

**A COMPARATIVE STUDY OF KENDALL-THEIL SEN, SIEGEL VS QUANTILE
REGRESSION WITH OUTLIERS**

by

AHMAD FAROOQI

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: EDUCATION & EVALUATION

Approved By:

Advisor

Date

© COPYRIGHT BY

AHMAD FAROOQI

2019

All Rights Reserved

DEDICATION

I would like to dedicate this dissertation to my loving mother **Mehmooda Akhter**, who always prays for my success in this life and here after. I have a special feeling of gratitude to my family, colleagues, and teachers who have supported me throughout my studies.

ACKNOWLEDGMENTS

All praise is due to almighty **Allah** (*Subhanahu Wa Ta'ala*) alone, the Sustainer of all the worlds. Peace be upon the last messenger of Allah, **Prophet Muhammad** (*Blessing of Allah be upon him and his family*), the educator and the mercy to all mankind.

Doing PhD was my dream and this dream comes true only by my dissertation mentor and supervisor at work. I would like to say thank you so much to my mentor **Professor Dr. Shlomo S. Sawilowsky**, for his continuous support and guidance during my PhD program. I really enjoyed learning Non-parametric Statistics from him. Dr. Sawilowsky was always there to listen and to give advice. I believe without Dr. Sawilowsky it would have been quite difficult for me to conduct this dissertation.

Very special thanks goes to my dissertation committee members, **Dr. Thomas G. Edwards**, **Dr. Barry S. Markman**, and **Dr. Ronald L. Thomas**, for their kind help and worthy guidance enabled me to complete my dissertation. I personally feel that I am so lucky enough to get a maximum encouragement, supportive guidance, nice behavior, extreme co-operation and a very nice affection of my work supervisor **Dr. Ronald L. Thomas**, without which I may not have been able to conduct this research and complete my dissertation.

I would like to express my thanks to *Wayne State University* for providing me a full tuition scholarship towards my PhD program. I am also very grateful to the *Children's Research Center of Michigan* and the Department of Pediatric *Wayne State University School of Medicine* for their support. Last but not the least, I would like to thank my wife and kids whose love and sacrifice were an unending source of inspiration during my PhD studies.

TABLE OF CONTENTS

Dedication	ii
Acknowledgments	iii
List of Tables	v
List of Figures	ix
Chapter 1 Introduction	1
Chapter 2 Literature Review	16
Chapter 3 Methodology	43
Chapter 4 Results	51
Chapter 5 Discussion	133
References	151
Abstract	155
Autobiographical Statement	157

LIST OF TABLES

Table 1: Descriptive Statistics of (X, Y) for Regression Model passing through origin under the Normality Assumption with no Outliers:.....	53
Table 2: Results from the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $Cor(X, Y) = 0.80$	54
Table 3: Results of Relative Root Mean Square Error of the four regression procedures at $n=10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $Cor(X, Y) = 0.80$	55
Table 4: Descriptive Statistics of (X, Y) in regression procedures with $n= 10, 30, 50, 100$,.....	57
Table 5: Results of the four regression procedures with $n = 10, 30, 50, 100$ $Nsim = 1000$,	58
Table 6: Results of Relative Root Mean Square Error of the four regression procedures with $n= 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $e \sim Normaln, 0, 2$, and $Y = 2 + 3 * X + e$:	59
Table 7: Descriptive Statistics of Y variable in regression procedures with $n= 10, 30$,.....	61
Table 8: Frequency distribution of X variable in regression procedures with $n= 10, 30, 50$,	61
Table 9: Results of the four regression procedures with $n = 10, 30, 50, 100$ $Nsim = 1000$,	62
Table 10: Results of Relative Root Mean Square Error of the four regression procedures with $n= 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Binomialn, 1, 0.5$, $e \sim Normaln, 0, 2$ and $Y = 2 + 3 * X + e$	63
Table 11: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in both X and Y variables:	65
Table 12: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:...	66
Table 13: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, $Nsim = 1000$, $X \sim Normaln, 0, 1$,	67
Table 14: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in both X and Y variables:	69
Table 15: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:.....	70
Table 16: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, $Nsim = 1000$, $X \sim Normaln, 0, 1$,	71
Table 17: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in both X and Y variables:	73

Table 18: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:.....	74
Table 19: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, $Nsim = 1000$, $X \sim Normaln, 0, 1$,	75
Table 20: Descriptive Statistics of (X, Y) in regression analysis with $n = 100$ and outliers of 10%, 20%, 30%, and 50% of $n = 100$ in both X and Y variables:.....	77
Table 21: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$,	78
Table 22: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, $Nsim = 1000$, $X \sim Normaln, 0, 1$,	79
Table 23: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 10$ in Y variable only:	81
Table 24: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$,	82
Table 25: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:.....	83
Table 26: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 30$ in Y variable only:	85
Table 27: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$,	86
Table 28: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:.....	87
Table 29: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 50$ in Y variable only:	89
Table 30: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$,	90
Table 31: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, in Y direction only with $Nsim = 1000$, $X \sim Normaln, 0, 1$, $Y \sim Normaln, 0, 1$, and $CorX, Y = 0.80$:.....	91
Table 32: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 100$ in Y variable only:	93

Table 33: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	94
Table 34: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $CorX, Y = 0.80$:.....	95
Table 35: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, and $Y \sim pois(n, lambda = \lambda)$:	97
Table 36: Results from the four regression procedures with $n = 10, 30, 50$, and 100 , $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, and $Y \sim pois(n, lambda = \lambda)$:	98
Table 37: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $Y \sim pois(n, lambda = \lambda)$:.....	99
Table 38: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth Symmetric(n)$:.....	101
Table 39: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth Symmetric(n)$:.....	102
Table 40: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth Symmetric(n)$:	103
Table 41: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:	105
Table 42: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:	106
Table 43: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:	107
Table 44: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Bimodal(n)$:.....	109
Table 45: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Bimodal(n)$:	110
Table 46: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Bimodal(n)$:	111
Table 47: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass at Zero(n)$:.....	113
Table 48: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass at Zero(n)$:	114

Table 49: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Mass at Zero}(n)$:.....	115
Table 50: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Mass at Zero with Gap}(n)$:.....	117
Table 51: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Mass at Zero with Gap } n$:	118
Table 52: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Mass at Zerowith Gap}(n)$:....	119
Table 53: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Multimodal Lumpy}(n)$:.....	121
Table 54: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Multimodal Lumpyn}$:	122
Table 55: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Multimodal Lumpy}(n)$:	123
Table 56: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry - Decay}(n)$:	125
Table 57: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry - Decay}n$:	126
Table 58: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry - Decay}(n)$:	127
Table 59: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Digit Preference}(n)$:.....	129
Table 60: Results of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Digit Preference } n$:	130
Table 61: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Digit Preference}(n)$:	131

LIST OF FIGURES

Figure 1: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:	56
Figure 2: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $e \sim Normal(n, 0, 2)$, and $Y = 2 + 3 * X + e$:	60
Figure 3: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Binomial(n, 1, 0.5)$, $e \sim Normal(n, 0, 2)$, and $Y = 2 + 3 * X + e$:	64
Figure 4: Four regression lines are shown in each plot with $n = 10$ and outliers of 10%, 30% and 50% of $n = 10$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	68
Figure 5: Four regression lines are shown in each plot with $n = 30$ and outliers of 10%, 30% and 50% of $n = 30$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	72
Figure 6: Four regression lines are shown in each plot with $n = 50$ and outliers of 10%, 30% and 50% of $n = 50$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	76
Figure 7: Four regression lines are shown in each plot with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:	80
Figure 8: Four regression lines are shown in each plot with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in Y only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	84
Figure 9: Four regression lines are shown in each plot with $n = 30$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	88
Figure 10: Four regression lines are shown in each plot with $n = 50$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	92
Figure 11: Four regression lines are shown in each plot with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in Y only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,	96
Figure 12: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $Y \sim pois(n, lambda = \lambda)$:	100
Figure 13: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Smooth Symmetric(n)$:	104
Figure 14: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:	108
Figure 15: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Extreme Bimodal(n)$:	112

Figure 16: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$,
 $X \sim Unif(n, 1, 10)$, $Y \sim \text{Mass at Zero}(n)$: 116

Figure 17: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$,
 $X \sim Unif(n, 1, 10)$, $Y \sim \text{Mass at Zero with Gap}(n)$: 120

Figure 18: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$,
 $X \sim Unif(n, 1, 10)$, $Y \sim \text{Multimodal Lumpy}(n)$: 124

Figure 19: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$,
 $X \sim Unif(n, 1, 10)$, $Y \sim \text{Extreme Asymmetry - Decay}(n)$: 128

Figure 20: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$,
 $X \sim Unif(n, 1, 10)$, $Y \sim \text{Digit Preference}(n)$: 132

CHAPTER 1 INTRODUCTION

Research in educational evaluation is an important strategy for improving the quality of education in universities, which may lead to a better understanding of the strengths of education. Researchers are always interested to study the relationship between a response variable and one or more independent predictors, either for the purpose of explanation or prediction. However, a weakness invariably occurring and often found in studies of educational research are the presence of data outliers in the response variable or both in response variable and predictor variable. The presence of outliers can increase error variance and reduce the power of statistical tests. Unfortunately, outliers can often times be impossible to prevent even when data has been carefully collected from respectable sources. Since outliers affect analysis and interpretation of statistical outcomes, a greater understanding of statistical procedures for handling the presence of outliers in educational datasets is needed. Specifically, identifying which procedures operate best in identifying and handling data corruption from outliers would be a welcome contribution to the field for scholars, data analysts, and educational researchers. Institutions can benefit from the results of this research by providing recommendations useful in practice and substantive interest, leading to more accurate and reliable conclusions and decision making.

In statistics, correlation and regression are commonly used statistical techniques to study the relationship between two or more numerical variables. Correlation is used to measure the linear relationship between the variables in terms of a correlation coefficient. The Pearson's linear correlation coefficient (r), measures the strength and direction of the relationship between Y_i and X_i . A positive value of r indicates an increasing linear relationship where as a negative value of r indicates a negative linear relationship. A value of r equal 0, indicate no linear relationship between the variables. A negative linear relationship is said to be strong if the value of r close to -1 and for a strong positive linear relationship the value of r is close to +1. On the other hand, a linear relation-

ship is said to be weak if the value of r close to zero. If the two variables are correlated, the value of a (the response) variable can be predicted from the value of other (the predictor) variable using the regression analysis.

One purpose of a regression analysis is to determine if there is a relationship between a response variable Y_i and one or more independent predictors X_i with a link function f , either for the purpose of explanation or prediction. The response variable Y_i cannot be predicted exactly from the independent predictors X_i . Normally, the behavior of response variable Y_i is summarized for each given predictor X_i with typically used measures of location called mean, median or mode.

For n pair of data (Y_i, X_i) this relationship can be modeled as

$$Y_i = f(X_i) + \varepsilon_i, i = 1, 2, \dots, n$$

where f could be a linear or nonlinear function of X_i , ε_i are the random errors, independently and normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$, and ε_i and ε_j are uncorrelated, i.e. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0; \forall i \neq j$.

There are two main approaches to model the link function f . They are (1) the parametric approach and (2) the nonparametric approach.

Ordinary Least Square (OLS) Regression is a parametric approach used to study the relationship between a response variable Y_i with at least one predictor X_i , also called an independent variable by describing the mean of response variable for each value of the given predictors, using a function called the conditional mean of the response variable. This relationship can be created by developing a statistical model with certain unknown population parameters called regression coefficients. The parameters are then estimated by the method of least square and the fitted model is used to get an approximate idea of the trend for prediction and forecasting of the data. A linear regression is reported when one of the estimated regression coefficients called slope of the fitted line is demonstrated to be statistically significant from zero, using a Student's t-test, a slope greater than zero indicates an increasing trend and a slope less than zero indicates a decreasing trend.

