

Spring 2021

A Simulation-Based Study of Location-Shift Models Under Non-Normal Conditions

Ummay Khayrunnesa Anika

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>



Part of the [Biostatistics Commons](#)

Recommended Citation

Anika, U. K. (2021). *A Simulation-Based Study of Location-Shift Models Under Non-Normal Conditions*. (Master's thesis). Retrieved from <https://scholarcommons.sc.edu/etd/6369>

This Open Access Thesis is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.

A SIMULATION-BASED STUDY OF LOCATION-SHIFT MODELS UNDER NON-NORMAL
CONDITIONS

by

Ummay Khayrunnesa Anika

Bachelor of Science
University of Dhaka, 2017

Submitted in Partial Fulfillment of the Requirements
For the Degree of Master of Science in Public Health in
Biostatistics

Arnold School of Public Health
University of South Carolina

2021

Accepted by:

Marco Geraci, Director of Thesis

Andrew Ortaglia, Reader

Stella Self, Reader

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

© Copyright by Ummay Khayrunnesa Anika, 2021
All Rights Reserved.

Acknowledgements

I am whole-heartedly grateful to my advisor and mentor Dr. Marco Geraci, who expertly navigated me in the proceedings of my MS thesis. Without his prudential guidance and valuable suggestions, it would have been difficult for me to finish this thesis properly.

I would also like to express my gratitude to Dr. Andrew Ortaglia and Dr. Stella Self for serving on my thesis committee. I am sincerely thankful to them for their suggestions, comments, and valuable time. I will be forever grateful to my graduate director Dr. Robert Moran for his support and co-operation throughout my MS journey.

Finally, I would like to thank my parents for their great blessings. A special thank goes to my husband Md Nazmul Al Imran, who always inspires me with great care and well-wishing.

Abstract

In this study, we compare ordinary least squares (OLS), generalized least squares (GLS), M- and quantile regression (QR) estimators for a continuous response variable under different scenarios by conducting a simulation study. We assess the performance of the estimators in terms of bias, average distance, mean squared error, coverage probability, and ratio of estimated standard error and empirical standard deviation. OLS estimator performs the best when the errors are homoscedastic normal or homoscedastic but skewed (exponential) having no outliers. GLS estimator shows good comparative results to QR when the errors are heteroscedastic normal or heteroscedastic heavy-tailed (t-distributed). The most satisfactory performance of the M-estimator is revealed when the errors are homoscedastic heavy-tailed with no outliers, and homoscedastic normal or homoscedastic exponential contaminated with outliers. In all of the scenarios with heavy-tailed-skewed (log-normal) errors, the QR estimator is shown to be more accurate and stable than the other estimators. Moreover, as a robust estimator, both M- and QR estimators become more reasonable than the others in scenarios with outliers contaminated errors which is also evident from real data analysis.

Table of Contents

Acknowledgements	iii
Abstract	iv
List of Tables	vii
List of Figures	x
Chapter 1: Introduction	1
1.1 Background	1
1.2 Objective of the Study	2
1.3 Outline of the Study	3
Chapter 2: Methodology	4
2.1 Ordinary Least Squares	4
2.2 Generalized Least Squares	6
2.3 M-Estimator	6
2.4 Quantile Regression Estimator	8
2.5 Criteria of Assessment	9
Chapter 3: Simulation Study	12
3.1 Simulation Procedure	12
3.2 Simulation Results	15
Chapter 4: Data Analysis	22

4.1	Exploratory Analysis	22
4.2	Regression Analysis	24
	Chapter 5: Discussion and Conclusions	28
	Bibliography	30
	Appendix A: Simulation Study Results	32
	Appendix B: Diagnostic Results for TcCB (ppb) Concentrations Data	47

List of Tables

Table 3.1	Summary of scenarios	14
Table 3.2	Summary of the results	16
Table 4.1	Descriptive statistics of TcCB (ppb) concentrations by Area. . .	24
Table 4.2	Estimates of the intercept and slope along with their estimated standard errors (given in parenthesis) for 4 different estimators (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with TcCB concentrations (ppb) data (TcCB \sim Area (ref: “Reference”); significant estimates are marked in bold)	25
Table 4.3	Summary statistics of TcCB (ppb) by Area after deletion of outliers from the data.	26
Table A.1	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 1 (normal, homoscedastic, no outliers).	33
Table A.2	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 2 (t_2 (heavy-tailed), homoscedastic, no outliers).	34
Table A.3	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 3 (exponential, homoscedastic, no outliers).	35

Table A.4	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 4 (normal, heteroscedastic, no outliers).	36
Table A.5	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 5 (normal, homoscedastic, outliers).	37
Table A.6	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 6 (log-normal, homoscedastic, no outliers).	38
Table A.7	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 7 (t_3 (heavy-tailed), heteroscedastic, no outliers).	39
Table A.8	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 8 (exponential, heteroscedastic, no outliers).	40
Table A.9	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 9 (exponential, homoscedastic, outliers).	41

Table A.10	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 10 (normal, heteroscedastic, outliers).	42
Table A.11	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 11 (log-normal, heteroscedastic, no outliers).	43
Table A.12	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 12 (log-normal, homoscedastic, outliers).	44
Table A.13	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 13 (exponential, heteroscedastic, outliers).	45
Table A.14	Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 14 (log-normal, heteroscedastic, outliers).	46
Table B.1	DFBETAS of Area with corresponding influential observations of TcCB (ppb) concentrations.	47

List of Figures

Figure 4.1	Distribution of TcCB concentrations (ppb) (before and after log-transformation).	24
Figure B.1	Plot of DFBETAS with a cutoff 0.18 to detect influential observation for Area (slope) of OLS model ($TcCB \sim Area$).	47

Chapter 1

Introduction

1.1 Background

Regression analysis is a very well known technique to explore the relationship between response and predictor variables. Traditional regression analysis summarizes the relationship between a set of predictors and the expected value of the response variable. The idea of modeling the conditional mean function is the foundation of a broad class of regression modeling approaches, including linear and generalized linear regression. Conditional mean modeling has been used broadly in different fields such as engineering, medical, and social sciences.

The conditional mean estimation is tied to least squares and maximum likelihood (ML) estimation. ML estimators are the best linear unbiased estimators (BLUE) as long as there are no serious violations of the model assumptions. Ordinary least squares (OLS) is the most common estimation technique used in classical linear regression analysis when the errors are independently and identically distributed, have constant variance (homoscedasticity), and are distributed according to the normal distribution [7]. But, there are many real situations in which the errors are not normal and have non-constant variability (heteroscedasticity). For example, the error distribution may be skewed, heavy-tailed, or heavy-tailed and skewed. Therefore, some alternative estimation methods should be adopted that are more robust in terms of departures from normal conditions.

Generalized or weighted least squares (WLS) is a modification of OLS that can be applied to account for non-constant variance [5]. WLS is also used iteratively in

generalized linear models [17] where heteroscedasticity follows from several exponential families. On the other hand, to mitigate the influence of the outliers, a robust estimation technique should be considered. A robust estimation has less restrictive assumptions than the least-squares estimation. A robust estimation procedure is insensitive to outliers and produces essentially the same results as least squares when the underlying distribution is normal and there are no outliers [19]. The most common method of robust regression is M-estimation introduced by Huber [15] which is used in our study.

An estimation method is highly desirable that deals with multiple violations of normality conditions. Quantile regression (Koenker and Bassett (1978)) can serve very well in this regard. Quantile regression provides a distribution-free approach to the modeling and estimation of the effects of covariates on different quantiles of the conditional distribution of a continuous response variable [16]. Since the median is a specific quantile, then conditional median regression is a special case of quantile regression in which the conditional 50th percentile of the response is modeled as a function of predictors. When the distribution of the response is skewed then the interpretation of the mean can be challenging while the median remains highly informative due to its robustness properties.

Applications of quantile regression are growing rapidly due to its advantage in investigating the impact of predictor variables on the entire distribution of the response. Applications range from the study of the conditional distribution of wages [4, 6] and schooling [10] to the demographics' effects on the distribution of infant birth weight [1]. Quantile regression has been applied in sociology [13, 14], ecological and environmental sciences [20, 3], and public health [2, 12, 18].

1.2 Objective of the Study

The main goal of this study is to compare different estimators for location-shift models (OLS, GLS, M- and quantile estimators) under different simulated scenarios.

We also conduct a real data analysis to compare those estimators. The response variable, 1,2,3,4-Tetrachlorobenzene (TcCB) concentrations (ppb) was measured from soil samples at a reference area and a cleanup area. Previously, the quantile test as a test for multiple outliers was performed in the cleanup-unit data set, where the standard for comparison is the data set for the site-specific reference area [11]. In our study, we make a contrast between quantile regression estimator and mean regression estimators by modeling TcCb as a function of area where reference area is the baseline category.

1.3 Outline of the Study

This study has been organized into five chapters. In Chapter 2, we discuss the methodology. The general form of linear regression models is introduced. The estimation methods of each estimator considered in our study are described along with their theoretical properties. We also introduce the criteria for the assessment of the estimators. Chapter 3 describes the simulation study along with simulation results. In Chapter 4, we perform a real data analysis using the estimators considered in the simulation study. In Chapter 5, we discuss the findings and draw conclusions based on the results.

