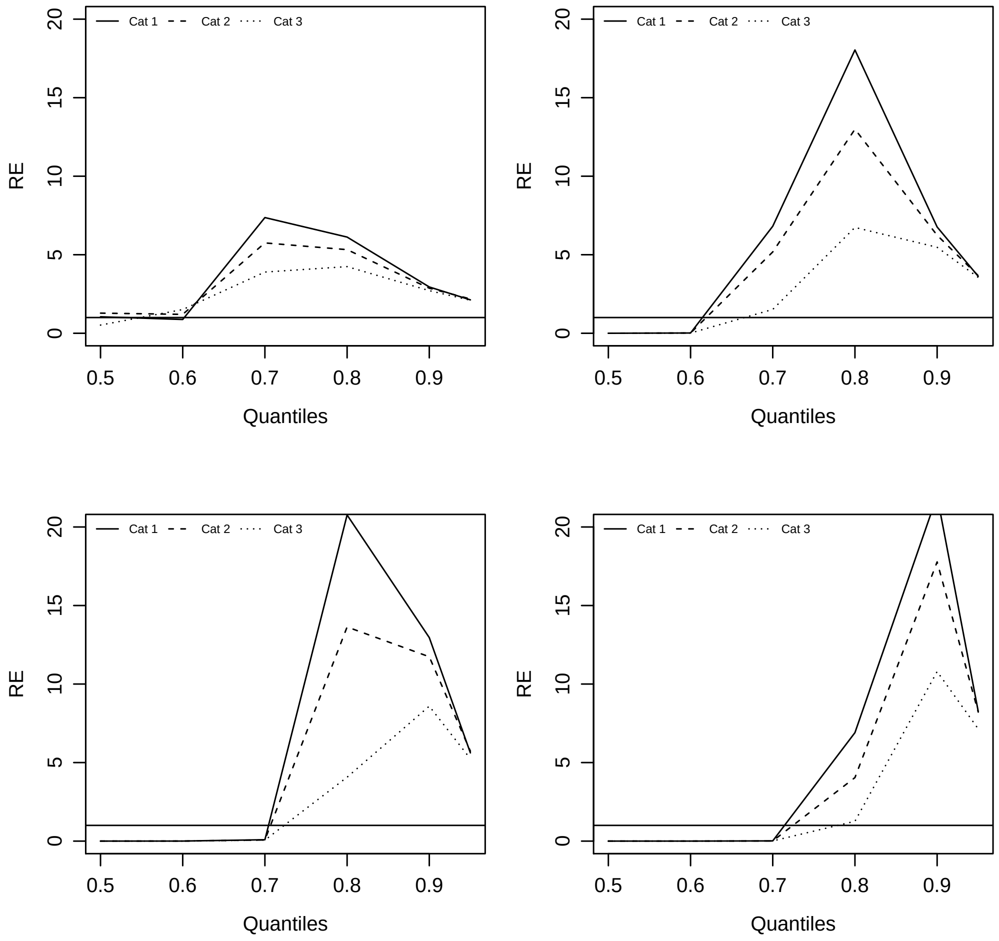

Figure 4.5: Relative Efficiency of SRS vs MNS for lumbar spine BMD follow-up prediction, when $k = (2, 3, 4, 5)$ and $\rho = 0.9$

Chapter 5

Concluding Remarks and Future Work

In this thesis, we proposed a new objective function which is built upon the distribution of minimum and maximum order statistics to estimate the τ^{th} regression quantile using minima and maxima-nominated samples. The MSE of our proposed method is compared with the MSE obtained using simple random samples, and we showed that the proposed method outperforms the estimates obtained using simple random samples, most especially at lower and upper quantiles of the distribution, say, $\tau \geq 0.7$, $\tau \leq 0.4$ respectively.

Also, we extended the result established in Nourmohammadi, M. et al. (2015) which showed the relationship between SRS and MNS to the quantile regression setting. In this case, we select samples from the population using MNS, samples obtained are then treated as simple random samples to estimate the τ^{th} regression coefficients. The MSE obtained using this approach is also shown to be smaller than the SRS for estimating lower or upper quantiles of the distribution of the variable of interest in the underlying population.

In Chapter 4, we applied our proposed methodology to the lumbar spine BMD measurement collected in Manitoba from 2000-2016. We predicted the lower and upper quantiles of the lumbar spine BMD follow-up measurements using the baseline BMD measurement as our concomitant variable. The results obtained are presented in tables and graphs. We proved that it is economical to use our proposed method by showing the relative efficiency of our method when compared with the established SRS approach. The analysis can be equally used to predict the follow-up of other BMD measurements in clinical studies.

It is important to note some of the limitations of our proposed method when applied on real data. The relationship between the concomitant and the response variable should be at least ≥ 0.7 to obtain better precision than estimates obtained using SRS. This limitation serves as a drawback for us to predict the change in BMD (as there were no variable in the real data that explained the change in BMD up to 70%), which is important in clinical studies. One way to address this issue is to use judgments from different experts for ranking. Also, the computation time in using our proposed method is longer when compared with quantile regression using SRS.