Mosteller and Tukey (1977) stated

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of X 's. We could go further and compute several different regression curves corresponding to the various percentage points of the distribution and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions. (p 266)

Ordinary Least Square (OLS) Regression has certain attractive properties

- OLS estimators are linear and unbiased to their correspond parameters.
- The variance of the estimators is the indicator of the accuracy of the estimators.
- OLS estimators are best linear unbiased estimators (BLUE) of their correspond parameters.
- OLS estimators are consistent estimators i.e. as n larger and larger, the estimator become closer and closer to their correspond estimator.
- OLS estimators have the minimum variance among all the unbiased linear or nonlinear estimators.

There are certain disadvantages of OLS. It provides only a partial view of the relationship between response and independent variable(s) through a conditional mean point of the response variable. There may be interest in studying the relationship at different points in the conditional distribution of response variable. Also, it does not fit well when there are some regular outliers present in the response variable, or if the data were sampled from a non-normal distribution. Therefore, the OLS regression estimator is not robust (e.g., Hampel et al., 1987; Huber & Ronchetti, 2009; Maronna et al. 2006; Staudte & Sheather, 1990; Wilcox, 2012a, b).

There are various alternatives to the OLS modeling has roots that can be traced to the mid of the 18th century. There are several alternative to the OLS, one of the alternative approaches can be referred to as median regression, where focus of the modeling is the median instead of mean. It is to me noted that median is the special case of quantile, which can be used to model the non-central position of a distribution. This idea can be extended to other quantile like quartile; decile and percentile can be used to specify any position of the distribution.

Quantile Regression is another very flexible approach that can be used to study the relationship between a response variable Y_i with at least one predictor X_i at different points in the conditional distribution of response variable, using the conditional median or other quantile functions, where the median is the 50th percentile; and the quartiles are the 25th, 50th and 75th percentiles. Similarly, the deciles are the 10th, 20th and so on until the 90th percentile of the empirical distribution can be used to study the response variable. For example, we can study different factors affecting student scores along their score distribution and we can imagine those factor could be different or they might affect differently those student scores for the students for their very high performance with their high scores and for those who have low student scores. In this case we can use regression procedure.

A generalization, aimed at dealing with any Quantile Regression model first introduced and derived by Koenker and Bassett (1978), which models the conditional quantiles of a response variable as a function of predictors. The regression procedure is the natural extension of OLS, where instead of specifying the change in conditional mean of the response variable associated with a unit change in the predictor variable, the regression procedure specifies the change in conditional quantile of the response variable associated with a unit change in the predictor variable. Regression procedure does not assume a particular distribution for the response, nor does it assume a constant variance for the response, unlike ordinary least squares regression.

OLS regression does not fit well when there are some regular outliers present in the response variable, or if the data were sampled from a non-normal distribution. Statisticians are used statistical methods in regression that are resistant to the outliers and non-normality of the distribution. Robust methods are invoked to refine the conditional mean and heterogeneity of the variance. They are not only sensitive to the non-normality of the data, but also minimize the effect of assumptions about data below detection limits, and the effect of outliers on the determination of relations between variables (Helsel & Hirsch, 2002).

The first step toward a robust regression estimator came from Edgeworth (1887), improving a proposal of Boscovich. Kendall (1938) provided a non-parametric method of detecting a relationship between two variables and to find a suitable fit when there is a problem of outliers in the data, it provides different option to examine lines between all pairs of points, and estimate the slope by the median of all slopes, and intercept by the median of all intercepts.

In a simple linear regression, Theil (1950) first proposed another robust linear regression method where there are one response and one predictor variable and is robust to outliers in the response variable. In this method, the slope of the regression line is estimated as the median of all pairwise slopes between each pair of points in the dataset. Sen (1968) extended this estimator to handle ties. The Theil-Sen estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses a high asymptotic efficiency.

A modified and preferred method is Siegel (1982), where the repeated median is used, the repeated median algorithm is a robustified U-statistic (Hoeffding, 1948), in which nested medians replaces the single median and has a 50% breakdown point. This is the best that can be expected, because for larger amounts of contamination, it becomes impossible to distinguish between the good and the bad parts of the sample. When the dependent variable is continuous, the Theil–Sen

estimator enjoys good theoretical properties and it performs well in simulations in terms of power and Type I error probabilities when testing hypotheses about the slope (e.g., Wilcox, 2012b).

Regular outliers and non-normality are problems in statistics, an outlier is a value that is far from the general distribution of the other observed values, and can often perturb the results of a statistical analysis (Michael Greenacre & H. Ayhan, 2015). As a result of these outliers there may be a breakdown in the model at the *ith* point produce a location shift and the variance exceed the error variance at the other data points, also there may be a large random disturbance that can be produced by chance. By contrast, an inlier is a data value that lies within a statistical distribution and is in error. Although the normality of the data takes center place in statistics, most of the data in behavioral and social sciences follow a non-normal data. A non-normality describes the shape of a data as being in that of not a bell curve when a variable along the x-axis and the corresponding frequencies of the variables along the y-axis is plotted.

As discussed, OLS estimates are seriously affected by outliers and non-normality, especially when the sample size is not very large. Although the log transformation can be used to handle the non-normality of the data, but this does not always work and still underline assumptions of OLS need to satisfy. Regression procedure with median not only more robust to outliers and non-normal errors but also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y , not merely its conditional mean (Christopher Baum, 2013). Numerical experiments indicate that the Kendall-Theil slope estimator is almost as efficient as OLS regression under ideal conditions for OLS and is much more efficient than OLS even when conditions do not depart substantially from the ideal (Hussain and Sprent, 1983; Dietz, 1987; Hirsch and others, 1991; Brauner, 1997; Nevitt and Tam, 1998; Helsel and Hirsch, 2002).

Objectives of the Study

Several comparisons are made between Ordinary Least Square Regression, Quantile Regression, Kendall–Theil Sen and Siegel Regression, but no direct comparison is yet made between Quantile Regression and Kendall–Theil Sen Siegel Regression in the presence of outliers. The research hypotheses of the dissertation are

- In the presence of outliers, Theil Sen Siegel Regression will have narrower confidence intervals for the regression coefficients than either Ordinary Least Square or Quantile Regression.
- In the presence of outliers, Theil Sen Siegel Regression will have lower Root Mean Square Error, Standard Errors and Average bias index values than either Ordinary Least Square or Quantile Regression
- In the presence of outliers, Theil Sen Siegel Regression will be more robust to maintain Type I and Type II error rates either Ordinary Least Square or Quantile Regression across three non-normal density functions.

In all four approaches, outliers will be modeled as (a) outliers, and (b) non-normal distributions. Different test statistics, such as biases, Standard Deviation (S.D), Standard Error (SE), R-squared, the overall F-test, Median Absolute Error (RMEDAE) and Root Mean Square Error (RMSE) will be used to evaluate the model fit.

Assumptions in Parametric and Non-Parametric Regressions

Linear regression as a parametric statistical technique makes several underlying assumptions. Among those the most important assumptions are

- Linearity of outcome variable with a predictor.
- Normality of residuals/errors.
- The variability in response variable at different levels of predictors should be homogeneous.

- Independence of residuals/errors.

Significant lack of symmetry and outliers can produce invalid and bias results. As a non-parametric regression, procedures do not require any specific underlying distribution for the given set of data.

The only assumption is

- The errors are assumed to be statistically independent.

Limitations

- With non-parametric regression approach, confidence Interval's construction is sometimes difficult.
- Non-parametric regression approach is not as powerful in terms of inference as parametric regression.
- This study did not evaluate effects of missing or inconsistent data.
- There are several other non-parametric regression methods, but our focus is on Quantile Regression and Kendall–Theil Sen Siegel Regression.
- This study focused on 50th percentile i.e. median in Quantile Regression only.
- This study focused on simple linear regression only.

Definitions and Terminologies

Alternative Hypothesis:

Any hypothesis which is different from null hypothesis and set parallel to the null hypothesis called an alternative hypothesis. An alternative hypothesis is usually denoted by H_1 (or H_a) and must contain a sign of inequality (\neq , $>$ or $<$).

Alpha Level (Level of significance):

An Alpha level or level of significance denoted by α is the probability that the test statistic will fall in the rejection region when the null hypothesis is actually true, in other words level of significance is the probability of making type-I error. Common levels of significant are 5%, 2%, and 1%. By 5%

level of significant means there is 1 chance out of 20 that a true null hypothesis H_0 is rejected this is rare that's not happen by chance but it is happen due to the intervention.

So,

$\alpha = P(\text{Type-I Error}) = P(\text{Reject } H_0 \mid H_0 \text{ is True}) = \text{Level of Significance}$

Assumptions:

In parametric statistical analysis, assumptions are the certain pre assumed characteristics about the data.

Confidence Interval:

Confidence Interval is an interval that contains the unknown population parameter (θ) with certain degree of confidence. In general, a confidence interval can be written as

Point estimate \pm (reliability coefficient) *(standard error of the point estimate)

Confidence Limits:

Confidence limits are the lower and upper boundaries / values of a confidence interval which define the range of a confidence interval.

Correlation:

The correlation is an association or relationship (in terms of variability) between two variables Y_i and X_i . Most people equate Y_i and X_i being correlated to mean that Y_i and X_i are associated, related, linearly overlap, or dependent upon each other. However, correlation is only a measure of the strength and direction of a linear relationship.

Deciles:

Deciles are the process of dividing the data in to ten equal parts. The 1st decile (D_1) has the 1/10(10%) of the data below it, the 2nd decile (D_2) has the 2/10(20%) of the data below it so on, the 9th decile (D_9) has the 9/10(90%) of the data below it.

Distribution Function:

The distribution function of a random variable Y is denoted by $F(Y)$ and defines as $F(Y) = P(Y \leq y)$ i.e. the function $F(Y)$ gives the probability of an event Y takes a value less than or equal to a specific value of y .

Estimation of Parameters:

Estimation of parameters is the process of making judgment about the unknown parameters on the bases of sample statistics.

Estimate:

The results obtained after applying the formula called an estimate.

Estimator:

A rule or formula that can be used to estimate the unknown population parameters called an estimator.

Inferential Statistics:

Inferential statistics allows a researcher to draw conclusion about the unknown population parameters based upon sample statistics.

Level of Confidence (or Confidence Coefficient):

The probability associated with a confidence interval called level of confidence; normally, 90%, 95%, or 99% level of confidence are used.

Linear Regression:

Linear regression is a statistical technique used to study the relationship between a response variable Y_i with at least one predictor variable X_i by developing a model with conditional mean and certain parameters, these parameters are estimated and the fitted model is then used for forecasting and prediction.

Linear Relationship:

If a scatter plot of dependent variable Y_i and an independent variable X_i shows a straight-line trend then the relationship between the two variables is said to be linear.

Linearity:

Linearity is one of the important assumptions in simple linear regression, when the given predictor X_i does have a linear relationship with outcome variable Y_i .

Mean:

Mean (or Arithmetic Mean) is just a sum of given values, divided by the number of values.

Mean Absolution Deviation:

Mean Absolution Deviation is the ratio of sum of absolute deviation from mean to the total number of observations.

Median:

Median is a value which divides the data in to two equal parts, after arranging the values into increasing or decreasing order of magnitude.

Method of Least Square:

It is a method by which the curves and equations are fitted to the data, this method consists of minimizing the sum of square of errors.

Model:

Model is a relationship between the variables in terms of mathematical equation i.e. $Y_i = a + bX_i + \varepsilon$.

Monte Carlo methods:

A method used for making inference or exploring distribution properties by repeated sampling; it is especially useful when an analytic solution is difficult to obtain (Nelson, 1998, p. 287).

Non-parametric statistical methods: Statistical methods designed to be used when the data being analyzed depart from the distributions that can be analyzed with parametric statistics (Vogt, 1993, p. 155).

Normal distribution:

A theoretical continuous probability distribution in which the horizontal axis represents all possible values of a variable and the vertical axis represents the probability of those values occurring (Vogt, 1993, p. 155).

Null Hypothesis:

A null hypothesis is any hypothesis which is to be tested for possible rejection under the assumption that it is true. A null hypothesis is denoted by H_0 and must contain a sign of equality ($=$, \leq or \geq).