Chapter 2

Methodology

Linear regression (LR) is a commonly used statistical technique. LR models the conditional mean of a continuous response variable. On the other hand, quantile regression (QR) facilitates the analysis of the full conditional distribution of the response variable. QR is an alternative approach to LR that is capable of handling heteroscedasticity, outliers and detects various forms of shape changes [16]. Generalized least squares (GLS) is a technique for estimating the parameters in a linear regression model when the variance of the error is not constant and there is a certain degree of correlation between the residuals in a regression model. Robust estimation methods such as M-estimation are used in the presence of outliers (heavy-tailed distribution).

2.1 Ordinary Least Squares

Consider a random sample of observations $\{y_i, x_{i1}, \dots, x_{ik}\}_{i=1}^n$ of size n ; a LR assumes that the relationship between the continuous response variable y_i and the k -vector of covariates or independent variables x_i is linear. This relationship is modeled through a stochastic error term ϵ_i which adds noise to the linear relationship between the response variable and covariates. Thus the model can be defined as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n, \quad (2.1)$$

where $\boldsymbol{\beta}$ is the $k \times 1$ dimensional vector of parameters of interest. In matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.2)$$

$$\text{where } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The main objective of LR is to estimate the parameter vector $\boldsymbol{\beta}$. Some assumptions are typically introduced. One of the assumptions is that $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ and the errors are uncorrelated.

OLS is the most common estimation technique which is conceptually simple and computationally straightforward for LR. The OLS method minimizes the sum of squared residuals (SSR), and leads to a closed-form expression for the estimated value of the unknown parameter $\boldsymbol{\beta}$ [8, 9]. Then using the analogue principle, the objective function is given by

$$L_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2, \quad (2.3)$$

Now, for the defined model (2.2)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \left(\sum \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum \mathbf{x}_i y_i \right). \quad (2.4)$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$ with vector of size $k \times 1$, normally distributed with mean $\boldsymbol{\beta}$ and finite variance $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. OLS estimators are said to be the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ as they are linear, unbiased and have minimum variance in the class of all such linear unbiased estimators. The OLS estimator is consistent if the errors have finite variance and are uncorrelated with the regressors $E[\mathbf{x}_i \varepsilon_i] = 0$. The OLS estimator is also the maximum likelihood estimator under the assumption that the errors are normally distributed.

2.2 Generalized Least Squares

The generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ of a linear regression is a generalization of the OLS estimator. Let \mathbf{V} be a symmetric nonsingular $n \times n$ covariance matrix. Now we have the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (2.5)$$

where $E[\boldsymbol{\varepsilon}] = 0$ and $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{V}$. The GLS method estimates $\boldsymbol{\beta}$ by minimizing the squared Mahalanobis length of the residual vector $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. The objective function and estimator are thus given by

$$L_{GLS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (2.6)$$

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, \quad (2.7)$$

respectively [5]. The GLS estimators are unbiased, consistent, efficient, and asymptotically normally distributed with $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$ and $\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$. If there is no correlation between the residuals, but still have heteroscedasticity then weighted least squares (WLS) is used which is a special case of GLS.

2.3 M-Estimator

This class of estimators can be regarded as a generalization of maximum-likelihood estimation where M stands for maximum likelihood, introduced by Huber [15]. The estimates of $\boldsymbol{\beta}$ are determined by minimizing an objective function

$$L_M(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\gamma(u_i) = \sum_{i=1}^n \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}), \quad (2.8)$$

where

$$\rho_\gamma(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \leq \gamma \\ \gamma|u| - \frac{1}{2}u^2, & \text{if } |u| > \gamma \end{cases} \quad (2.9)$$

is the dispersion function and γ is a tuning constant that reflects the possible proportion of outliers in the data. Smaller positive values of γ produce more resistance to outliers. But, if the errors are normally distributed, the choice of smaller γ reduces the estimation efficiency. This study uses the value of γ as 1.345.

Let $\Psi = \rho'$ be the derivative of ρ . Ψ is called the influence function. A system of $(k + 1)$ estimating equations for the coefficients is obtained by differentiating ρ with respect to $\boldsymbol{\beta}$ and setting the partial derivatives to 0:

$$\sum_{i=1}^n \Psi(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i^T = \mathbf{0}. \quad (2.10)$$

Let define the weight function $w(u) = \Psi(u)/u$ and let $w_i = w(u_i)$. Then the estimating equations can be written as

$$\sum_{i=1}^n w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i^T = \mathbf{0}, \quad (2.11)$$

where

$$w(u) = \begin{cases} 1, & \text{if } |u| \leq \gamma \\ \frac{\gamma}{|u|}, & \text{if } |u| > \gamma. \end{cases} \quad (2.12)$$

But, the weights rely on the residuals, the residuals rely on the estimates, and the estimates rely on the weights. Therefore, an iterative algorithm technique called iteratively re-weighted least squares (IRLS) is used to solve the equations defined

in Equation (2.11) as no closed form solution exists. At iteration t , the estimated coefficient matrix is given by

$$\boldsymbol{\beta}_{(t)} = (\mathbf{X}^T \mathbf{W}_{(t-1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{(t-1)} \mathbf{y}. \quad (2.13)$$

Iterations are continued until the estimated coefficients converge. M-estimators are consistent and asymptotically normally distributed.

2.4 Quantile Regression Estimator

QR assumes no parametric form for the conditional distribution of the response. For any $\tau \in (0, 1)$, the τ^{th} conditional quantile function can be written as

$$Q_Y(\tau \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau, \quad (2.14)$$

where $\boldsymbol{\beta}_\tau$ is the τ^{th} quantile regression coefficients. The QR estimator $\hat{\boldsymbol{\beta}}_\tau$ can be obtained by solving

$$L_{QR}(\boldsymbol{\beta}_\tau) = \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}), \quad (2.15)$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is called the check function and $I(\cdot)$ is the indicator function. Linear programming algorithms are applied to estimate the parameters for quantile regression.

A property of the τ^{th} quantile regression estimator is that the proportion of data points lying below the fitted line is τ , and the proportion lying above is $1 - \tau$. Under some regularity conditions, QR estimators are consistent and asymptotically normal.

2.5 Criteria of Assessment

We evaluate the performance of the estimators for each scenario and sample size considered in our simulation study by using the criteria described below. Let θ be a true value of a parameter and $\hat{\theta}$ be an estimator for θ or the actual estimate.

Bias: For any estimator $\hat{\theta}$, bias can be defined by

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta, \quad (2.16)$$

which represents the deviation of results from the truth. If $E(\hat{\theta}) = \theta$, the estimator $\hat{\theta}$ is said to be unbiased. Unbiasedness is one of the desirable properties of an estimator.

Average distance: In addition to bias, another interesting criteria is average distance between the point estimate and the parameter being estimated. By taking the absolute value of the difference between the estimator and the true parameter for each simulation, and then averaging that across all simulations, we calculate this measure. For R replications, average distance of $\hat{\theta}$ can be presented as

$$Average\ distance(\hat{\theta}) = \frac{\sum_{i=1}^R |\hat{\theta}_i - \theta|}{R}. \quad (2.17)$$

Variance: The variance of an estimator is a measure of precision given by

$$Var(\hat{\theta}) = E\{\hat{\theta} - E(\hat{\theta})\}^2. \quad (2.18)$$

Mean squared error: Mean squared error (MSE) is a combination of variance and bias of an estimator. By definition,

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \{Bias(\hat{\theta})\}^2 + Var(\hat{\theta}). \quad (2.19)$$

It is a measure of the overall quality of an estimator, $\hat{\theta}$. If the estimator $\hat{\theta}$ is unbiased, then $MSE(\hat{\theta}) = Var(\hat{\theta})$. But, if $\hat{\theta}$ is biased, then $MSE(\hat{\theta}) > Var(\hat{\theta})$. When $\hat{\theta}$ is

unbiased and has low variance, it will have small MSE. In that case, we say that the estimator is accurate and precise.

Coverage probability: The coverage probability of a technique is the proportion of the time that the confidence interval contains the true value of interest. The confidence interval contains the unknown value of interest with a given probability which is the nominal coverage probability for constructing confidence intervals.

In our study, we set the nominal coverage probability at 0.95 and calculate the 95% confidence interval of θ in each simulation for each sample size. Then, we calculate the proportion of the confidence intervals which contain the true value of θ across all R replications. That proportion is an estimate for the empirical coverage probability for the confidence interval. Let $\hat{\pi}$ be the estimated coverage probability. Then for $n \leq 30$, $\hat{\pi}$ can be defined by

$$\hat{\pi} = \frac{\#(\hat{\theta} - t_{0.975,18} \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + t_{0.975,18} \times SE(\hat{\theta}))}{R}, \quad (2.20)$$

where $t_{0.975,18}$ is the upper critical value for the t-distribution with degrees of freedom 18.