In the future, we would like to explore quantile regression with randomized nomination sampling which is a random combination of the minima and maxima NS in a finite population (Jafari Jozani, M. and Johnson, B., 2012). Also, we would like to study some statistical properties of the quantile regression "estimator" using rank-based samples. Another area of interest to explore in future is quantile regression with size-biased sampling to predict change in BMD measurement. In terms of real application which is of more interest from the clinical perspective, we would like to study change in BMD using our proposed approach. This is a more challenging problem that requires meticulous research.

Bibliography

Abreveya, J. (2001). The effects of demographics and maternal behaviour on the distribution of birth outcomes. *Empirical Economics 26*, 247–257. (Cited on page 4.)

Al-Omari, A.I. and Al-Nasser, A.D. (2012). Estimation of the population mean and median using truncation based ranked set samples. *Journal of Statistical Computation and Simulation*, 1–19. (Cited on page 33.)

Amiri, S., Jafari Jozani, M., and Moderres, R. (2013). Resampling unbalanced ranked set samples with application in testing hypothesis about the population mean. *Journal of Agricultural, Biological, and Environmental Studies 19*(1), 1–17. (Cited on pages 42 and 44.)

Austin, P.C. and Schull, M.J. (2003). Quantile regression: A statistical tool for out-of-hospital research. *Academic Emergency Medicine 10*(7), 789–797. (Cited on page 5.)

Bachrach, L.K., Hastie, T., Wang, M.C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy asian, hispanic, black and caucasian youth.

a longitudinal study. *Journal of Clinical Endocrinology and Metabolism* (84), 4702–12. (Cited on pages 2, 6 and 25.)

Barreto, M. C. M. and Barnett, V. (1999). Best linear unbiased estimators for the simple linear regression model using ranked-set sampling. *Environmental and Ecological Statistics* (6), 119–133. (Cited on page 31.)

Barrodale, I. and Roberts, F. D. K. (1973). An improved algorithm for discrete l1 linear approximation. *SIAM Journal of Numerical Analysis 10*, 839–848. (Cited on page 14.)

Bera, A. K., Antonio, F. G. Jr., Montes-Rojas, G. V., and Park, S. Y. (2016). Asymmetric laplace regression: Maximum likelihood, maximum entropy and quantile regression. *Journal of Econometric Methods 5*(1), 79–101. (Cited on page 51.)

Berman, A. (1973). *Cones, Matrices and Mathematical Programming.* New York: Springer. (Cited on page 12.)

Bianchi, A. and Salvati, N. (2015). Asymptotic properties and variance estimators of the m-quantile regression coefficients estimators. *Communications in Statistics-Theory and Methods 44*(11), 2416–2429. (Cited on page 15.)

Boyles, R.A. and Samaniego, F.J. (1986). Estimating a distribution function based on nomination sampling. *Journal of the American Statistical Association 81*(396), 1039–1045. (Cited on page 48.)

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika 75*(4), 761–771. (Cited on pages 15 and 54.)

Chambers, R. and Tzavidis, N. (2006). M-quantile model for small area estimation. *Biometrika 93*(2), 255–268. (Cited on page 15.)

Chen, C. (2003). A finite smoothing algorithm for quantile regression. *Preprint.*. (Cited on page 10.)

Chen Zehua, Zhidong Bai, and Bimal K. Sinha (2004). *Ranked Set Sampling: Theory and Applications*, Volume 176. Newyork: Springer. (Cited on pages 33, 34, 36, 37, 40 and 42.)

Cole, T.J. and Green, P.J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine 11*, 1305–1319. (Cited on page 5.)

Cristina, D., Marilena, F., and Domenico, V. (2013). *Quantile Regression Theory and Applications.* WILEY. (Cited on page 22.)

Dell, T.R. and Clutter, J.L. (1972). Ranked set sampling theory with order statistics background. *Biometrics 28*, 545–555. (Cited on pages 31, 34, 57 and 60.)

Dresassi, E., Ranalli, M.G., and Salvati, N. (2014). Semiparametric m-quantile regression for count data. *Statistical Methods in Medical Research 23*(6), 591–610. (Cited on page 15.)

Efron, B. and Tibshirani, R.J. (1998). *An Introduction to the Bootstrap.* CRC Press LLC, Boca Raton, FL. (Cited on page 22.)

Farcomeni, A (2012). Quantile regression for longitudinal data based on latent markov subject-specific parameters. *Statistics and Computing 22*(1), 141–152. (Cited on page 51.)

Geraci, M. and Bottai, M. (2006). Quantile regression for longitudinal data using the asymmetric laplace disribution. *Biostatistics 8*(1), 140–154. (Cited on page 51.)

Gill, P., Murray, W., and Wright, M. (1991). *Numerical Linear Algebra and Optimization*. Redwood City, CA: Addison Wesley. (Cited on page 14.)