Outliers:

Outliers are the data points that do not fit well with the pattern of the rest of the data.

Predictor Variable:

A predictor (Independent, covariate, explanatory, regressor, or factor) variable is a presumed cause in an experimental study. The values of the independent variable are under researcher control.

P-value:

A p-value is the probability of getting a value of the sample test statistic that is at least as extreme as the one found from the sample data, assuming that null hypothesis is true.

Parameter:

Parameters are the numerical information obtained from a population data.

Parametric statistical Methods:

Statistical techniques designed when data have certain characteristics-usually when they approximate a normal distribution and are measurable with interval or ratio scales (Vogt, 1993, p. 165).

Percentiles:

Percentiles are the process of dividing the data in to hundred equal parts. The 1st percentile (P_1) has the 1/100(1%) of the data below it, the 2nd percentile (P_2) has the 2/100(2%) of the data below it so on, the 99th percentile (P_{99}) has the 99/100(99%) of the data below it.

Probability Density Function:

A function $f(x)$ is said to be a probability density function (pdf) over an interval $[a, b]$ where $a < b$, if the area between the given points $x = a$ to $x = b$ gives the probability that x lies between a and b , i.e.

$$P(a \leq x \leq b) = F(b) - F(a) = \int_a^b f(x_i) dx$$

Population:

In statistics, a population consist of all possible observations (or individuals or subjects, or content) of interest.

Quartiles:

Quartiles are the process of dividing the data in to four equal parts. The 1st quartile (Q_1) has the 1/4(25%) of the data below it, the 2nd quartile (Q_2 or Median) has the 2/4(50%) of the data below it, the 3rd quartile (Q_3) has the 3/4(75%) of the data below it.

Quantile Regression:

Quantile Regression is a non-parametric statistical technique used to study the relationship between a response variable Y_i with at least one predictor variable X_i by developing a model with conditional quantiles and certain parameters, these parameters are estimated and the fitted model is then used for forecasting and prediction.

Regression: Regression is a statistical technique used to study the relationship between a response variable Y_i with at least one predictor variable X_i by developing a model with certain parameters, these parameters are estimated and the fitted model is then used for forecasting and prediction.

Regression Models:

The mathematical models that allows predicting the value of response variable from known values of one or more predictor variables.

Regression Coefficient (or Slope):

In statistics, regression coefficient is a change in response variable Y_i due to unit changes in predictor X_i . A slope of 2 means that for every 1-unit change in a predictor, yields a 2-unit change in response variable.

Relative Root Mean Square Error (RRMSE):

A statistic used to measure the relative performance of two estimation methods based on mean.

Relative Median Absolute Error (RRMEDAE):

A statistic used to measure the relative performance of two estimation methods based on median.

Response Variable:

A response (outcome, explained, dependent) variable is a presumed effect in an experimental study whose values depend upon another variable, called independent variable.

Robust:

A statistical method that is relatively insensitive to departures from a postulated assumption (Hollander & Wolfe, 1996, p. 132).

Robust Estimator:

A robust estimator is one that has a high breakdown point.

Sample:

A sample is a set of observations drawn from a population.

Sampling Distribution:

The probability distribution of a sample statistic is called sampling distribution.

Scatter Plot:

A scatter plot is used to show the score on one variable plotted against score of a second variable.

Standard Deviation:

Standard Deviation is the ratio of sum of squared deviation from mean to the total number of observation and is used to measure the variability of a set of data.

Standard Error (S.E):

It is the standard deviation of sampling distribution of a sample statistic.

Statistic:

Statistics are the numerical information obtained from sample data.

Test of Significance:

Test of Significance is a process of assessing evidence provided by the data in favor of, or against some claim about the characteristics of population.

Test Statistic:

A test statistic is a rule or formula used in test of significance and its value is calculated from the sample data that is used in making the decision about the rejection of the null hypothesis.

Type I error:

A Type-I error occurs when we reject a null hypothesis H_0 when, in fact, H_0 is True.

Type II error:

A Type-II error occurs when we fail to reject a null hypothesis H_0 when, in fact, H_0 is False.

CHAPTER 2 LITERATURE REVIEW

Ordinary Least Square Regression

Ordinary Least Square Regression is a statistical method which helps in estimating an average relationship between two or more variables by using the method of least square. This method allows explanation and prediction of the unknown value of one variable called response variable Y_i from known value of related variable called predictor variable X_i . Normally in Ordinary Least Square Regression the behavior of response variable Y_i is summarized for each given predictor X_i with typically used measures of location called mean. In order to estimate this average relationship, the concept of sampling distribution must be employed.

Sampling Distribution of Sample Statistic:

The sampling distribution of a sample statistic $\hat{\theta}$ is the probability distribution or the relative frequency distribution of all possible random samples of the same size that could be selected from a given population.

Sampling Distribution of Sample Mean:

The sampling distribution of sample mean \bar{X} is the probability distribution or the relative frequency distribution of all possible random samples of the same size that could be selected from a given population? The mean of this distribution represented by $\mu_{\bar{X}}$ and the standard deviation of this distribution is called standard error of estimate denoted by $\sigma_{\bar{X}}$ which indicates the variability of the distribution of all possible sample means.

Standard Error of Estimates in Ordinary Least Square Regression:

The degree of scatterness of the observed values about the regression line is measured by what is called standard error of regression or standard deviation of Y on X, denoted by $S_{y.x}$ and define as

$$S_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

It is to be noted that if $S_{y.x} = 0$, this shows that all the given data points lie of the Ordinary Least Square Regression line.

Ordinary Least Square Regression Model with Single Independent Predictor:

Ordinary Least Square (OLS) regression is a parametric approach used to study the relationship between a response variable (Y_i) with at least one predictor or independent variable (X_i). In a simple linear regression model, suppose a sample of n pair of observation (X_i, Y_i), $i = 1, 2, \dots, n$ was taken from a normal population to fit a model such that it will best fit the data. In order to do that a straight line with an equation below is used

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Where, ε_i 's are random errors (residuals) and assumed to be independent of X_i and normally distributed with zero mean i.e. $E(\varepsilon_i) = 0$ and variance σ^2 i.e. $\text{Var}(\varepsilon_i) = \sigma^2$, a constant for all X_i . These assumptions also imply that Y_i also have common variance σ^2 as the only element in the model is ε_i .

A method of least square (LS) in which curves and equations are fitted to the data is used. This method consists of minimizing the sum of squared error ε_i from the fitted straight line to the observed outcome variable Y_i , i.e. $\sum_{i=1}^n \varepsilon_i^2$.

Let the equation of the least square model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Then the conditional mean of Y_i given X_i can be written as,

$$\hat{Y} = E(Y_i | X_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i)$$

$$\hat{Y} = E(Y_i | X_i) = \beta_0 + \beta_1 X_i,$$

as $E(\varepsilon_i) = 0$ and $E(X_i) = X_i$.

The sum of squared error is

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y})^2.$$

To minimize it, partially differentiate it with respect to β_0 and β_1 and equate to zero the derivatives.

Estimation of parameters in Ordinary Least Square Regression with Single Independent Predictor:

Suppose, $\hat{\beta}_0$ & $\hat{\beta}_1$ are the estimates of the corresponding parameters β_0 and β_1 in an Ordinary Least Square Regression model. The method of least square estimates the parameters by minimizing the sum of squared errors, denoted by S^2 and given by

$$\begin{aligned} S^2 &= \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

Now to minimize sum of squared errors, partially differentiate S^2 with respect to β_0 and β_1 and equate to zero the, i.e.

$$\frac{\partial S^2}{\partial \beta_0} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$\frac{\partial S^2}{\partial \beta_1} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0.$$

After solving the equations, the estimates of β_0 and β_1 are given below,

$$\hat{\beta}_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_0 = \bar{Y}_i - b \bar{X}_i.$$

Hence, the estimated ordinary least square regression line is given as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

It is evident in an Ordinary Least Square Regression line, the quantities $\hat{\beta}_0$, $\hat{\beta}_1$, \hat{Y} and \bar{Y} will vary from one sample of data to another. They are thus random variables and hence have their sampling distribution.

Interpretation of Ordinary Least Square Regression Estimates with Single Independent Predictor:

In an Ordinary Least Square Regression, $\hat{\beta}_1$ is the estimated rate of change in the average value of the response variable Y_i for a one-unit change in predictor variable X_i . Similarly, $\hat{\beta}_0$ is the estimated average value of Y_i when the value of X_i is zero (which may not always be sensible).

Reference and comparison in Ordinary Least Square Regression Estimates:

When a predictor variable in an Ordinary Least Square Regression is categorical, and to facilitate the interpretation of the categorical OLS estimates, uses the notation of reference and comparison with some ideas related to the quantification of effects can be used. For example, in a dichotomous categorical variable, one category may be used as reference and compared other category to study the effect of a unit change in response variable. This idea can be extended to more than categories in a categorical predictor. In OLS, the fitted categorical coefficient can be interpreted as an estimated effect i.e. estimates of the change in the mean of the response variable that results from a 1-unit change between reference category and comparison category.

Statistical Inference in Ordinary Least Square Regression with Single Independent Predictor:

For purpose of statistical inference, the concept of means, variances and the shapes of these sampling distributions must be known. The mean and variance of the sampling distribution of estimated regression coefficient $\hat{\beta}_1$ can be driven as

$$\begin{aligned} \text{Mean}(\hat{\beta}_1) &= \mu_{\hat{\beta}_1} = \beta_1 \\ \text{Var}(\hat{\beta}_1) &= \sigma_{\hat{\beta}_1}^2 = \sigma^2 / \sum (X_i - \bar{X})^2. \end{aligned}$$

Because the random errors ε_i 's are assumed to be independent of X_i and normally distributed, therefore the distribution of $\hat{\beta}_1$ is also normally distributed with mean $\mu_{\hat{\beta}_1} = \beta_1$ and

$\sigma_{\hat{\beta}_1}^2 = \sigma^2 / \sum(X_i - \bar{X})^2$. Generally, σ^2 is unknown, it is, therefore, requiring an estimate for σ^2

from the sample data by $S^2_{y.x} = \frac{\sum(Y_i - \hat{Y})^2}{n-2}$. Similarly, the mean and variance of the sampling distribution of y-intercept $\hat{\beta}_0$, can be driven as

$$\text{Mean}(\hat{\beta}_0) = \mu_{\hat{\beta}_0} = \beta_0$$

$$\text{var}(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right]$$

Because the random errors ε_i 's are assumed to be independent of X_i and normally distributed, leads to the fact that the distribution of $\hat{\beta}_0$ is also normally distributed with mean $\mu_{\hat{\beta}_0} = \beta_0$ and

$$\sigma_{\hat{\beta}_0}^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right],$$

where σ^2 can be estimated from the sample data by $S^2_{y.x} = \frac{\sum(Y_i - \hat{Y})^2}{n-2}$.

Interval Estimation in Ordinary Least Square Regression with Single Independent Predictor:

Confidence Interval for Population regression coefficients can be obtained as follows. A 100(1 - α) % Confidence Interval for Population regression coefficient β_0 is given by

$$(\hat{\beta}_0 \pm t_{\alpha/2(n-2)} S.E(\hat{\beta}_0)).$$

Similarly,

A 100(1 - α) % Confidence Interval for Population regression coefficient β_1 is given by

$$(\hat{\beta}_1 \pm t_{\alpha/2(n-2)} S.E(\hat{\beta}_1)).$$

Test of Significance in Ordinary Least Square Regression with Single Independent Predictor:

To test the null hypothesis about the slope of the regression line, $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ a test statistic $t = \frac{b - \beta_1}{S.E(b)}$ with $(n - 2)$ d.f can be used. Similarly, to test the null hypothesis about the y-intercepts of the regression line $H_0: \beta_0 = 0$ against $H_1: \beta_0 \neq 0$, a test statistic $t =$

$\frac{a - \beta_0}{S.E(a)}$ with $(n - 2)$ d.f. can be used.