Under the assumption of asymptotic normality of the estimators, for $n > 30$, $\hat{\pi}$ can be defined by

$$\hat{\pi} = \frac{\#(\hat{\theta} - Z_{0.975} \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + Z_{0.975} \times SE(\hat{\theta}))}{R}, \quad (2.21)$$

where $Z_{0.975}$ is the upper critical value for the standard normal distribution. For this application, estimated coverage should be close to the nominal level (0.95).

Ratio of the estimated standard error and the empirical standard deviation: In our study, we also compare the estimated standard error (SE) with the empirical standard deviation (ESD) of $\hat{\theta}$. By calculating the average SE and the ESD of the estimates, the SE/ESD ratio is reported by

$$\frac{SE}{SD}(\hat{\theta}) = \frac{\overline{SE_{\hat{\theta}}}}{\sqrt{Var(\hat{\theta})}}. \quad (2.22)$$

This ratio represents how the estimated SE of $\hat{\theta}$ is underestimated or overestimated with respect to the empirical standard deviation of $\hat{\theta}$. A ratio of 1 indicates that the estimated SE and ESD of $\hat{\theta}$ is equal which is the most desirable condition.

Chapter 3

Simulation Study

3.1 Simulation Procedure

To generate data, we used simple linear models for a continuous response variable Y and a continuous predictor w where $w \sim Uniform(0, 2)$. For the errors ϵ , we used different normal and non-normal distributions. The details of the models used to generate the data are shown in table 3.1.

Scenario 1 and Scenario 4 represent the homoscedastic and heteroscedastic (HS) normal errors with no outliers respectively. On the other hand, Scenario 5 and Scenario 10 represent the homoscedastic and HS normal errors with outliers contamination (OC) separately. In Scenario 2, the errors are homoscedastic but heavy-tailed (HT) with $t_{(2)}$ distribution whereas in Scenario 7, the errors are HS and HT with $t_{(3)}$ distribution. The distribution of the error term in Scenario 3 is homoscedastic exponential, that is, skewed (SK). On the contrary, the error distribution is HS exponential in Scenario 8. The errors are homoscedastic and HS exponential with OC in Scenario 9 and Scenario 13 respectively. In Scenario 6, the errors are homoscedastic log-normal, that is, heavy-tailed-skewed. On the other hand, a model with HS log-normal errors was considered in Scenario 11. The homoscedastic and HS log-normal errors with OC are presented in Scenarios 12 and 14 respectively.

The mean and standard deviation of normal distribution was chosen as 0 and 1 respectively. The rate of the exponential distribution was kept 1. Again, the mean and standard deviation of log-normal distribution was chosen as 0 and 1 respectively.

However, these parameterizations were changed in scenarios where outliers were generated. Outliers contaminated errors follow a mixture of specific distributions; for example, in Scenario 5, normal distribution was contaminated in which the majority of observations are from a specified normal distribution $[N(0, 1)]$, but a small proportion are from a normal distribution with much higher variance $[N(0, 16^2)]$. We generated outliers with the help of a Bernoulli distribution where $B \sim \text{Bernoulli}(0.1)$ such that if the outcome of a Bernoulli trial is 1, then the variance of the error will be 256 instead of 1. That means, the outliers were generated in a way that the noise applied to some points (10%) follows the same distribution but with a higher variance than the noise that are applied to most other points (90%). The ratio of the variances obtained from the two error distributions (i.e., 256:1) was kept constant in each corresponding scenario (5, 9, 10, 12-14) where the errors were outlier contaminated (OC). But, in scenarios 10, 13, and 14, we kept the same ratio as 256:1 on average. The errors are heteroscedastic in those scenarios; therefore, $1 \leq \text{var}(y) \leq 9$ when $\epsilon \sim N(0, 1)$ and $w \in [0, 2]$. Thus, we took the average value from 1 to 9 which is 5.

In our study, we conducted all simulations and analyses using the statistical software package R. Samples (y_i, w_i) of size $n \in \{20, 100, 500\}$ were independently drawn according to each model presented in table 3.1 for 5000 replications. With the R functions **lm**, **gls**, **rlm**, and **rq**, we performed linear (OLS), generalized least squares (GLS), robust (M), and quantile regression (QR) respectively. Then, we obtained the OLS, GLS, M- and QR estimates of the true coefficients and their corresponding standard errors in each iteration for each sample size. To calculate the standard errors of QR estimates, we used **boot** with **rq** function. We also calculated and obtained the 95% confidence intervals to assess coverage of the coefficients under all scenarios when data were generated. Forty-two configurations were produced in total from the

Table 3.1: Summary of scenarios

No.	Scenario	Model	Error
1.	Normal	$Y = 2w + \epsilon$	$\epsilon \sim N(0, 1)$
2.	HT	$Y = 2w + \epsilon$	$\epsilon \sim t(2)$
3.	SK	$Y = 2w + \epsilon$	$\epsilon \sim Exp(1)$
4.	HS	$Y = 2w + (w + 1)\epsilon$	$\epsilon \sim N(0, 1)$
5.	OC	$Y = 2w + \epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim N(0, 1) \\ 1, & \text{if } \epsilon \sim N(0, 16^2). \end{cases}$
6.	HT-SK	$Y = 2w + \epsilon$	$\epsilon \sim LogN(0, 1)$
7.	HT-HS	$Y = 2w + (w + 1)\epsilon$	$\epsilon \sim t(3)$
8.	SK-HS	$Y = 2w + (w + 1)\epsilon$	$\epsilon \sim Exp(1)$
9.	SK-OC	$Y = 2w + \epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim Exp(1) \\ 1, & \text{if } \epsilon \sim Exp(\lambda). \end{cases}$ where $\lambda(= 0.0625)$ is such that $\frac{1}{\lambda^2} = 256$.
10.	HS-OC	$Y = 2w + (w + 1)\epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim N(0, 1) \\ 1, & \text{if } \epsilon \sim N(0, 35^2). \end{cases}$ where $\sigma = 35$ is such that $\sigma^2 = 5 \times 256 = 1280$.
11.	HT-SK-HS	$Y = 2w + (w + 1)\epsilon$	$\epsilon \sim LogN(0, 1)$
12.	HT-SK-OC	$Y = 2w + \epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim LogN(0, 1) \\ 1, & \text{if } \epsilon \sim LogN(0, \sigma). \end{cases}$ where $\sigma(= 1.886187)$ is such that $[exp(\sigma^2) - 1] \cdot exp(\sigma^2) = (4.670774)(256) = 1195.718$. $[exp(1) - 1] exp(1) = 4.670774$.
13.	SK-HS-OC	$Y = 2w + (w + 1)\epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim Exp(1) \\ 1, & \text{if } \epsilon \sim Exp(\lambda). \end{cases}$ where λ is such that $\frac{1}{\lambda^2} = 1280$.
14.	HT-SK-HS-OC	$Y = 2w + (w + 1)\epsilon$	$B = \begin{cases} 0, & \text{if } \epsilon \sim LogN(0, 1) \\ 1, & \text{if } \epsilon \sim LogN(0, \sigma). \end{cases}$ where $\sigma(= 2.787369)$ is such that $[exp(\sigma^2) - 1] \cdot exp(\sigma^2) = (21878.05)(256) = 5600781$. $[exp(5) - 1] exp(5) = 21878.05$.

fourteen scenarios for three sample sizes. Each configuration contains 5000 estimates of the true intercept, 5000 estimates of true slope, 5000 standard errors of estimated intercepts, 5000 standard errors of estimated slopes, 5000 95% confidence intervals of true intercept, and 5000 95% confidence intervals of true slope for four regression techniques (OLS, GLS, M, QR) respectively. Except for GLS, the other three estimators did not produce any missing value on any simulations. GLS occasionally failed to converge, mostly when the sample size was small, but no more than 2% in any scenarios.

3.2 Simulation Results

We assessed the performance of those four estimators in terms of bias, average distance, MSE, coverage probability, and ratio of estimated standard error (SE) and empirical standard deviation (ESD) described in chapter 2. These measures were used to evaluate the accuracy and stability of an estimator. If an estimator have estimated bias close to zero, minimum average distance and MSE, 0.95 coverage, and a ratio of estimated standard error and empirical standard deviation (SE/ESD) close to one, that is considered as a good one. Estimated bias, average distance, MSE, coverage probability, and SE/ESD for four different estimators (OLS, GLS, M, QR) of the intercept and slope are shown in Tables A.1–A.14. In addition, Table 3.2 gives a clear overview of the results from the simulation study. The most suitable scenario for each estimator is presented in this table.