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Statistics 35*(1), 73–101. (Cited on pages 15 and 54.)

Jafari Jozani, M. and Johnson, B. (2011). Design based estimation for ranked set sampling in finite populations. *Environmental and Ecological Statistics 18*(4), 663–685. (Cited on page 31.)

Jafari Jozani, M. and Johnson, B. (2012). Randomized nomination sampling for finite populations. *Journal of Statistical Planning and Inference 142*, 2103–2115. (Cited on pages 48 and 110.)

Jafari Jozani, M. and Mirkamali, S.J. (2011). Control charts for attributes with maxima nominated samples. *Journal of Statistical Planning and Inference 141*, 2386–2398. (Cited on page 48.)

Kocherginsky, M. and He, X. (2007, July). Extensions of the markov chain marginal bootstrap. *Probability Letters 77*(12), 1258–1268. (Cited on page 22.)

Koenker, R. and Bassett, G. (1978, January). Regression quantiles. *Econometrica 46*(1), 33–50. (Cited on pages 1, 4 and 16.)

Koenker, R. and Bassett, G. (1982, January). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica 50*(1), 43–61. (Cited on pages 20 and 21.)

Koenker, R. and D'Orey, V. (1985). Computing regression quantiles. Technical Report 1141, University of Illinois at Urban Champaign, April. (Cited on page 14.)

Koenker, R. and Machado, J.A.F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* (94), 1296–1310. (Cited on pages 21 and 50.)

Koenker Roger (2005). *Quantile Regression.* Cambridge University Press. (Cited on pages 4, 5, 11 and 14.)

McIntyre, G.A. (1952). A method of unbiased selective sampling, using ranked sets. *Aust. J. Agri. Res.* (3), 385–390. (Cited on pages 6, 30 and 31.)

Modarres, R., Hui, T., and Gang, Z. (2006). Resampling methods for ranked set samples. *Computational Statistics and Data Analysis* (51), 1039–1050. (Cited on pages 43 and 44.)

Muttlak, H. A. (1995). Parameters estimation in a simple linear-regression using rank set sampling. *Biometrical Journal,* (37), 799–810. (Cited on page 31.)

Muttlak, H. A. (1996). Estimation of parameters of one-way layout with rank set sampling. *Biometrical Journal 38*, 507–515. (Cited on page 31.)

Muttlak, H. A. (1998). Median ranked set sampling with size biased probability of selection. *Biometrical Journal 40*(4), 455–465. (Cited on page 33.)

Nourmohammadi, M., Jafari Jozani, M., and Brad, J. (2015). Nonparametric confidence intervals for quantiles with randomized nomination sampling. *Sankhya 77*(2), 408–432. (Cited on pages 49, 52, 55, 57 and 109.)

Ozturk, O. (2002, December). Rank regression in ranked-set samples. *Journal of the American Statistical Association 97*(460). (Cited on page 31.)

Park, S. (1996). Fisher information in order statistics. *Journal of American Statistical Association 91*(433), 385–390. (Cited on page 33.)

Philip, L.H. and Lam, K. (1997). Regression estimator in ranked set sampling. *International Biometric Society 53*(3), 1070–1080. (Cited on page 31.)

Pratesi, M., Ranalli, M. G., and Salvati, N. (2009). Nonparametric m-quantile regression using penalized splines. *Journal of Nonparametric Statistics 21*(3), 287–304. (Cited on page 15.)

Raisz, L. (2005). Clinical practice. screening for osteoporosis. *The New England Journal of Medicine 352*(2), 164–171. (Cited on page 96.)

Samawi,H. M., Ahmed, M.S., and Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal 38*(5), 577–586. (Cited on pages 31 and 33.)

Stephen, P. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science 12*(4), 279–300. (Cited on page 15.)

Stigler, G. (1945). The cost of subsistence. *Journal Farm Economics* (37), 1249–1258. (Cited on page 11.)

Terpstra, T. and P. Wang (2008). Confidence intervals for a population proportion based on a ranked set sample. *Journal of Statistical Computation and Simulation 78*(4), 351–366. (Cited on page 31.)

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Stat Med 25*(8), 1369–1382. (Cited on page 5.)

Wells, M.T. and Tiwari, R.C. (1990). Estimating a distribution function based on minima-nomination sampling. *Institute of Mathematical Statistics*, 471–479. (Cited on page 48.)

Willemain, T.R. (1980). Estimating the population median by nomination sampling. *Journal of the American Statistical Association 75*(372). (Cited on pages 47 and 48.)

Wolfe, D.A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *International Scholarly Research Network*. (Cited on page 35.)

Yu, K. and Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics and Probability Letters 54*, 437–447. (Cited on pages 50 and 51.)

Yu, K. and Stander, J. (2007). Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics 137*(1), 260–276. (Cited on page 51.)

Yu, K. and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics- Theory and Methods 34*, 1867–1879. (Cited on page 50.)