Measures of Variability in Ordinary Least Square Regression:

To determine the regression coefficients for a given set of data by the method of least square, there are normally three important measures of variability are used, named as total variability measured by total sum of square (TSS), explained variability measured by regression sum of square (RSS) and unexplained variability measured by error sum of square (SSE). Note that the total variability of the data can be partitioned into two explained variability and unexplained variability, i.e. $TSS = RSS + SSE$.

The Coefficient of Determination in Ordinary Least Square Regression with Single Predictor:

A coefficient of determination is the proportion of total variability in the response variable Y_i that is explained by the (variability in X_i) regression model, denoted by R^2 and given by $R^2 = 1 - (SSR/SST)$. The value of R^2 varies from 0 to 1 and is free from the unit of measurement. The value of R^2 close to 1 implies that most of the variability in the response variable Y_i is explained by the regression but this does not mean that the regression is predicting accurately.

Ordinary Least Square Regression Model with Multiple Independent Predictors:

Suppose a response variable (Y_i) along with k independent predictors ($X_{i1}, X_{i2}, \dots, X_{ik}$) were selected with n pair of observation, then a multiple linear regression model with k predictors can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i,$$

where, ε_i 's are again random errors (residuals). The random errors ε_i 's are again assumed to be independent of X_{ik} and normally distributed with zero mean i.e. $E(\varepsilon_i) = 0$ and variance σ^2 i.e. $\text{Var}(\varepsilon_i) = \sigma^2$, a constant for all X_{ik} . These assumptions also imply that Y_i also have common variance σ^2 as the only element in the model is ε_i .

Again, by the method of least square (LS) which consists of minimizing the sum of squared error from the fitted plane instead of straight line to the observed outcome variable Y_i , i.e. $\sum_{i=1}^n \varepsilon_i^2$.

Let the equation of the least square model be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i.$$

Then, the conditional mean of Y_i given X_{ik} can be written as,

$$\hat{Y} = E(Y_i | X_{ik}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + E(\varepsilon_i)$$

$$\hat{Y} = E(Y_i | X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

as $E(\varepsilon_i) = 0$ and $E(X_{ik}) = X_{ik}$.

The sum of squared error is

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y})^2$$

To minimize it, partially differentiate with respect to β_i 's equate to zero the derivatives. Computer can be used to estimates for β_i 's and called partial regression coefficients denoted by $\hat{\beta}_i$'s.

Estimation of parameters in Ordinary Least Square Regression with Multiple Independent Predictors:

Suppose a response variable (Y_i) with k independent predictors ($X_{i1}, X_{i2}, \dots, X_{ik}$). A multiple linear regression model with k predictors can be written as

$$Y = X\beta + \xi$$

Where, Y is an $(n \times 1)$ matrix of the response variable, X is an $n \times k$ matrix of the predictor variables, β is a $(k \times 1)$ matrix of the regression coefficients and ξ is a $(n \times 1)$ matrix of the error terms with $E(\xi) = 0$, $var(\xi) = \sigma^2 I$, $E(Y) = X\beta$ and I is the identity matrix.

Now,

$$\begin{aligned} S^2 &= \xi' \xi = (Y - X\beta)'(Y - X\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \end{aligned}$$

Partially differential with respect to β and equate to zero the derivative

$$\frac{\partial S^2}{\partial \beta} = -2X'Y + 2X'X\beta = 0$$

$$X'Y = X'X\beta$$

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

So, the fitted plane is

$$\hat{Y} = X\hat{\beta}$$

Interpretation of Ordinary Least Square Regression Estimates with Multiple Independent Predictors:

In multiple least square regressions $\hat{\beta}_j$ is the estimated rate of change in the average value of the response variable Y_i for a one-unit change in predictor variable X_j , keeping all other X_{k-j} predictors constant. For example, $\hat{\beta}_1$ is the estimated rate of change in the average value of the response variable Y_i for a one-unit change in predictor variable X_1 , keeping all other X_{k-1} predictors constant.

Confidence Interval in Ordinary Least Square Regression with Multiple Independent Predictors:

The Confidence intervals for individual parameters are given by

$$\hat{\beta}_j \pm t_{\alpha/2(n-k-1)} S.E(\hat{\beta}_j), j = 0, 1, 2, \dots, k.$$

The Coefficient of Determination in Multiple Least Square Regression with Multiple Predictors:

A coefficient of determination is the proportion of total variability in the response variable Y_i that is explained by the (variability in X_i) regression model, denoted by R^2 and given by $R^2 = SSR/SST$. The value of R^2 varies from 0 to 1 and is free from the unit of measurement. The value of R^2 close to 1 implies that most of the variability in the response variable Y_i is explained by the regression but this does not mean that the regression is predicting accurately.

Goodness of Fit in Multiple Least Square Regression:

In ordinary least squares regression models, the goodness of fit of the model is measured by coefficient of determination method after the different regression models have been estimated to select the most appropriate model. A coefficient of determination is denoted by R^2 and defined as

$$R^2 = \frac{SSR}{SST} = \frac{\sum_I(\hat{Y}_I - \bar{Y})^2}{\sum_I(Y_I - \bar{Y})^2}$$

Akaike Information Criterion (AIC) and Schwarz Criterion (SC) can be used for the goodness of fit of the regression model.

Outliers and Least Square Regression:

Regular outliers and non-normality are the well-known problems in statistics, an outlier is a value that is far from the general distribution of the other observed values, and can often perturb the results of a statistical analysis (Michael Greenacre & H. Ayhan, 2015). As a result of these outliers there may be a breakdown in the model at the *ith* point produce a location shift and the variance exceed the error variance at the other data points, also there may be a large random disturbance that can be produced by chance. Least square estimates are seriously affected by outliers and non-normality, especially when the sample size is not very large. Although log transformation can be used to handle the non-normality of the data, but this does not always work and still underline assumptions of OLS need to satisfy.

Quantile Regression

Quantile Regression is another very flexible approach developed by Koenker and Bassett (1978), that can be used an alternative of ordinary least square regression and allows researcher to study the relationship between a response variable Y_i with at least one predictor X_i at different points in the conditional distribution of response variable Y_i , at several points using the conditional median function $Q_q(Y_i|X_i)$, or other quantile function where median is the 50th percentile and is the best-known quantile, similarly, the other quartiles, e.g. 25th, 30th, 75th, and so on 95th percentiles or

simply a q th quantile q , of the empirical distribution $F(Y)$ can be defined. It is to be noted that quantile and percentiles are synonymous i.e. the 0.90 quantile is the 90th percentile. Using Quantile Regression investigator/researcher can see a more comprehensive picture of the effect of the predictors on the response variable.

Cumulative Distribution Function (cdf) and Quantile Function:

In order to describe the empirical distribution of a random variable Y , in Quantile Regression, the concept of distribution function or cumulative distribution function (cdf) can be used. The cumulative distribution function of a random variable Y is denoted by $F(Y)$ and defines as

$$F(Y) = P(Y \leq y),$$

i.e. the function $F(Y)$ gives the probability of an event Y takes a value less than or equal to a specific value of y . It is to be noted that $F(-\infty) = 0$ and $F(\infty) = 1$. Also $F(Y)$ is a non-decreasing function and is continuous at least on the right of each y .

$$F(Y) = P(Y \leq y) = \sum_i f(y_i), \text{ if } Y \text{ is discrete random variable}$$

$$F(Y) = P(Y \leq y) = \int_{-\infty}^{+\infty} f(y_i) dy, \text{ if } Y \text{ is continuous random variable}$$

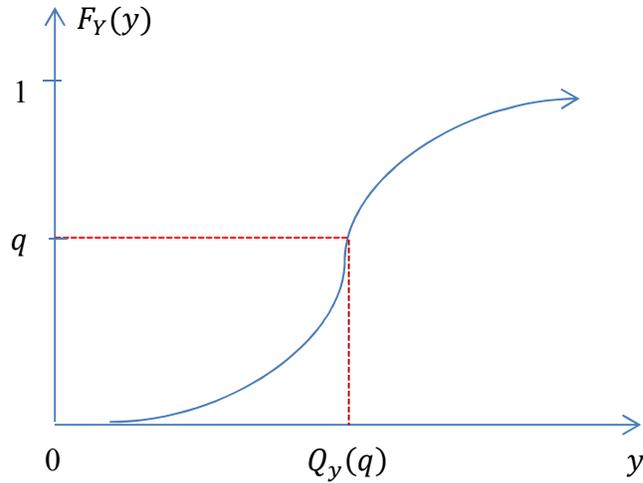
Note

$$P(a \leq y \leq b) = F(b) - F(a) = \int_a^b f(y_i) dy.$$

Continuing the explication of a cdf, F , for some population characteristics, the q th quantile of the distribution, denoted by Q_q such that $Q_q = q^*$. In a standard normal distribution, $F(1.96) = .96$. So $Q_{0.95} = 1.96$. The quantile function for q th quantile, $0 < q < 1$ split the response variable Y_i into proportion q below and $1 - q$ above such that $F(Y_q) = P(Y \leq y) = q$ and $Y_q = F^{-1}(q)$. Another way to express the quantile function is

$$Q_y(q) = F_y^{-1}(q) = \inf\{y/F_y(y) \geq q\}$$

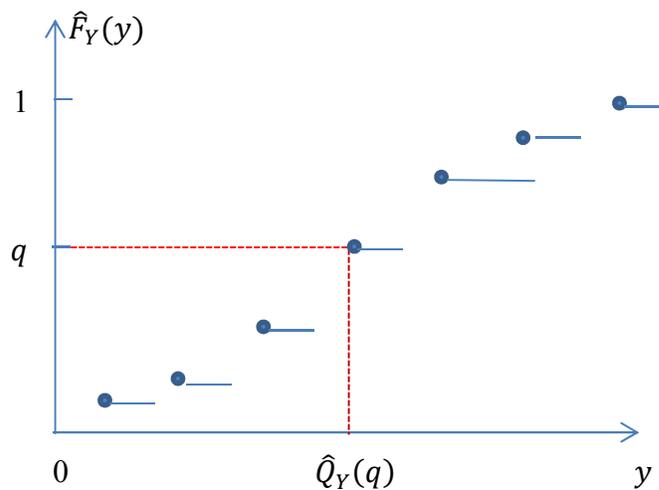
A graph of a typical quantile function is shown in Figure.



The empirical quantile function can be given as

$$\hat{Q}_y(q) = \hat{F}_y^{-1}(q) = \inf \left\{ y / \frac{\#(Y_i \leq y)}{n} \geq q \right\}.$$

A graph of a typical empirical quantile function is shown in Figure.



Sampling Distribution of Sample Quantile:

Suppose y_1, y_2, \dots, y_n , is a large random sample from a distribution function of Q_q and probability density function of $f(y) = \frac{\partial}{\partial y}(F(y))$, then the distribution of \hat{Q}_y is approximately normal with mean Q_q and variance $\frac{q(1-q)}{n} \cdot \frac{1}{f(Q_q)^2}$. The variance of the sampling distribution is completely determined by the probability density evaluated at the quantile.

Standard Error of Estimates in Quantile Regression:

Koenker and Bassett (1982) described the method of estimating standard errors of the Quantile Regression coefficients. However, Rogers (1992), reported this method performs well for homoscedastic distribution of residuals (i.e., the distribution of the residuals has uniform variance), but it appears to underestimate the standard errors when the distribution of the residuals is heteroscedastic (i.e., the distribution of the residuals has non-uniform variance). Efron and Tibshirani (1993) suggested bootstrap methods as an alternative approach to estimate the standard errors of estimates for Quantile Regression.

Quantile Regression Model with Single Independent Predictor:

The q^{th} quantile, $0 < q < 1$ split the response variable Y_i into proportion q below and $1 - q$ above such that $F(Y_q) = q$ and $Y_q = F^{-1}(q)$ for the median $q = 0.5$. Quantile Regression uses an asymmetric weighting system of data points and therefore, all data points are weighted based on their distance from the researcher-specified quantile for that estimation. Consequently, Quantile Regression is not synonymous with fitting a separate OLS regression line at each quantile (Petscher & Logan, 2014; Petscher et al., 2013).