In Scenario 1, overall the performance of OLS estimator was the best among those four estimators as the scenario was for normal homoscedastic errors with no outliers. We observed that the bias of all estimators for both intercept and slope was nearly zero in each sample size. Both the average distance and MSE of OLS estimator were the smallest among those four estimators irrespective of sample sizes. The coverage was good for OLS estimator across all sample sizes. For sample size 20, the coverage

Table 3.2: Summary of the results

Estimators	Suitable scenarios
OLS	homoscedastic normal, homoscedastic skewed.
GLS	heteroscedastic normal ($n = 500$), heteroscedastic heavy-tailed ($n = 500$). (comparative to QR)
M	homoscedastic heavy-tailed, heteroscedastic heavy-tailed ($n = 20$), homoscedastic normal with outliers, heteroscedastic normal with outliers ($n = 20$), homoscedastic skewed with outliers, heteroscedastic skewed with outliers ($n = 20$).
QR	heteroscedastic normal, heteroscedastic normal with outliers ($n = 100, 500$), heteroscedastic skewed, heteroscedastic skewed with outliers ($n = 100, 500$), heteroscedastic heavy-tailed ($n = 100, 500$), homoscedastic heavy-tailed-skewed, heteroscedastic heavy-tailed-skewed, homoscedastic heavy-tailed-skewed with outliers, heteroscedastic heavy-tailed-skewed with outliers.

of OLS estimator for both intercept (0.948) and slope (0.951) was nearer to 0.95 than the coverages of other estimators. But, for large sample sizes, the coverages of all estimators were very comparative and close to 0.95. Again, SE/ESD of OLS estimator for both intercept (0.971) and slope (0.967) was more close to 1 than the other three estimators when sample size was 20. When sample size was 100, SE/ESD of M-estimator for both intercept (0.986) and slope (0.988) was nearer to 1 than the other three estimators. However, the ratio of SE and ESD of OLS estimator for both intercept (0.982) and slope (0.986) was very comparative to M-estimator for sample size 100. For sample size 500, SE/ESD of all estimators (OLS:0.992, GLS:0.986, M:0.987, QR:1.014) was nearly 1.

In Scenario 2, all of the estimators were nearly zero biased for both intercept and slope across sample sizes. However, we noticed that the OLS estimator had the largest

average distance and MSE in each sample size for both intercept and slope. That means the OLS estimator had larger empirical variability than the other estimators. Also, when we looked at SE/ESD, we observed that the estimated standard error of OLS estimates was underestimated than the empirical standard deviation of the OLS estimates for all sample sizes. However, still we got a good coverage of near 0.95 for the OLS estimator as it was not biased. The same things happened with GLS estimator though the the performance of GLS estimator was somewhat better than the OLS. That implies, for a homoscedastic heavy-tailed distribution with no outliers, both OLS and GLS estimators might be accurate but not precise. However, both M- and QR estimators performed well in this scenario though the performance of the M-estimator was preeminent than the QR estimator. M-estimator had a smaller average distance and smaller MSE than QR. For large sample sizes, the coverage for both M- and QR estimators was very comparative. However, for a small sample size 20, the coverage of the QR estimator was higher for both intercept (0.979) and slope (0.982) than the nominal level whereas the coverage of the M- estimator was 0.945 for intercept and 0.944 for slope. Again, from SE/ESD we observed that the estimated SE of QR estimates was overestimated more for sample size 20 than 100 and 500. On the other hand, the SE/ESD of the M-estimator was nearer to 1 than the QR estimator for all sample sizes. Overall, for this scenario, the performance of the M-estimator was outstanding than the other three estimators in terms of accuracy and stability. Nevertheless, for sample size 500, the performance of both M- and QR estimators was nearly equal.

For a homoscedastic exponential model with no outliers defined in Scenario 3 with sample size 20, the OLS estimator performed best among those estimators as it had the least bias, coverage of 0.941 for intercept and 0.952 for slope which was the closest to 0.95, and a SE/ESD of 0.943 for intercept and 0.935 for slope which was most near to 1 than any other estimators. Again for sample size 100, the performance of OLS

was much satisfactory though the performance of both OLS and QR were relative. For sample size 500, all four estimators gave very comparative results for both intercept and slope except for M-intercept. We observed that M-estimator for intercept had slightly larger bias across all sample sizes which had a great impact on coverage probability. The coverage of the M-intercept was very poor for all sample sizes. In fact, the coverage was getting worse with the increase in sample size as the bias was getting larger for a large sample size. However, a good ratio of SE and ESD and small MSE revealed that the M-estimator had less variability.

In Scenario 4 dealing with a heteroscedastic normal model with no outliers, QR had a smaller bias than OLS, GLS, and M-estimator when the sample size was 20. When the sample size was 100 and 500, all of the estimators were nearly zero biased. GLS had the least average distance and smallest MSE across sample sizes. But, the coverage of GLS was not very good for a small sample size because of being biased. Again, the estimated standard error of GLS was greatly underestimated relative to the empirical standard deviation when the sample size was 20. Overall, for all sample sizes, QR gave more reasonable results though both GLS and QR were comparative for sample size 500.

When looking at Scenario 5, we observed that M-estimator performed best for both intercept and slope of a homoscedastic normal model contaminated with outliers with sample size 20 as it had nearly zero bias, small average distance, and MSE, better coverage, and SE/ESD than any other estimators. In fact, the performance of M-estimator was overwhelmingly better in terms of all criteria across sample sizes. Though for large sample sizes 100 and 500 all of the estimators gave comparative results, OLS and GLS had larger average distance and MSE than M- and QR estimators.

In Scenario 6, QR gave more reasonable results for a homoscedastic log-normal model with no outliers. Overall, QR and OLS had smaller bias than M- and GLS

estimators. In sample sizes 100 and 500 for slope, both M- and QR estimators had exactly zero bias. But, when looking at the intercept, we noticed that M-estimator was highly biased across sample sizes and also had very poor coverages even the opposite results were observed for the slope coefficient. On the other hand, the OLS estimator maintained good coverage for both the intercept and slope of the model. The coverage of QR somewhat larger than 0.95 for sample size 20, but it also maintained good coverage across sample sizes. Again, in terms of SE/ESD, OLS gave a better performance than QR for 20 though for 100 and 500 their performance was relative. GLS did not perform well in this scenario in terms of any criteria.

In Scenario 7, the performance of the M-estimator was the best for both intercept and slope of a heteroscedastic t_3 model with no outliers for sample size 20. For sample sizes 100 and 500, GLS, M- and QR estimators showed very comparative performances. OLS estimator did not perform well in terms of both accuracy and precision.

In Scenario 8 dealing with a heteroscedastic exponential model with no outliers, QR was seemed more reasonable than the other estimators. Though OLS had a small bias, it had a large average distance and MSE. Also, the coverage and SE/ESD of the OLS intercept were not very good. Overall, the performance of the GLS was also not satisfactory. Though M-estimator gave comparative results for the intercept, it was highly biased and had poor coverage for the slope for all sample sizes. On the other hand, QR produced good coverage and SE/ESD for sample sizes 100, 500, and the results for sample size 20 were also relative to the M-estimator.

In Scenario 9, M-estimator produced the most suitable results for a homoscedastic exponential model with outliers. QR also gave comparable results with M-estimator. In contrast, OLS and GLS estimators had a large bias, average distance, and MSE than M- and QR estimators. GLS estimator had very poor coverage and SE/ESD than the other estimators across sample sizes. Interestingly, OLS intercept had good

coverage in sample size 20 and very bad coverage in 500. If we looked at the bias, we could see that the bias of the OLS intercept was increasing with sample size. But, in terms of other criteria except for bias, the results of OLS was improved with the increase in sample size.

In Scenario 10, M- and QR estimators had a smaller bias, average distance, and MSE than the OLS and GLS estimators for both intercept and slope of a heteroscedastic normal model having outliers across sample sizes. For sample size 20, M-estimator gave better results than QR and for 100 and 500, the performance of QR was more preminent in terms of coverage and SE/ESD.

In Scenario 11, QR performed very well in terms of all criteria for a heteroscedastic log-normal model with no outliers. The average distance and MSE of OLS were comparatively large. GLS showed very poor coverage and SE/ESD across sample sizes. M-estimator was highly biased for all sample sizes.

For the homoscedastic log-normal model contaminated with outliers defined in Scenario 12, overall both QR and M-estimators gave more reasonable results than the other estimators in each sample size. The performance of OLS was very bad in terms of bias, average distance, MSE and SE/ESD. Also, GLS estimator behaved poorly in this scenario.

In Scenario 13, for sample size 20, the performance of the M-estimator was better than the other estimators for the heteroscedastic exponential model contaminated with outliers. For sample sizes 100 and 500, both QR and M-estimators performed better than OLS and GLS estimators. Both OLS and GLS estimators were highly biased and had larger MSE.

When looking at the heteroscedastic log-normal model contaminated with outliers in Scenario 14, we observed that for sample size 20, M-estimator showed the best performance among those four estimators for the intercept of the model. On the other hand, QR performed best for the slope of the model in each sample size and

intercept of the model for sample sizes 100 and 500. Both OLS and GLS estimators performed poorly in this scenario.

Chapter 4

Data Analysis

In this chapter, we performed a real data analysis to compare the median regression estimator with mean regression estimators considered in our study. We used 1,2,3,4-Tetrachlorobenzene (TcCB) concentrations (ppb) data which is available under R package **EnvStats** (named “EPA.94b.tccb.df”). Measurements of TcCB concentrations (ppb) in soil samples was taken at a site-specific reference area and a cleanup area (contaminated site) [11]. Among 124 observations, there are 47 observations for the reference area and 77 for the cleanup area. As one observation for cleanup area was not detected in the data set, we restricted our analysis by discarding that censored observation. Therefore, we used 123 measurements of TcCB concentrations (ppb) in total. The data set has 4 variables: one character variable with the original tetrachlorobenzene concentrations (ppb), one numeric variable of tetrachlorobenzene with < 0.99 coded as 0.99, one factor indicating the area (cleanup vs. reference) and one censoring indicator defining which observations are not detected. We assumed that the data were representative of the two areas like [11] although the samples were not located on a triangular grid.

4.1 Exploratory Analysis

In our study, we considered continuous “TcCB” concentrations (ppb) as a response variable and “Area” as a predictor. The density of TcCB concentrations (ppb) overlapped with a normal curve (mean = 2.68, standard deviation = 15.88) is shown at the left side of Figure 4.1. It is obvious from the plot that the distribution of the

TcCB (ppb) concentrations is not normally distributed. The shape of the distribution of TcCB concentrations (ppb) is asymmetric with a long tail to the right. It also appears that there are few unusually large observations at the right tail of TcCB distribution.

Table 4.1 shows the descriptive statistics of TcCB for cleanup and reference areas in the second and third columns separately. By looking at the table, it was observed that there was higher variability of TcCB measurements in the cleanup area than the reference area. The standard deviation of TcCb in the cleanup area is 20.14 (ppb) whereas, in the reference area it is only 0.28 (ppb). The high positive value of skewness and Kurtosis of TcCB in the cleanup area suggests that the non-normality of TcCB shown in Figure 4.1 mostly comes from the cleanup area. Moreover, the maximum value of TcCB is 168.6 (ppb) in the cleanup area. On the other hand, the maximum value of TcCB is 1.33 (ppb) in the reference area which is interestingly nearer to the third quartile (1.12) of TcCB concentrations (ppb) in the cleanup area. Again, the minimum value of TcCB distribution is 0.09 belongs to the cleanup area which also indicates the high range of variability among measurements in the cleanup area.

Although the means of TcCB differ largely by those areas, the medians are closer. This is because the median is robust to extreme observations whereas, mean is very susceptible to extreme values. In many real application areas dealing with non-normality, the robustness property of median is very useful to conduct statistical analysis. We also displayed the distribution of log-transformation of TcCB overlapped with a normal curve (mean = -0.56 , standard deviation = 1.10) in the right-side plot of Figure 4.1. We observed that log-transformed TcCB tended to be normal. However, in our regression analysis, we compared the models without doing log-transformation of TcCB to be consistent with the simulation study.



Figure 4.1: Distribution of TcCB concentrations (ppb) (before and after log-transformation).

Table 4.1: Descriptive statistics of TcCB (ppb) concentrations by Area.

Statistic	Cleanup	Reference
N	76	47
Mean	3.97	0.60
Median	0.43	0.54
Standard deviation	20.14	0.28
Minimum	0.09	0.22
Maximum	168.6	1.33
25 th Quantile	0.24	0.39
75 th Quantile	1.12	0.75
Skewness	7.67	0.90
Kurtosis	61.86	0.13

4.2 Regression Analysis

We performed both mean and median regressions on TcCB data as a function of “Area”. We considered reference area as baseline category. Four different estimators (OLS, GLS, M, QR) used for simulation study in Chapter 3 were applied on TcCB data to estimate the intercept and slope of Area. As we observed that the distribution of TcCB carried few larger observations on the right tail, the influence of those points on regression results might be a great concern. Of course, it was not possible to discover the influence of those extreme observations until fitting a model. Therefore, we fit the model of TcCB verses area using those four estimators mentioned above.

Table 4.2: Estimates of the intercept and slope along with their estimated standard errors (given in parenthesis) for 4 different estimators (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with TcCB concentrations (ppb) data (TcCB \sim Area (ref: “Reference”); significant estimates are marked in bold) .

Estimator	With Outliers		Without Outliers	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	0.5985 (2.3134)	3.3670 (2.9431)	0.5985 (0.2684)	0.4930 (0.3432)
GLS	0.5985 (2.3134)	3.3670 (2.9431)	0.5985 (0.2684)	0.4930 (0.3432)
M	0.5963 (0.0650)	0.0084 (0.0827)	0.5950 (0.0616)	-0.0241 (0.0788)
QR	0.5400 (0.0422)	-0.1100 (0.0789)	0.5400 (0.0412)	-0.1201 (0.0793)

Again, the whole analysis was conducted using R software. For the QR estimator, we fit the model at the 50th quantile (median regression) like the simulation study. The standard error of QR estimator was estimated by bootstrap resampling technique using `boot` with `rq` function.

The second and third columns of Table 4.2 represent the estimates of intercepts and slopes as well as their estimated standard errors allowing all observations under each estimation technique respectively. Two interesting points came out when we observed the estimates of both intercept and slope coefficients for all estimation methods. First, for both M-and QR estimators, the intercept was significant whereas both OLS and GLS estimators, the intercept was not significant. Second, The estimated values of the slope coefficients of OLS and GLS estimators were completely different from the QR estimator not only by magnitude but also by direction. Although the magnitudes of QR and M estimates were somewhat similar given that both were close to zero and neither was significant, their directions were different. As there were no heteroscedasticity and auto-correlation issues in this data, the OLS and GLS gave

Table 4.3: Summary statistics of TcCB (ppb) by Area after deletion of outliers from the data.

Statistic	Cleanup	Reference
N	74	47
Mean	1.0910	0.5985
Median	0.4150	0.5400
Standard deviation	2.3387	0.2836
Minimum	0.0900	0.2200
Maximum	18.400	1.3300

the same estimates. The reason for which we got the estimates of slope for OLS, GLS and M estimators with the same directions is that those estimators are all involved to mean regression. In contrast, QR is involved in median regression which is robust to extreme observations. Since M-estimator is also robust, it gave estimates of the slope coefficient closer to QR.

To detect the influence of extreme measurements of TcCB, we then performed diagnostic checking for the OLS model. We calculated DFBETAS for the slope (Area) coefficient and plotted them. DFBETAS is a useful tool that measures the change in regression coefficients when one observation is deleted [19]. A rule of thumb for choosing a cutoff is $2/\sqrt{n}$ which is size-adjusted. If the value of DFBETAS is greater than the cutoff, the corresponding observation is considered as an influential observation. A plot of DFBETAS with a cutoff 0.18 for Area is attached in Figure B.1. Table B.1 shows the DFBETAS of Area with corresponding influential observations of TcCB concentrations (ppb). From both Figure B.1 and Table B.1, it appeared that the last two observations of TcCB (51.97, 168.64) in the cleanup area were influential.

After that, we fit the model again using those four estimators by discarding the influential observations from the data. The results of new regression estimates are also included in Table 4.2. This time the intercept was significant for all four estimators. Although the magnitudes of OLS and GLS slope estimates changed by 85.34%, the directions remained the same as positive. Interestingly, the direction of the slope

coefficient for M-estimator was changed from positive to negative. Still, the direction of the slope coefficient for the QR estimator remained the same as negative. It is obvious from the study that the QR estimator is a reasonable good alternative if the data is affected by the influential observations or outliers.

Table 4.3 summarizes the basic statistics of TcCB by Area after the deletion of outliers from the data. The mean and median of TcCB in the cleanup area changed to 3.97 from 1.09 and 0.43 from 0.42 respectively. The variability of TcCB in the cleanup area was also decreased.

Chapter 5

Discussion and Conclusions

This study illustrates a simulation study in which we compared four different estimators (OLS, GLS, M, QR) for a continuous response variable under different scenarios. We also made a contrast among those estimators using TcCB concentrations (ppb) data. Our results from the simulation study reveal interesting behaviors of those estimators. When the errors are homoscedastic normal having no outliers, OLS performs the best among those four estimators. In contrast, when the errors are heteroscedastic normal having no outliers, QR performs the best among those four estimators. Again, if the homoscedastic normal errors are affected by outliers, then the performance of M-estimator is the best. But, both M- and QR estimators give much better performances than OLS and GLS estimators when the errors are heteroscedastic normal contaminated with outliers. When the error distribution is heavy-tailed but symmetric (t-distribution), with small n , M-estimator gives the most satisfactory performances and with large n , both QR and M-estimators give better comparative results. In addition, the M-estimator is the best in scenarios with heteroscedastic heavy-tailed distributed errors for small n and for large n , the performance of the QR estimator is more reasonable than the other estimators.

Again, the best performance of the OLS estimator is found for a model with homoscedastic skewed (exponentially distributed) errors. But, if the homoscedastic skewed errors affected by the outliers, then the M-estimator performed the best. On the other hand, when the errors are heteroscedastic skewed whether or not contaminated with outliers, the performance of the QR estimator is preeminent. The most satisfactory results of QR are found when the errors are log-normal irrespective of

homoscedasticity or having outliers. Also, M-estimator gives very comparable results to QR when the errors are log-normally distributed.