In Quantile Regression the fitted model is

$$Y_i = \beta_{q0} + \beta_{q1}X_i + \varepsilon_{qi}, i = 1, 2, \dots, n,$$

Where ε_{qi} 's are random errors

β_{q0} and β_{q1} are the unknown parameters associated with q^{th} quantile, and $0 < q < 1$.

Recall in the Ordinary Least Square Regression model, the conditional mean is

$$E(Y_i|X_i) = \beta_0 + \beta_1X_i.$$

In contrast, for the corresponding Quantile Regression model, the q^{th} conditional quantile given X_i is specified as,

$$Q_q(Y_i|X_i) = \beta_{q0} + \beta_{q1} X_i.$$

Thus, the q^{th} quantile is determined by the quantile specific parameters β_{q0} and β_{q1} , with a specific predictor value of X_i . Like, Ordinary Least Square Regression the $E(\varepsilon_i) = 0$, in Quantile Regression $Q_q(\varepsilon_{qi}) = 0$. It is to be noted that for different values of the quantile q of interest, the error terms ε_{qi} for fixed i are related. By extending the idea of several equations can be expressed at different quantiles. For example, if the Quantile Regression model specifies the 9th quantiles, the 9 different models yields 9 Quantile Regression coefficients for X_i , one at each of the 9 conditional quantiles, i.e. $\beta_{0.10}, \beta_{0.20}, \dots, \beta_{0.90}$.

In ordinary least square regression, the least squares (LS) method tries to minimize $\sum_{i=1}^n \varepsilon_i^2$, the sum of squared error from the fitted straight line to the observed outcome variable Y_i whereas, in Quantile Regression absolute sum of error from the fitted q^{th} line to the observed outcome variable Y_i , is tried to minimize, i.e. $\sum_{i=1}^n |\varepsilon_{qi}|$ is to minimize.

Estimation of parameters in Quantile Regression with Single Independent Predictor:

Suppose, in a Quantile Regression $\hat{\beta}_{q0}$, and $\hat{\beta}_{q1}$ are the estimates of the corresponding unknown parameters β_{q0} and β_{q1} respectively. A method of the absolute sum of error is used to estimate the parameters by minimizing the sum of absolute errors. The attempt is to minimize $\sum_{i=1}^n |\varepsilon_{qi}|$. The Quantile Regression minimizes the $\sum_{i=1}^n q |\varepsilon_{qi}| + \sum_{i=1}^n (1 - q) |\varepsilon_{qi}|$, a sum that gives the asymmetric penalties $q |\varepsilon_{qi}|$ for under prediction and $(1 - q) |\varepsilon_{qi}|$ for over prediction. For example, in a median regression, if $q = 0.5$ then the quantity $\sum_{i=1}^n |\varepsilon_{qi}|$ will collapse to a median regression.

In order to find the quantile regression coefficients, a criterion function is defined for q^{th} Quantile Regression estimator $\hat{\beta}_{q1}$ that minimizes $Q(\beta_q)$ with objective function along with penalty q when response variable is higher than the predicted values i.e. $Y_i \geq X_i \beta$ and penalty $1 - q$ when response variable is higher than the predicted values i.e. $Y_i < X_i \beta$.

$$Q(\beta_q) = \sum_{i=1}^n q|\varepsilon_{qi}| + \sum_{i=1}^n (1-q)|\varepsilon_{qi}|$$

$$Q(\beta_q) = \sum_{i:Y_i \geq X_i\beta} q|Y_i - \beta_{q0} - \beta_{q1}X_i| + \sum_{i:Y_i < X_i\beta} (1-q)|Y_i - \beta_{q0} - \beta_{q1}X_i|,$$

Where $\varepsilon_{qi} = Y_i - \beta_{q0} - \beta_{q1}X_i$ equivalently, it can be written as

$$Q(\beta_q) = \sum_{i=1}^n [I_{\{Y_i \geq X_i\beta\}} q|Y_i - \beta_{q0} - \beta_{q1}X_i| + I_{\{Y_i < X_i\beta\}} (1-q)|Y_i - \beta_{q0} - \beta_{q1}X_i|],$$

Where $0 < q < 1$ and I is an indicator function.

In contrast to ordinary least square regression or maximum likelihood, the Quantile Regression computational implementation uses linear programming method to find the regression coefficients.

Quantile Regression Model with Multiple Independent Predictors:

Given that, $q_1, q_2, \dots, q_k \in (0,1)$, are the quantiles of a response variable (Y_i) with k independent predictors ($X_{i1}, X_{i2}, \dots, X_{ik}$), then a multiple Quantile Regression model with k predictors can be written as

$$Q_{qk}(Y_i|X_{ki}) = X\beta_{qk} + \xi,$$

where, Y is an $(n \times 1)$ matrix of the response variable, X is an $n \times (k + 1)$ matrix of the predictor variables, β_{qk} is a $(k \times 1)$ matrix of the regression coefficients at q_k th quantile level and ξ is a $(n \times 1)$ matrix of the error.

Interpretation of Quantile Regression Estimates with Multiple Independent Predictors:

In a multiple Quantile Regression, $\hat{\beta}_{qj}$ estimates the change in specific quartile q of the response variable Y_i produces by one unit change in the predictor variable X_j , keeping all other X_{k-j} predictors constant. For example, $\hat{\beta}_{q1}$ is the estimated rate of change in specific quartile q in a response variable Y_i for a one unit change in predictor X_1 , keeping all other X_{k-1} predictors constant.

Again, it is to be noted that unlike to the Ordinary Least Square Regression, the interpretation of Quantile Regression results needs to specify which quantile of response variable are used.

Reference and comparison in Quantile Regression Estimates:

Again, when a predictor variable in Quantile Regression is categorical, and to facilitate the interpretation of the categorical Quantile Regression estimates, the notation of reference and comparison with some ideas related to the quantification of effects can be used. For example, in a dichotomous categorical variable, one category as reference can be used to compare other category to study the effect of a unit change in response variable, this idea can be extended to more than categories in a categorical predictor. In Quantile Regression, the fitted categorical coefficient can be interpreted as an estimated effect i.e. estimates of the change in the q^{th} quantile of the response variable that results from a 1-unit change between reference category and comparison category.

Goodness of Fit in Quantile Regression:

Koenker and Machado (1999) suggest measuring goodness of fit by comparing the sum of the weighted absolute deviations when the explanatory variables are not used in the prediction of the q^{th} quantile (i.e. the reduced model) with the sum of the weighted absolute deviations when the explanatory variables are used in the prediction of the q^{th} quantile (i.e. the full model).

Suppose, $\hat{V}(q)$ is the sum of the weighted absolute deviations when the explanatory variables are used in the prediction of the q^{th} quantile (i.e. the full model) and $\tilde{V}(q)$ is the sum of the weighted absolute deviations when the explanatory variables are not used in the prediction of the q^{th} quantile (i.e. the reduced model), then goodness of fit for Quantile Regression defined as

$$R(q) = 1 - \frac{\tilde{V}(q)}{\hat{V}(q)}.$$

The value of $R(q)$ varies from 0 to 1 and is free from the unit of measurement, with larger $R(q)$ indicating a better model fit.

Test of Significance in Quantile Regression:

There are mainly two types of significance that are important in Quantile Regression coefficients.

- Test of significance about the quantile coefficients different from zero.
- Test of significance based on the quantile coefficients different from OLS coefficients, showing the significant effect along the distribution of the response variable.

To test the significance about the regression coefficient of the Quantile Regression line $H_0: \beta_q = 0$ against $H_0: \beta_q \neq 0$, there are normally three kinds of tests are used, the likelihood ratio (LR) tests, Wald test (W) and Lagrange multiplier (LM). Koenker and Machado (1999) recommended the asymptotic test statistics called likelihood ratio (LR) test for the contribution of response variable to the prediction of the q^{th} quantile as,

$$LR = L_n(q) = \frac{2(\tilde{V}(q) - \hat{V}(q))}{q(1-q)[f_{eq}(0)]^{-1}},$$

where, $\tilde{V}(q)$ is the sum of the weighted absolute deviations when the explanatory variables are not used in the prediction of the q^{th} quantile (i.e. the reduced model).

$$\tilde{V}(q) = \sum_i |Y_i - \beta_0 + \beta_1 X_{1i} + \dots + \beta_{q-1} X_{q-ki}|,$$

$\hat{V}(q)$, is the sum of the weighted absolute deviations when the explanatory variables are used in the prediction of the q^{th} quantile (i.e. the full model).

$$\hat{V}(q) = \sum_i |Y_i - \beta_0 + \beta_1 X_{1i} + \dots + \beta_q X_{qi}|$$

The term $s(q)$ represents the sparsity function which measures the density of observations near the quantile of interest. The test statistics LR is asymptotically distributed as a χ^2_q , where q is the difference in the number of predictors included in the full model and the number of predictors included in the reduced model.

Suppose four independent predictors were selected, then a multiple Quantile Regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i.$$

After fitting OLS and Quantile Regression at different quantiles suppose X_2 and X_3 are not statistically significant in any of the fitted equations, and it is relevant to test the null hypothesis

$$H_0 = \beta_2 = \beta_3 = 0.$$

This hypothesis involves a reduced model. The test function for LR test can be defined as

$$LR = \frac{2(\tilde{V}(q) - \hat{V}(q))}{q(1-q)[f_{eq}(0)]^{-1}}.$$

At the median, to test the null hypothesis $H_0 = \beta_2 = \beta_3 = 0$, the two-objective function with restricted and unrestricted models can be define as

$$\tilde{V}(q) = \sum_i |Y_i - \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i}|$$

$$\hat{V}(q) = \sum_i |Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}|.$$

After the estimation process, the value of LR can be estimated and compared with the critical value of χ^2 with 2 degrees of freedom at 5% level of significance. A decision can be made whether to reject the null hypothesis. If so, it means the restricted and unrestricted models yield similar results, and drop X_2 and X_3 safely.

Koenker and Basset (1982) explained another test to test the significance of more than on Quantile Regression coefficient at a time is LM test. It considers the gradient g which is a function of the sign of the errors, of their position above and below the Quantile Regression line excluding the variables under test. The test function is defined as

$$LM = g^T [D_{kk}]^{-1} g.$$

Weiss (1990) suggested the LM test can be implemented by estimating an auxiliary regression. The residuals of the reduced model become the dependent variable of the additional regres-

sion having an explanatory variable those predictors excluded from the model. The term nR^2 is asymptotically χ^2 with degree of freedom equal to the number of predictors under test.

Consider four independent predictors were selected. The multiple Quantile Regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i.$$

The objective function of median regression is given by

$$\sum_i \rho_q(e_i) = 0.5 \sum_i |e_i|.$$

It considers the gradient g at median which is a function of the sign of the errors, of their position above and below the Quantile Regression line. Under the null hypothesis $H_0 = \beta_2 = \beta_3 = 0$, the residuals of the constraint model estimated at median are given by

$$\tilde{e}_i(q) = Y_i - \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i},$$

and is the dependent variable in the auxiliary equation:

$$\tilde{e}_i(q) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{4i} + \lambda_i.$$

nR^2 is estimated and compared with χ^2 with 2 degree of freedom at 5% level of significance under test.

Finally, Koenker and Basset (1982b) suggested the Wald test, denoted by W , considers estimates of the model including the predictors under the tests. The test function is defined as

$$\begin{aligned} W &= n \left[q(1-q) [f_{eq}(0)]^{-1} \right]^2 \hat{\beta}(q)^T [D^{kk}]^{-1} \hat{\beta}(q) \\ &= n [q(1-q) f_{eq}(0)]^{-2} \hat{\beta}(q)^T [D^{kk}]^{-1} \hat{\beta}(q). \end{aligned}$$

Consider four independent predictors were taken. A multiple Quantile Regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i.$$

To test the null hypothesis $H_0 = \beta_2 = \beta_3 = 0$, the estimated median regression coefficients under the test is given by

$$\hat{\beta}(0.5) = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}, \quad \text{matrix } D_{22} = \begin{pmatrix} \sum X_{2i}^2 & \sum X_{2i}X_{3i} \\ \sum X_{3i}X_{2i} & \sum X_{3i}^2 \end{pmatrix} \quad \text{and } q(1-q)f_{eq}(0) = 0.25(.25), \quad \text{where } f_{e0.5}(0) = 0.25 \text{ at } q = 0.5.$$

The function W in equation 27 can be estimated and compared with χ^2 with 2 degree of freedom at 5% level of significance under test. Again, a decision can be made whether to reject the null hypothesis. If it is rejected, it means the restricted and unrestricted models yield similar results, and drop X_2 and X_3 safely.