Our results give a clear overview of those four estimators (OLS, GLS, M- and QR) to the researchers for analyzing different normal and non-normal continuous response data. Clearly, classical OLS estimator would not be a proper choice when the data is heteroscedastic or heavy-tailed or heavy-tailed-skewed, and/or contaminated with outliers. Because of having robustness, both M and QR estimators excel in all scenarios affiliated with outliers. The flexibility of QR is revealed when the errors are heteroscedastic. Also from the data analysis, the promising behavior of the QR estimator is revealed when dealing with a skewed-heavy-tailed data contaminated with outliers. As the mean is not robust to extreme observations, OLS can not behave properly when dealing with outliers. However, it should be suggested that outliers should not be removed without knowing their actual impact on regression results. Rather, it can be recommended that another suitable choice would be the QR estimator to deal with skewed-heavy-tailed outliers contaminated data.

Bibliography

- [1] J Abrevaya. The effects of demographics and maternal behavior on the distribution of birth outcomes. In *Economic Applications of Quantile Regression*, pages 247–257. Springer, 2002.
- [2] PC Austin, JV Tu, PA Daly, and DA Alter. The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in Medicine*, 24(5):791–816, 2005.
- [3] Per Bergström and Mats Lindegarth. Environmental influence on mussel (*mytilus edulis*) growth—a quantile regression approach. *Estuarine, Coastal and Shelf Science*, 171:123–132, 2016.
- [4] M Buchinsky. Changes in the us wage structure 1963-1987: Application of quantile regression. *Econometrica: Journal of the Econometric Society*, pages 405–458, 1994.
- [5] Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.
- [6] G Chamberlain. Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress*, volume 2, pages 171–209, 1994.
- [7] JM Chambers and TJ Hastie. Linear models. chapter 4 of statistical models in s. *Wadsworth & Brooks/Cole*, 1992.
- [8] In Choi. Econometrics: by fumio hayashi, princeton university press, 2000. *Econometric Theory*, 18(4):1000–1006, 2002.
- [9] Norman Richard Draper, Harry Smith, and Elizabeth Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.
- [10] ER Eide, MH Showalter, and DP Sims. The effects of secondary school quality on the distribution of earnings. *Contemporary Economic Policy*, 20(2):160–170, 2002.
- [11] RO Gilbert and JC Simpson. Statistical methods for evaluating the attainment of cleanup standards. volume 3, reference-based standards for soils and solid media, revision 1. Technical report, Pacific Northwest Lab., Richland, WA (United States), 1992.

- [12] Lucy Griffiths, Marco Geraci, Mario Cortina-Borja, Francesco Sera, Catherine Law, Heather Joshi, Andrew Ness, and Carol Dezateux. Associations between children’s behavioural and emotional development and objectively measured physical activity and sedentary time: Findings from the uk millennium cohort study. *Longitudinal and Life Course Studies*, 7(2):124–143, 2016.
- [13] MS Handcock and M Morris. *Relative distribution methods in the social sciences*. Springer Science & Business Media, 2006.
- [14] L Hao. Sources of wealth inequality: Analyzing conditional distribution. In *Invited seminar at The Center for Advanced Social Science Research*, 2006.
- [15] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.
- [16] R Koenker and G Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.
- [17] Peter McCullagh. *Generalized linear models*. Routledge, 2018.
- [18] Samantha M McDonald, Andrew Ortaglia, Christina Supino, and Matteo Bottai. The role of energy intake on fitness-adjusted racial/ethnic differences in central adiposity using quantile regression. *Journal of Racial and Ethnic Health Disparities*, 6(2):292–300, 2019.
- [19] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.
- [20] FS Scharf, F Juanes, and M Sutherland. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. *Ecology*, 79(2):448–460, 1998.

Appendix A

Simulation Study Results

In this Appendix, we display the results deduced from our simulation study. The results of each scenario is presented through Tables A.1-A.14. The illustration of the results can be found in Chapter 3.

Table A.1: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 1 (normal, homoscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.009	-0.002	0.002	OLS	0.012	0.000	-0.001
GLS	-0.004	-0.002	0.002	GLS	0.006	0.000	-0.001
M	-0.006	-0.002	0.001	M	0.010	0.001	-0.001
QR	0.001	-0.001	0.001	QR	0.006	0.001	0.000
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.372	0.163	0.072	OLS	0.328	0.142	0.063
GLS	0.413	0.167	0.072	GLS	0.368	0.144	0.063
M	0.382	0.166	0.074	M	0.336	0.144	0.065
QR	0.464	0.203	0.091	QR	0.408	0.174	0.079
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.221	0.042	0.008	OLS	0.170	0.031	0.006
GLS	0.278	0.044	0.008	GLS	0.221	0.033	0.006
M	0.233	0.044	0.009	M	0.179	0.033	0.006
QR	0.341	0.066	0.013	QR	0.261	0.048	0.010
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.948	0.942	0.949	OLS	0.951	0.940	0.950
GLS	0.855	0.925	0.947	GLS	0.902	0.934	0.948
M	0.938	0.944	0.951	M	0.947	0.943	0.950
QR	0.962	0.935	0.950	QR	0.968	0.946	0.947
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.971	0.982	0.993	OLS	0.967	0.986	0.992
GLS	0.750	0.937	0.982	GLS	0.796	0.957	0.986
M	0.967	0.986	0.986	M	0.964	0.988	0.987
QR	1.125	1.043	1.004	QR	1.142	1.063	1.014

Table A.2: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 2 (t_2 (heavy-tailed), homoscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.001	-0.008	0.003	OLS	0.018	0.008	-0.004
GLS	0.012	0.001	0.002	GLS	-0.009	0.002	-0.002
M	-0.003	-0.004	0.001	M	0.006	0.004	-0.001
QR	0.002	-0.005	0.000	QR	0.001	0.004	0.000
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.922	0.444	0.217	OLS	0.827	0.387	0.189
GLS	0.773	0.379	0.195	GLS	0.742	0.340	0.171
M	0.552	0.226	0.100	M	0.486	0.197	0.087
QR	0.570	0.232	0.100	QR	0.504	0.202	0.087
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	3.133	0.469	0.100	OLS	3.468	0.346	0.078
GLS	1.244	0.275	0.071	GLS	1.324	0.221	0.052
M	0.517	0.081	0.016	M	0.403	0.061	0.012
QR	0.555	0.085	0.016	QR	0.440	0.064	0.012
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.953	0.952	0.951	OLS	0.956	0.950	0.953
GLS	0.836	0.934	0.951	GLS	0.906	0.946	0.955
M	0.945	0.943	0.948	M	0.944	0.944	0.948
QR	0.979	0.947	0.946	QR	0.982	0.955	0.948
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.630	0.798	0.859	OLS	0.525	0.805	0.842
GLS	0.667	0.844	0.884	GLS	0.664	0.883	0.945
M	0.932	0.996	1.003	M	0.926	0.992	0.998
QR	1.267	1.062	1.037	QR	1.303	1.079	1.043

Table A.3: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 3 (exponential, homoscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.002	0.003	-0.001	OLS	0.001	-0.002	0.000
GLS	-0.054	-0.018	-0.006	GLS	0.037	0.012	0.004
M	-0.133	-0.136	-0.138	M	0.005	0.000	0.000
QR	0.050	0.010	0.002	QR	0.001	0.000	-0.001
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.369	0.160	0.072	OLS	0.321	0.139	0.063
GLS	0.394	0.167	0.073	GLS	0.348	0.142	0.063
M	0.338	0.177	0.141	M	0.259	0.106	0.048
QR	0.366	0.162	0.072	QR	0.323	0.139	0.064
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.221	0.041	0.008	OLS	0.171	0.031	0.006
GLS	0.246	0.044	0.009	GLS	0.201	0.033	0.006
M	0.176	0.045	0.025	M	0.114	0.018	0.004
QR	0.222	0.041	0.008	QR	0.170	0.030	0.006
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.941	0.948	0.948	OLS	0.952	0.953	0.956
GLS	0.777	0.898	0.938	GLS	0.877	0.928	0.950
M	0.888	0.805	0.491	M	0.946	0.954	0.954
QR	0.962	0.943	0.944	QR	0.979	0.954	0.949
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.943	0.990	0.982	OLS	0.935	0.988	0.987
GLS	0.717	0.906	0.950	GLS	0.776	0.938	0.971
M	0.901	0.942	0.911	M	0.929	1.005	0.977
QR	1.151	1.043	1.005	QR	1.165	1.058	1.003

Table A.4: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 4 (normal, heteroscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.017	-0.001	0.002	OLS	0.026	-0.002	-0.001
GLS	-0.011	-0.004	0.002	GLS	0.012	0.001	-0.001
M	-0.014	-0.001	0.002	M	0.022	-0.001	0.000
QR	-0.003	0.000	0.001	QR	0.015	-0.001	0.002
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.606	0.259	0.114	OLS	0.705	0.299	0.134
GLS	0.570	0.233	0.100	GLS	0.687	0.272	0.119
M	0.587	0.250	0.111	M	0.697	0.299	0.134
QR	0.680	0.287	0.128	QR	0.798	0.337	0.152
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.595	0.106	0.020	OLS	0.780	0.141	0.028
GLS	0.544	0.086	0.016	GLS	0.771	0.117	0.022
M	0.559	0.099	0.019	M	0.767	0.140	0.028
QR	0.762	0.133	0.026	QR	1.006	0.182	0.036
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.984	0.987	0.990	OLS	0.943	0.929	0.943
GLS	0.778	0.891	0.943	GLS	0.884	0.916	0.949
M	0.981	0.987	0.990	M	0.932	0.921	0.933
QR	0.978	0.944	0.948	QR	0.970	0.946	0.948
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.230	1.284	1.306	OLS	0.931	0.964	0.966
GLS	0.675	0.875	0.963	GLS	0.772	0.921	0.977
M	1.236	1.288	1.299	M	0.915	0.937	0.935
QR	1.205	1.060	1.008	QR	1.156	1.062	1.014

Table A.5: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 5 (normal, homoscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.034	0.010	-0.006	OLS	0.022	-0.007	0.003
GLS	-0.011	0.003	-0.004	GLS	-0.011	-0.001	0.001
M	-0.004	-0.002	0.000	M	0.006	0.000	-0.001
QR	-0.002	-0.003	0.002	QR	0.001	0.000	-0.002
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.646	0.800	0.369	OLS	1.460	0.706	0.324
GLS	1.194	0.683	0.347	GLS	1.208	0.630	0.310
M	0.477	0.196	0.087	M	0.424	0.170	0.075
QR	0.516	0.225	0.100	QR	0.458	0.197	0.086
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	5.873	1.085	0.216	OLS	4.433	0.833	0.165
GLS	3.970	0.846	0.196	GLS	3.524	0.673	0.153
M	0.413	0.061	0.012	M	0.330	0.045	0.009
QR	0.441	0.081	0.016	QR	0.349	0.061	0.012
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.964	0.950	0.948	OLS	0.962	0.956	0.945
GLS	0.821	0.922	0.948	GLS	0.908	0.949	0.946
M	0.949	0.949	0.951	M	0.946	0.950	0.950
QR	0.976	0.952	0.947	QR	0.980	0.957	0.954
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.837	0.959	0.985	OLS	0.841	0.949	0.976
GLS	0.575	0.865	0.972	GLS	0.662	0.942	0.989
M	0.919	0.996	0.989	M	0.898	1.000	0.995
QR	1.475	1.041	1.011	QR	1.512	1.053	1.016

Table A.6: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 6 (log-normal, homoscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.013	0.006	0.004	OLS	0.020	-0.007	-0.002
GLS	-0.143	-0.061	-0.017	GLS	0.080	0.035	0.012
M	-0.359	-0.388	-0.388	M	0.006	0.000	0.000
QR	0.083	0.016	0.004	QR	0.004	0.000	0.000
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.733	0.340	0.154	OLS	0.630	0.288	0.134
GLS	0.735	0.350	0.157	GLS	0.643	0.295	0.135
M	0.581	0.404	0.388	M	0.368	0.147	0.066
QR	0.480	0.203	0.091	QR	0.424	0.175	0.079
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.016	0.196	0.038	OLS	0.806	0.141	0.029
GLS	0.914	0.199	0.040	GLS	0.785	0.147	0.030
M	0.492	0.206	0.162	M	0.242	0.035	0.007
QR	0.414	0.068	0.013	QR	0.303	0.049	0.010
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.922	0.935	0.945	OLS	0.953	0.949	0.950
GLS	0.723	0.837	0.898	GLS	0.880	0.906	0.929
M	0.799	0.515	0.038	M	0.948	0.944	0.948
QR	0.974	0.936	0.949	QR	0.983	0.949	0.948
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.857	0.935	0.976	OLS	0.838	0.956	0.974
GLS	0.625	0.809	0.909	GLS	0.690	0.867	0.939
M	0.840	0.897	0.893	M	0.898	0.986	0.985
QR	1.240	1.053	1.006	QR	1.284	1.074	1.015

Table A.7: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 7 (t_3 (heavy-tailed), heteroscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.016	-0.012	0.005	OLS	0.015	0.011	-0.005
GLS	0.006	-0.010	0.001	GLS	-0.016	0.008	0.000
M	-0.009	-0.008	0.004	M	0.006	0.010	-0.003
QR	0.006	-0.007	0.004	QR	-0.011	0.010	-0.003
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.966	0.436	0.193	OLS	1.090	0.504	0.224
GLS	0.833	0.356	0.163	GLS	1.029	0.425	0.195
M	0.754	0.317	0.137	M	0.849	0.359	0.158
QR	0.771	0.313	0.138	QR	0.896	0.369	0.162
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.715	0.345	0.059	OLS	2.226	0.488	0.082
GLS	1.345	0.214	0.042	GLS	1.889	0.259	0.060
M	0.953	0.160	0.030	M	1.197	0.204	0.039
QR	1.032	0.156	0.030	QR	1.348	0.219	0.041
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.983	0.985	0.989	OLS	0.945	0.942	0.947
GLS	0.795	0.908	0.941	GLS	0.893	0.929	0.946
M	0.978	0.982	0.985	M	0.930	0.938	0.939
QR	0.986	0.949	0.944	QR	0.983	0.952	0.949
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.163	1.187	1.299	OLS	0.885	0.865	0.959
GLS	0.645	0.873	0.963	GLS	0.750	0.937	0.977
M	1.188	1.243	1.271	M	0.919	0.951	0.961
QR	1.309	1.085	1.029	QR	1.252	1.073	1.040

Table A.8: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 8 (exponential, heteroscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.005	0.003	-0.001	OLS	0.002	-0.001	-0.001
GLS	0.001	-0.028	-0.013	GLS	-0.066	0.004	0.004
M	-0.031	-0.006	-0.006	M	-0.253	-0.315	-0.328
QR	0.079	0.017	0.004	QR	0.020	0.002	-0.001
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.592	0.256	0.114	OLS	0.685	0.297	0.134
GLS	0.556	0.234	0.101	GLS	0.666	0.278	0.119
M	0.487	0.204	0.093	M	0.610	0.357	0.329
QR	0.530	0.228	0.103	QR	0.630	0.270	0.123
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.583	0.105	0.021	OLS	0.777	0.141	0.028
GLS	0.513	0.087	0.016	GLS	0.725	0.122	0.022
M	0.396	0.065	0.014	M	0.570	0.177	0.123
QR	0.488	0.083	0.017	QR	0.658	0.115	0.024
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.985	0.985	0.988	OLS	0.934	0.931	0.946
GLS	0.717	0.860	0.927	GLS	0.862	0.903	0.944
M	0.971	0.981	0.984	M	0.871	0.730	0.204
QR	0.980	0.949	0.944	QR	0.975	0.947	0.947
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.199	1.282	1.290	OLS	0.901	0.955	0.963
GLS	0.642	0.835	0.930	GLS	0.740	0.875	0.956
M	1.135	1.190	1.147	M	0.870	0.944	0.934
QR	1.253	1.060	1.006	QR	1.187	1.057	1.002

Table A.9: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 9 (exponential, homoscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.475	1.481	1.507	OLS	0.026	0.006	-0.012
GLS	0.801	1.173	1.385	GLS	0.217	0.218	0.082
M	0.111	0.060	0.058	M	0.005	0.000	-0.001
QR	0.183	0.119	0.109	QR	0.004	0.000	0.000
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	2.153	1.588	1.507	OLS	1.774	0.902	0.416
GLS	1.302	1.289	1.388	GLS	1.311	0.933	0.455
M	0.450	0.189	0.093	M	0.369	0.144	0.063
QR	0.456	0.218	0.125	QR	0.401	0.174	0.076
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	12.341	4.056	2.644	OLS	7.785	1.434	0.276
GLS	5.892	3.106	2.347	GLS	4.140	1.461	0.334
M	0.469	0.059	0.014	M	0.301	0.033	0.006
QR	0.418	0.079	0.024	QR	0.284	0.048	0.009
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.941	0.813	0.284	OLS	0.964	0.956	0.954
GLS	0.782	0.813	0.244	GLS	0.876	0.848	0.886
M	0.925	0.916	0.885	M	0.947	0.956	0.950
QR	0.980	0.941	0.846	QR	0.987	0.954	0.948
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.757	0.939	0.983	OLS	0.753	0.929	0.992
GLS	0.500	0.698	0.831	GLS	0.628	0.785	0.869
M	0.763	0.873	0.897	M	0.821	0.977	1.002
QR	1.617	1.029	1.018	QR	1.739	1.045	1.018

Table A.10: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 10 (normal, heteroscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.096	0.013	-0.020	OLS	0.065	-0.004	0.006
GLS	-0.113	0.023	-0.010	GLS	0.038	-0.016	-0.005
M	-0.010	-0.003	0.001	M	0.018	-0.002	-0.002
QR	-0.006	-0.002	0.003	QR	0.004	-0.004	-0.003
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	5.835	2.800	1.274	OLS	6.306	3.243	1.502
GLS	3.253	1.752	0.927	GLS	4.021	2.330	1.212
M	0.824	0.316	0.140	M	0.916	0.355	0.158
QR	0.770	0.323	0.141	QR	0.913	0.382	0.168
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	75.487	13.165	2.553	OLS	94.644	18.088	3.581
GLS	41.618	6.530	1.472	GLS	46.170	9.348	2.352
M	1.703	0.159	0.031	M	2.173	0.199	0.040
QR	1.089	0.167	0.032	QR	1.405	0.229	0.045
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.982	0.981	0.988	OLS	0.968	0.953	0.939
GLS	0.791	0.887	0.940	GLS	0.876	0.948	0.946
M	0.980	0.984	0.987	M	0.937	0.943	0.936
QR	0.989	0.957	0.950	QR	0.984	0.960	0.950
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.001	1.224	1.285	OLS	0.775	0.905	0.940
GLS	0.485	0.795	0.940	GLS	0.590	0.941	0.963
M	0.927	1.230	1.225	M	0.712	0.952	0.944
QR	2.135	1.059	1.017	QR	2.055	1.060	1.018