In testing the significance of the quantile coefficients different from OLS coefficients which showing the significant effects along the distribution of the response variable. It may have lower effect for lower quantile and higher effect of higher quantile or the reverse; this differential effect across the quantiles can be showed and may be important that the quantiles coefficients differ from the OLS coefficients.

Interval Estimation in Quantile Regression:

In Quantile Regression, the confidence interval for β_q can be estimated by using three different methods named as sparsity, rank, and resampling.

Sparsity:

The sparsity method is the fastest, but it involves estimation of the sparsity function, which is not robust for data that are not independently and normally distributed.

Rank Method:

The rank method, which computes confidence intervals by inverting the rank score test, does not suffer from the problem of data that are not independently and normally distributed, but it uses the simplex algorithm and is computationally expensive with large data sets.

Resampling (Bootstrapping):

Resampling or bootstrapping may give two alternatives to make inferences about parameters. The first alternative is counting standard deviation of parameters and using it to obtain a t-value and its p-value of related parameters. Confidence intervals (CI) can be approximated using this method. The second alternative is by constructing 95% CI (or other CIs) using 97.5th percentile and 2.5th percentile of the samples of bootstrap estimates. If the CI captured the parameter, an inference can be made that the parameter is significant at $\alpha = .05$ (Hao and Naiman, 2007).

Robustness and Regression Modeling:

Robustness is a degree to which a statistical test maintain Type I error and Type II errors rates in the presence of the violation of assumptions relates to the outcome variable Y_i . As mention earlier, OLS does not fit well when there are some regular outliers present in the response variable, or if the data were sampled from a non-normal distribution. Statistical methods in regression are generally considered resistant to the outliers and non-normality of the distribution.

In contrast, the Quantile Regression model estimates are not sensitive to outliers also they are robust to the distributional assumptions because the regression coefficients weigh the local behavior of the distribution near the specific quantile more than the remote behavior of the distribution. Robust method invoked in order to refine the conditional mean and heterogeneity of the variance. Robust methods are not only sensitive to the non-normality of the data, but also minimize the effect of assumptions about data below detection limits, and the effect of outliers on the determination of relations between variables (Helsel & Hirsch, 2002). The first step toward a robust regression estimator came from Edgeworth (1887), improving a proposal of Boscovich. Kendall (1938) provided a non-parametric method of detecting a relationship between two variables and to find a suitable fit when there is a problem of outliers in the data, it provides different option to examine

lines between all pairs of points, and estimate the slope by the median of all slopes, and intercept by the median of all intercepts.

As mentioned, in a simple linear regression, Theil (1950) first proposed another robust linear regression method where there are one response and one predictor variable and is robust to outliers in the response variable. In this method, the slope of the regression line is estimated as the median of all pairwise slopes between each pair of points in the dataset. Sen (1968) extended this estimator to handle ties. The Theil-Sen estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses a high asymptotic efficiency. A modified and preferred method is named after Siegel (1982), who proposed the repeated median with a 50% breakdown point. Indeed, 50% is the best that can be expected (for larger amounts of contamination, it becomes impossible to distinguish between the good and the bad parts of the sample). When the dependent variable is continuous, the Theil–Sen estimator enjoys good theoretical properties and it performs well in simulations in terms of power and Type I error probabilities when testing hypotheses about the slope (e.g., Wilcox, 2012b).

Kendall–Theil Sen Siegel Regression

Kendall–Theil (1950) regression is another completely nonparametric approach to linear regression with one predictor and one response variable. The Theil estimator provides a robust estimator for linear regression and outliers in the response variable. When the estimator is a line, then the Ordinary Least Square estimate corresponds with the mean, and is not robust estimate. A single point can easily affect the slope of the line. The Theil estimator is a robust version of a linear regression. It simply computes all the lines between each pair of points and uses the median of the slopes of these lines. This procedure is sometimes called Theil–Sen procedure.

A modified, and more robust, method is named after Siegel. The method yields a slope and intercepts for the fit line, and a p-value for the slope can be determined as well. Typically, no

measure analogous to r-squared is reported. Theil-Sen single (1950) median method computes slopes of lines crossing all possible pairs of points, when x coordinates differ. After calculating these $n(n-1)/2$ slopes (these values are true only if X_i is distinct), the median of them is taken as slope estimator. Next, the intercepts of n lines, crossing each point and having calculated slope are calculated. The median from them is intercept estimator.

Sen (1968) extended this estimator to handle ties and obtained unbiasedness and asymptotic normality of the estimator for absolutely continuous error distribution and a no identical covariate.

A variability of the Theil-Sen estimator due to Siegel (1982) determines, for each sample point, the median m_i of the slopes of lines through that point, and then determines the overall estimator as the median of these medians. These repeated medians are sometime more complicated. For each point, the slopes between it and the others are calculated (resulting $(n-1)$ slopes) and the median is taken. This results in n medians and median from these medians is slope estimator. Intercept is calculated in similar way. The breakdown point of Theil-Sen method is about 29% and, Siegel extended it to 50%, so these regression methods are very robust. Additionally, if the errors are normally distributed and no outliers are present, the estimators are very similar to classic least squares.

Estimation of parameters Kendall–Theil Sen Siegel Regression with Single Independent Predictor:

In the Kendall-Theil Sen Regression, suppose a sample of n pair of observation (X_i, Y_i) , $i = 1, 2, \dots, n$ were taken. Suppose a straight line given below is used as a best fit model to the given set of data.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, 2, \dots, n$$

where, Y_i is the response variable for each data point (i), X_i is the predictor variable for each data point (i), ε_i is the residual in prediction of Y for each data point (i), $\hat{\beta}_1$ is the estimated regression coefficient, $\hat{\beta}_0$ is the estimated Y intercept, and n is the number of XY data points in the sample.

In Kendall–Theil Sen Regression, the regression coefficient β_1 can be estimated by the median of all pairwise slopes between each pair of points in the given data set (Theil, 1950; Sen, 1968; Helsel and Hirsch, 2002). Each regression coefficient passing through (X_i, Y_i) and (X_j, Y_j) , the data point can be estimated by the

$$\hat{\beta}_1 = \text{Median}_j \left\{ \frac{(Y_j - Y_i)}{(X_j - X_i)}, \text{ for } i = 1, 2, \dots (n - 1), X_j \neq X_i \text{ and } j = 2, \dots n \right\}$$

The number of possible regression coefficients between data pairs can be calculated by

$$N_p = \frac{n(n-1)}{2}.$$

All possible estimated b_{ij} are sorted and ranked by ascending order. Sorting is a computationally intensive process because each slope estimate in the array of slopes must be compared to other values and put in the proper order. If N_p is an odd number, the median slope is selected as the middle value of the array otherwise, the median is calculated as the arithmetic average of the two center points.

The Y-intercept of the line can be estimate by the equation used by Conover (1980) as

$$\hat{\beta}_0 = \{\tilde{Y} - \hat{\beta}_1 \tilde{X}\}$$

Where, $\hat{\beta}_0$ is the estimated Y-intercept, \tilde{Y} is the median of the response variable, $\hat{\beta}_1$ is the estimated slope, and \tilde{X} is the median of the predictor variable. The error tem ε_i are the random errors, should be independently and normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$. However, In Kendall-Theil regression model these assumptions associated with error term are not bounded to fulfill.

If N_p is an odd number, the median regression coefficient is selected as the middle value of the array, otherwise, median regression coefficient is selected by taking the mean of the two middle values of an array. Hence the estimated Theil Sen Regression line is

$$TS: \hat{Y}_{TS} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Siegel (1982) considered repeated medians. For each observation (X_i, Y_i) , the regression coefficients between it and the others $(n - 1)$ are calculated and the median is taken. This results in n medians and median from these medians is a regression coefficients estimator. A robust estimator $\tilde{\beta}_n$ of the regression coefficient $\hat{\beta}_1$ can be estimated by taking the medians of these least square estimates i.e.

$$\tilde{\beta}_n = \text{Median}_i \left\{ \hat{\beta}_1 = \frac{(Y_j - Y_i)}{(X_j - X_i)} : X_i \neq X_j, 1 \leq i < j \leq n \right\}.$$

Similarly, the y-intercept can be estimated by the medians of all possible least square estimates

$$\tilde{\alpha}_n = \text{Median}_i \left\{ \hat{\beta}_0 = \frac{(Y_j X_i - Y_i X_j)}{(X_j - X_i)} : X_i \neq X_j, 1 \leq i < j \leq n \right\}.$$

Hence the estimated Theil Sen Siegel Regression line is

$$TSS: \hat{Y}_{TSS} = \tilde{\alpha}_n + \tilde{\beta}_n X_i$$

Interval Estimation in Kendall–Theil Sen Regression with Single Independent Predictor

Three methods were considered to estimate confidence interval for Theil Sen Regression coefficient β_1 . They are,

Theil Sen method to handle ties: Confidence interval for Theil regression coefficient β_1 can be calculated based on ordering the b_{ij} values such that $b_{(1)} \leq b_{(2)} \dots \leq b_{(N_p)}$. Let $R_l = (N_p - w)/2$ and $R_u = (N_p + w)/2 + 1$, where w is the $1 - \alpha/2$ quantile of Kendall's statistics and is given by $w =$

$$Z_{\alpha/2} \left(\sqrt{\frac{n(n-1)(2n+5)}{18}} \right). \text{ So,}$$

A $100(1 - \alpha)$ % confidence interval for Theil Sen Regression coefficient β_1 can be calculated by using large sample approximation equations describe by Helsel and Hirsch (2002). The lower R_l and upper R_u confidence limits can be estimated as,

$$R_l = \left[\frac{N_p - Z_{\alpha/2} \left(\sqrt{\frac{n(n-1)(2n+5)}{18}} \right)}{2} \right]$$

and

$$R_u = \left[\frac{N_p + Z_{\alpha/2} \left(\sqrt{\frac{n(n-1)(2n+5)}{18}} \right)}{2} \right] + 1$$

where, R_l is the lower rank order of the regression coefficient, R_u is the upper-rank order of the regression coefficient, N_p is the number of pairwise slopes calculated from $N_p = \frac{n(n-1)}{2}$, Z is the table value taken from a standard normal table.

Percentile bootstrap method:

In this method, let (Y_i^*, X_i^*) , $i = 1, 2, \dots, n$ bootstrap sample obtained by randomly sampling, with replacement, of n pair of observation (X_i, Y_i) : $i = 1, 2, \dots, n$. Label the resulting estimates of the slope as b^* and repeat this process B times resulting $b_{(1)}^*, b_{(2)}^*, \dots, b_{(B)}^*$.

A $100(1 - \alpha)$ % confidence interval for Theil Sen Regression coefficient β_1 given by (Wilcox, 1998) is $(b_{(L)}^*, b_{(U)}^*)$ where $L = \alpha B / 2$ and $U = (1 - \frac{\alpha}{2})B$ and $b_{(1)}^* \leq b_{(2)}^* \dots \leq b_{(B)}^*$ are the B bootstrap values written in ascending order.

Bootstrap estimated method:

This method uses the bootstrap estimate of $var(b)$, say $\hat{\sigma}_b^2$ based on $B=200$ bootstrap samples.

Then,

A $100(1 - \alpha) \%$ confidence interval for Theil Sen Regression coefficient β_1 given by (Efron, 1987) is $(b \pm z_{\alpha/2} \hat{\sigma}_b^2)$.