Table A.11: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 11 (log-normal, heteroscedastic, no outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	-0.026	0.013	0.006	OLS	0.046	-0.019	-0.002
GLS	0.017	-0.061	-0.033	GLS	-0.225	-0.019	0.009
M	-0.177	-0.180	-0.179	M	-0.566	-0.661	-0.675
QR	0.130	0.028	0.006	QR	0.041	0.002	0.003
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.184	0.538	0.246	OLS	1.357	0.612	0.286
GLS	1.015	0.483	0.222	GLS	1.216	0.554	0.260
M	0.755	0.334	0.204	M	0.947	0.681	0.675
QR	0.709	0.287	0.129	QR	0.830	0.338	0.152
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	2.869	0.491	0.097	OLS	3.944	0.628	0.132
GLS	2.117	0.394	0.079	GLS	2.700	0.508	0.107
M	0.965	0.171	0.059	M	1.358	0.582	0.483
QR	0.951	0.137	0.026	QR	1.174	0.185	0.037
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.979	0.984	0.987	OLS	0.917	0.929	0.939
GLS	0.691	0.803	0.871	GLS	0.853	0.878	0.918
M	0.957	0.940	0.838	M	0.823	0.507	0.020
QR	0.986	0.944	0.947	QR	0.978	0.947	0.948
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.054	1.215	1.272	OLS	0.778	0.930	0.945
GLS	0.561	0.749	0.857	GLS	0.678	0.834	0.898
M	1.039	1.114	1.012	M	0.854	0.944	0.949
QR	1.344	1.073	1.009	QR	1.307	1.076	1.015

Table A.12: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 12 (log-normal, homoscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.451	0.421	0.416	OLS	-0.059	0.002	0.005
GLS	-0.018	0.151	0.278	GLS	0.131	0.103	0.062
M	-0.324	-0.359	-0.361	M	0.003	-0.001	-0.001
QR	0.084	0.015	0.005	QR	0.006	0.000	-0.002
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	1.263	0.812	0.531	OLS	1.097	0.704	0.391
GLS	0.882	0.532	0.389	GLS	0.859	0.520	0.328
M	0.592	0.385	0.362	M	0.397	0.157	0.069
QR	0.500	0.213	0.095	QR	0.447	0.186	0.082
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	19.221	3.486	0.754	OLS	14.767	2.542	0.487
GLS	2.016	0.762	0.333	GLS	2.074	0.730	0.232
M	0.524	0.194	0.143	M	0.295	0.039	0.008
QR	0.454	0.075	0.014	QR	0.353	0.055	0.011
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.935	0.944	0.892	OLS	0.961	0.958	0.956
GLS	0.721	0.854	0.898	GLS	0.866	0.897	0.908
M	0.811	0.856	0.090	M	0.949	0.952	0.951
QR	0.974	0.950	0.947	QR	0.981	0.962	0.954
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.350	0.523	0.724	OLS	0.347	0.518	0.685
GLS	0.521	0.622	0.709	GLS	0.564	0.651	0.752
M	0.844	0.888	0.889	M	0.879	0.987	1.000
QR	1.335	1.054	1.013	QR	1.345	1.068	1.019

Table A.13: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 13 (exponential, heteroscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	3.327	3.445	3.519	OLS	3.645	3.456	3.406
GLS	2.937	2.095	2.985	GLS	0.686	2.947	3.738
M	0.430	0.314	0.305	M	-0.128	-0.221	-0.231
QR	0.249	0.138	0.118	QR	0.144	0.113	0.112
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	7.781	4.747	3.609	OLS	8.233	4.895	3.551
GLS	3.802	3.342	3.016	GLS	4.407	4.407	3.870
M	0.870	0.397	0.308	M	0.885	0.359	0.243
QR	0.691	0.302	0.156	QR	0.816	0.351	0.175
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	150.808	36.351	17.160	OLS	197.221	45.713	17.966
GLS	72.905	26.332	13.174	GLS	58.729	31.604	20.301
M	2.598	0.264	0.122	M	2.570	0.196	0.079
QR	1.083	0.154	0.039	QR	1.229	0.198	0.048
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.961	0.919	0.781	OLS	0.961	0.937	0.736
GLS	0.768	0.821	0.553	GLS	0.863	0.748	0.530
M	0.966	0.915	0.628	M	0.914	0.879	0.665
QR	0.991	0.953	0.896	QR	0.988	0.951	0.914
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.921	1.201	1.284	OLS	0.695	0.886	0.964
GLS	0.422	0.623	0.728	GLS	0.555	0.759	0.788
M	0.719	1.019	1.057	M	0.606	0.936	0.980
QR	2.580	1.054	1.017	QR	2.688	1.051	1.013

Table A.14: Estimated bias, average distance, mean squared error (MSE), coverage probability, and ratio of standard error (SE) and empirical standard deviation (ESD) for 4 different estimators of the intercept and slope (ordinary least squares, OLS; generalized least squares, GLS; M-; quantile regression (QR) estimator) with data generated as in scenario 14 (log-normal, heteroscedastic, outliers).

<i>Bias</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	5.387	3.926	4.012	OLS	2.042	4.801	4.930
GLS	1.007	1.380	2.431	GLS	-0.300	0.541	1.437
M	-0.093	-0.104	-0.106	M	-0.551	-0.663	-0.679
QR	0.141	0.030	0.009	QR	0.040	-0.001	-0.002
<i>Average distance</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	12.56	11.64	9.22	OLS	12.76	12.33	9.703
GLS	2.216	2.036	2.770	GLS	2.573	2.563	3.424
M	0.834	0.348	0.174	M	1.054	0.695	0.679
QR	0.757	0.308	0.137	QR	0.903	0.366	0.162
<i>Mean squared error</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	17765	3407	979.5	OLS	30562	4622	1977
GLS	135.6	48.88	56.16	GLS	67.79	43.21	44.61
M	1.292	0.188	0.046	M	1.861	0.616	0.493
QR	1.121	0.157	0.030	QR	1.464	0.212	0.042
<i>Coverage probability</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.982	0.967	0.931	OLS	0.939	0.965	0.968
GLS	0.704	0.822	0.864	GLS	0.850	0.893	0.851
M	0.956	0.955	0.933	M	0.830	0.558	0.041
QR	0.986	0.959	0.947	QR	0.982	0.958	0.953
<i>Ratio of SE & ESD</i>							
Intercept	20	100	500	Slope	20	100	500
OLS	0.130	0.276	0.434	OLS	0.085	0.205	0.264
GLS	0.167	0.302	0.359	GLS	0.282	0.417	0.520
M	0.997	1.080	1.076	M	0.786	0.940	0.962
QR	2.237	1.077	1.018	QR	2.055	1.083	1.018

Appendix B

Diagnostic Results for TcCB (ppb) Concentrations Data

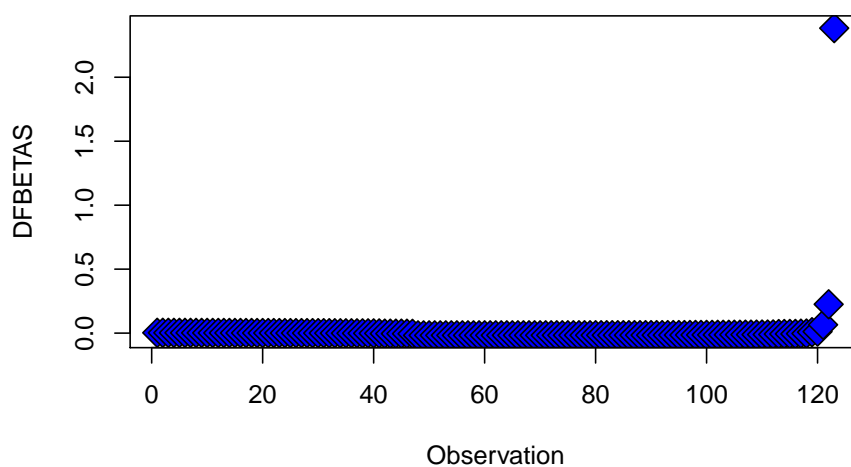


Figure B.1: Plot of DFBETAS with a cutoff 0.18 to detect influential observation for Area (slope) of OLS model ($TcCB \sim Area$).

Table B.1: DFBETAS of Area with corresponding influential observations of TcCB (ppb) concentrations.

Observation	TcCB	Area	DFBETAS
122	51.97	Cleanup	0.2254
123	168.64	Cleanup	2.3836