Root Mean Square Error:

The RMSE, also known as standard error of the regression or standard deviation of residuals, indicates lack of precision (spread) in the population of residual errors (Helsel and Hirsch, 2002). The RMSE is calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n - 2}}$$

Test of Significance in Kendall–Theil Sen Regression with Single Independent Predictor:

Sen (1968), Helsel & Hirsch (2002) provided a method for testing a hypothesis about the regression coefficient. In Testing the hypothesis $H_0: \tilde{\beta}_n = 0$ for Kendall–Theil Sen Regression can be done by testing the hypothesis $H_0: \tau = 0$, where τ is the Kendall rank correlation coefficient. The τ can be expressed as follows

$$\tau = \frac{2S}{n(n-1)},$$

where, S is the Kendall's S statistic and is given by

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(X_j - X_i),$$

and,

$$\text{sgn}(x) = \begin{cases} +1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$$

z testing $H_0: \tau = 0$ is equivalent to $H_0: S = 0$ (Helsel & Hirsch, 2002).

Estimation of parameters Kendall–Theil Sen Regression with Multiple Independent Predictors

Suppose a response variable (Y_i) with p independent predictors ($X_{i1}, X_{i2}, \dots, X_{ip}$) were selected. Xin Dang et al., (2009) proposed after following the above procedure defined in Kendall-Theil Sen Regression model, first an estimator of $\theta = (\beta_0, \beta_1^t)^t$ can be found as the solution to the $p + 1$ equations

$$Y_i = \beta_0 + \beta_1 X_i^t, \quad i \in K_{p+1} = \{i_1, \dots, i_{p+1}\},$$

Where K_{p+1} is the $(p + 1)$ -subset of $\{1, \dots, n\}$ such that $(p + 1) * (p + 1)$ matrix $(X_k: k \in K_{p+1})$ is invertible. To stress the dependence on the $p + 1$ observations, they denote this estimator by $\widehat{\theta}_{K_{p+1}}$.

Then a natural extension of the Theil-Sen estimator from a simple linear regression to a multiple linear regression is the multivariate median

$$\widehat{\theta}_n = Mmed\{\widehat{\theta}_{K_{p+1}}: \forall K_{p+1}\}.$$

This $\widehat{\theta}_{K_{p+1}}$ is also the least square estimator of θ based on $p + 1$ observation $\{(X_i, Y_i) : i \in K_{p+1}\}$.

From this point of view and slightly more generally, an arbitrary combination of m distinct observations $\{(X_i, Y_i) : i \in K_m\}$, where $p + 1 \leq m \leq n$ may be chosen to construct a least squares estimator $\widehat{\theta}_{K_m}$. Then a multiple Theil-Sen estimator $\widehat{\theta}_n$ of the parameter θ is naturally defined to be the multivariate median of all possible least square estimators:

$$\widehat{\theta}_n = Mmed\{\widehat{\theta}_{K_m}: \forall K_m\}$$

CHAPTER 3 METHODOLOGY

Monte Carlo simulation technique will be used to generate from randomly sampled observation with replacement from different distributions and estimate of regression coefficients, standard errors, median absolute deviation, p-values, confidence intervals and test of significance will be calculated, based on Ordinary Least Square Regression, Quantile Regression, Theil Sen Regression and Theil Sen Siegel Regression. A comparison will be made using these four regression methods. The author will write several essential codes in R in order to compare regression coefficient, confidence intervals, test of significance and to generate different figures.

Three theoretical distributions and eight empirical distributions identified by Micceri (1989) will be randomly sampled. The simulations will be run on a Dell PC with an Intel (R) Core (TM) i5-4590 CPU processor.

Procedure:

Observations for the Monte Carlo simulations will be randomly generated with replacement from the Normal, Uniform, and Poisson distributions using statistical software R. Similarly, observations for the Monte Carlo simulations will be randomly generated with different sample sizes in the presence of 10% and 20%, 30% and 50% outliers. Values of regression coefficients, confidence intervals and test of significance will be obtained and tested by fitting four regression models to the simulated data. Parametric values of regression coefficients were set at certain values to generate response variable Y_i .

Specific Procedure:

Observations will be randomly sampled with replacement from the Normal, Uniform, and Poisson distributions. Similarly, observations will be randomly sampled with replacement from the Micceri family distributions. After each sample has been generated, the regression coefficients, confidence intervals and test of significance will be constructed based on Ordinary Least Square

Regression, the Quantile Regression and the Theil Sen Siegel Regression. A comparison based on Biasedness through mean and median, Standard Deviation (S.D), Standard Errors (S.E), Median Absolute Error (MEDAE), Root Mean Square Error (RMSE), Relative Root Mean Square Error (RMSE), and Relative Median Absolute Error (RMEDAE) of the four regression methods will be used to evaluate the model fit. A negative value of Relative Root Mean Square Error (RRMSE) refer to a proportional increase in RMSE of β_1 obtained by other regression model, on the other hand positive value of Relative Root Mean Square Error (RRMSE) indicates a proportional decrease in RMSE of β_1 obtained other regression model (Syed et al., 2016). This procedure will be repeated some hundred thousand times for different sample sizes.

Selected Distributions:

In statistics there are several statistical distributions that could have been selected for this study; however, only eleven of them are selected and the observations are randomly generated from these distributions with replacement. They are the Normal distribution, a Uniform distribution, a Poisson distribution, and eight Micceri family distributions.

The Normal Distribution:

If X is a continuous random variable with mean μ and variance σ^2 , then a normal distribution is defined by the *pdf* as (Chaudhary and Kamal, 2000)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty \leq x \leq +\infty$$

Where,

μ = Mean of normal distribution

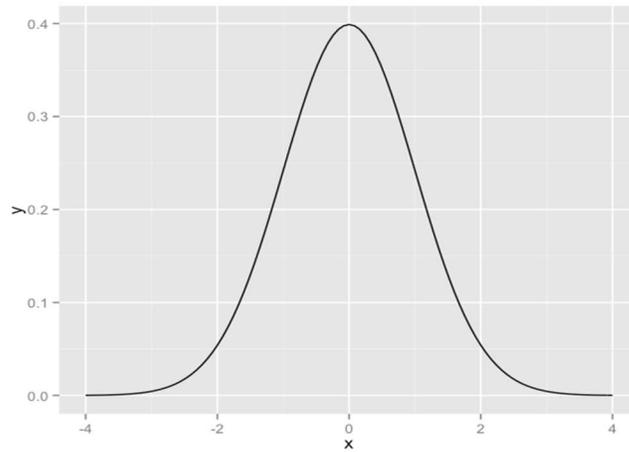
σ = Standard deviation of the normal distribution

π = A constant having value 3.1414

e = A constant having value 2.718

$x = \text{Value of continuous random variable}$

The shape of normal distribution is given in Figure below:

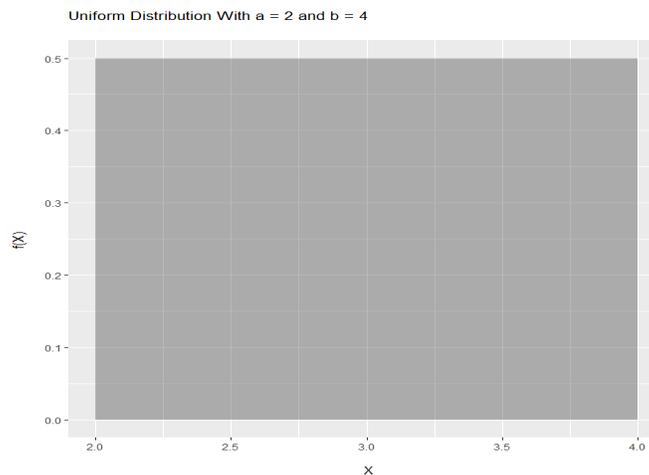


The Uniform Distribution:

If X is a continuous random variable over the interval $[a, b]$, with mean $\mu = \frac{a+b}{2}$ and variance $\sigma^2 = \frac{(b-a)^2}{12}$, then a uniform distribution is defined by the *pdf* as (Chaudhary and Kamal, 2000)

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

The shape of uniform distribution is given in Figure below.



The Poisson Distribution:

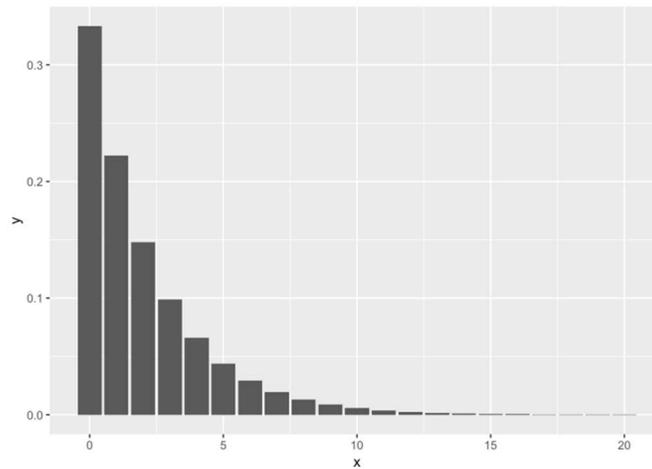
Suppose that we can expect some independent event to occur λ time over a specified period of time or space. Let a discrete random variable X be the occurrence of event, we call it a Poisson random variable. Then, its probability function is:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0, 1, 2, \dots$$

Where,

λ = average(or mean) number of events in a given time t, e = a constant = 2.718

The shape of exponential distribution is given in Firure below.

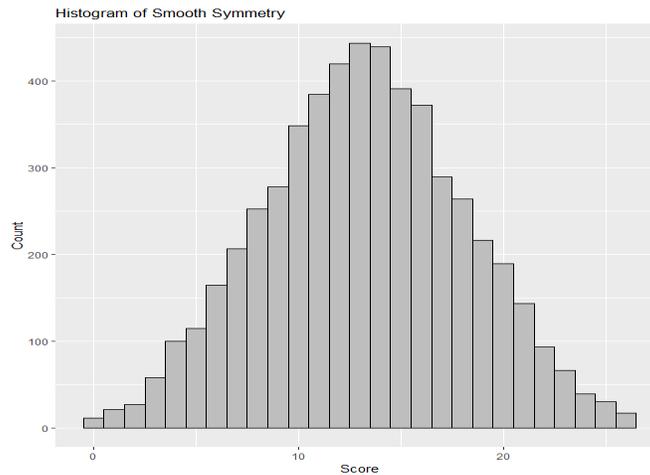


The Micceri Data Sets:

Random samples will be generated from eight of the empirical distributions identified by Micceri (1989). The large sample distributions were categorized into general achievement/ability tests, criterion/mastery tests, psychometric measures, pre-test measures, and post-test measures. It included 265 distributions that were derived from journal articles, 30 from national tests, 64 from statewide tests, 65 from district-wide tests, and 17 from college entrance and GRE tests (Lawson, 2006). A brief description of the eight distributions with mean, median, and variance are summarized below.

Smooth Symmetric Distribution:

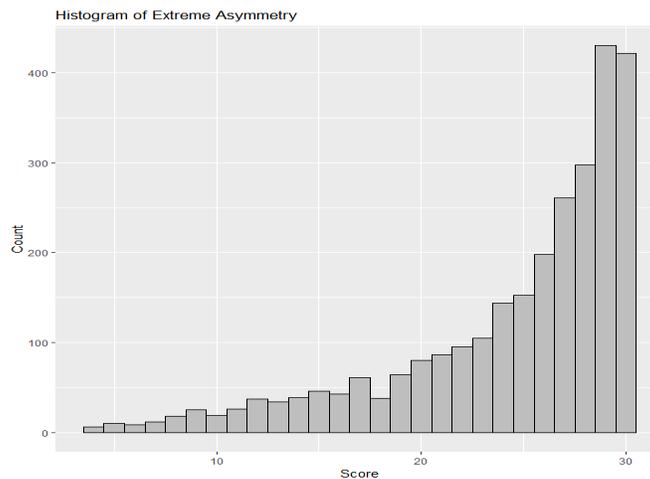
The smooth symmetric distribution consists of achievement observations with a light skew. This distribution has a mean = 13.91, median = 13, and variance = 24.11 (Sawilowsky, 1992). The shape of smooth symmetric distribution is given in figure below.



Extreme Asymmetric Distribution:

The extreme asymmetric distribution consists of achievement observations with a fairly large skew. This distribution has a mean = 24.5, median = 27, and variance = 33.53 (Sawilowsky, 1992).

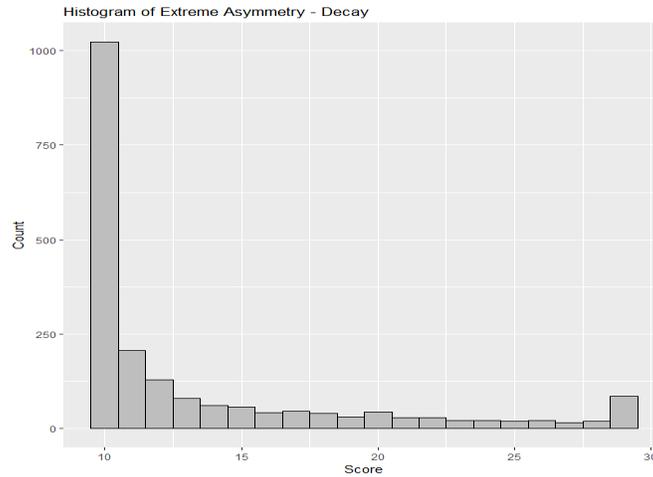
The shape of extreme asymmetric distribution is given in figure below.



Extreme Asymmetric-Decay Distribution:

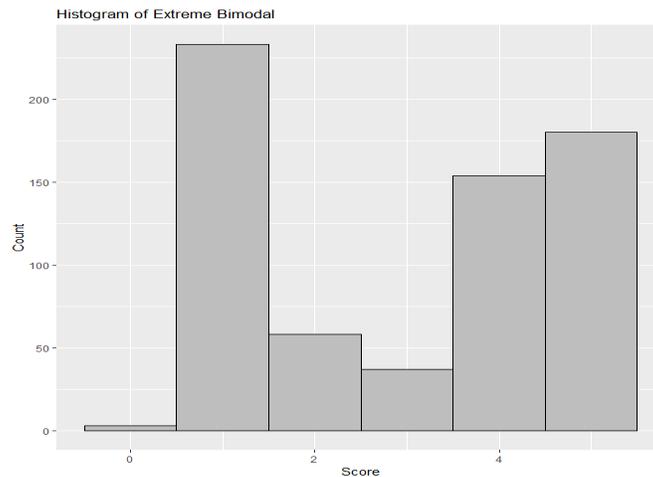
The extreme asymmetric distribution consists of achievement observations with a fairly large skew on right hand side. This distribution has a mean = 13.67, median = 11, and variance = 33.06 (Sawilowsky, 1992).

The shape of extreme asymmetric distribution is given in given below



Extreme Bimodal Distribution:

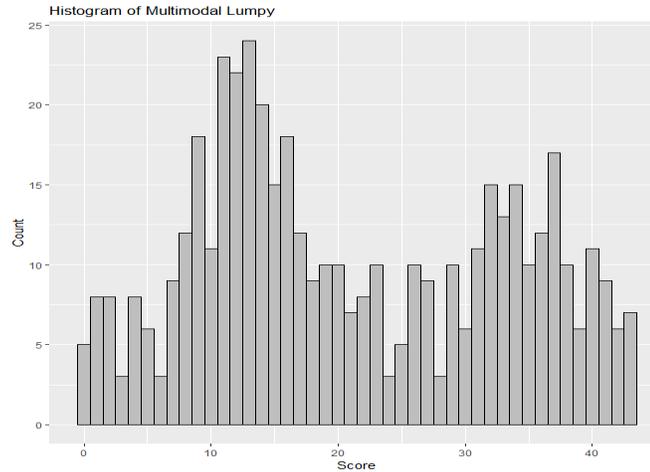
The extreme bimodal distribution consists of psychometric observations with a mean = 2.97, median = 4, and variance = 2.86 (Sawilowsky, 1992). The shape of extreme bimodal distribution is given in Figure below.



Multi-modal Lumpy Distribution:

The multi-modal and lumpy distribution consists of achievement observations with a mean = 21.15, median = 18, and variance = 141.61(Sawilowsky, 1992).

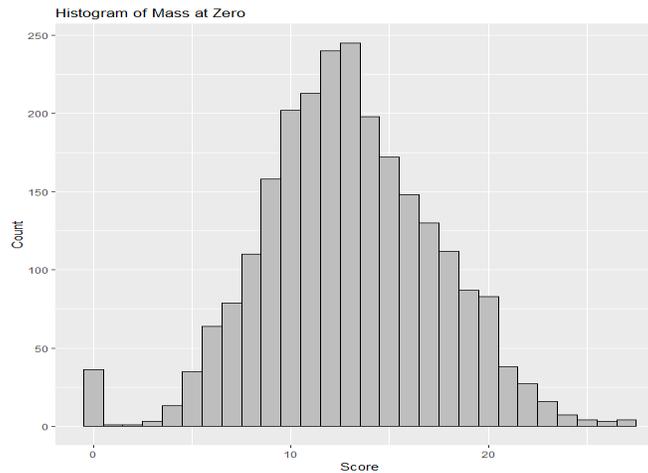
The shape of multi-modal lumpy distribution is given in Firure below.



Mass at Zero Distribution:

The Mass at Zero distribution consists of achievement observations with a mean = 12.92, median = 13, and variance = 19.54 (Sawilowsky, 1992).

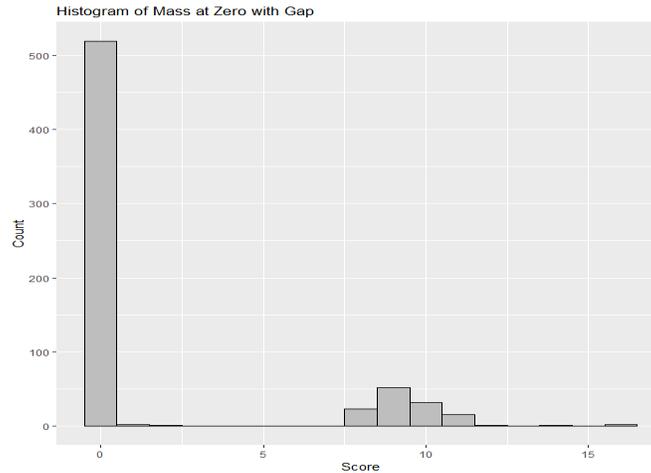
The shape of multi-modal lumpy distribution is given in Firure below.



Mass at Zero with Gap Distribution:

The Mass at Zero distribution consists of achievement observations with a mean = 1.85, median = 0.00, and variance = 14.44 (Sawilowsky, 1992).

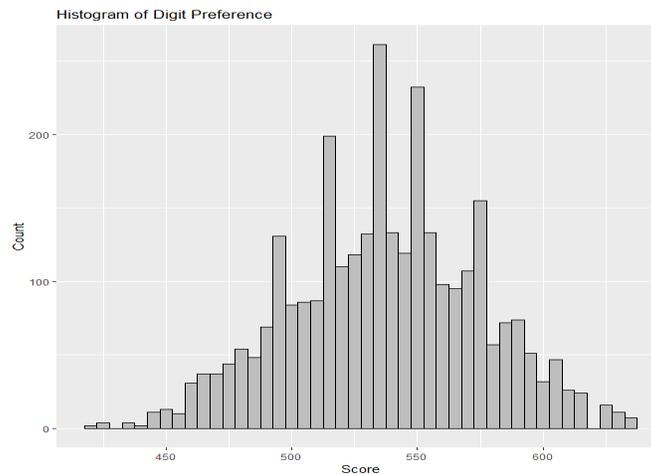
The shape of multi-modal lumpy distribution is given in Firure below.



Digit Preference Distribution:

The Mass at Zero distribution consists of achievement observations with a mean = 536.95, median = 535.00, and variance = 1416.77 (Sawilowsky, 1992).

The shape of multi-modal lumpy distribution is given in Firure below.



CHAPTER 4 RESULTS

The Monte Carlo technique was used to estimate the regression coefficients, standard errors, median absolute deviation, p-values, confidence intervals and test of significance based on Ordinary Least Square Regression, the Quantile Regression, the Theil Sen Regression and Theil Sen Siegel Regression. A visual as well as numerical comparison was made using these four regression methods. For visual comparison scatter plots with fitted regression lines using all four regression procedures were used. For numerical comparison, standard errors, median absolute deviation, confidence intervals, mean bias, median bias, root mean square error (RMSE), median absolute error (MEDAE), relative mean square error and relative median absolute error were used.

Observations for the Monte Carlo simulations were randomly generated with replacement and the process was repeated 1000 times to generate independent sample of size n for an outcome variable Y and a predictor variable X . The sample sizes studied were $n=10, 30, 50$ and 100 . In order to study the effect of various situations of the regression coefficients and robustness of the selected regression models, samples were classified in to different cases as.

Regression Model with slope and intercept under the Normality Assumption with no Outliers:

If the errors (e_i) are independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a uniform distribution with $\min=0$ and $\max=1$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$.

Regression Model with slope, intercept and dichotomous predictor variable with no Outliers:

If the errors (e_i) are independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a binomial distribution with a single trial and $p = 0.5$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$.

Regression Model under the Normality Assumption with Outliers in both X and Y direction:

If the errors (e_i) are independent and normally, then a random sample of size n was generated from a bivariate normal distribution with mean (0, 0) and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 30% 50% and 100% of n was generated in both X and Y variables from a bivariate normal distribution with means (2, 6) with variances $0.1 \times$ variance of the above bivariate normal distribution, i.e. the variances (0.1, 0.1).

Regression Model under the Normality Assumption with Outliers in Y direction only:

If the errors (e_i) were independent and normally, then a random sample of size n was generated from a bivariate normal distribution with mean (0, 0) and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 30% 50% and 100% of n was introduced in Y variable only from a bivariate normal distribution with means (0, 6) with variances $0.1 \times$ variance of the above bivariate normal distribution, i.e. the variances (0.1, 0.1).

Regression Model under the Non-Normality Assumption:

If an outcome variable Y is non- normally distributed, a random sample of size n was generated from non-normal distribution using a log link function for a predictor variable X. We assumed $Y \sim \text{Poisson}(\lambda)$, and $\log(\lambda) = 1 + 0.2 * X$. It was assumed X is uniformly distributed with $\min=0$ and $\max=1$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression Model under the Micceri distributions:

Eight Micceri distributions were used to generate a random sample for an outcome variable Y, with a uniform distribution of predictor variable X with $\min=0$ and $\max=1$.

Finally, these Micceri distributions were used to generate a random sample for an outcome variable Y, with a binomial distribution of predictor variable X.

For simulation, an open source software program R version 3.5.1 was downloaded from the CRAN website link and several codes were written with package include “mblm”, “quantreg”, and

“MASS”. The packages were then verified for their accuracy and reliability with some textbook examples.

Regression Model passing through origin under the Normality Assumption with no Outliers:

If the errors (e_i) were independent and normally distributed, then a random sample of size n from a bivariate normal distribution was generated with mean $(0, 0)$, variances equal to 1, and a correlation coefficient equal to 0.80.

Table 1: Descriptive Statistics of (X, Y) for Regression Model passing through origin under the Normality Assumption with no Outliers:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	-1.05	1.56	-0.24	-0.40	0.77	0.68
X	10	-1.61	0.98	-0.26	-0.47	0.83	1.06
Y	30	-1.74	2.44	0.19	0.37	0.95	1.13
X	30	-1.79	2.19	0.20	0.03	0.99	1.30
Y	50	-1.74	2.44	0.19	0.37	0.95	1.13
X	50	-1.79	2.19	0.20	0.03	0.99	1.30
Y	100	-1.74	2.44	0.19	0.37	0.95	1.13
X	100	-1.79	2.19	0.20	0.03	0.99	1.30

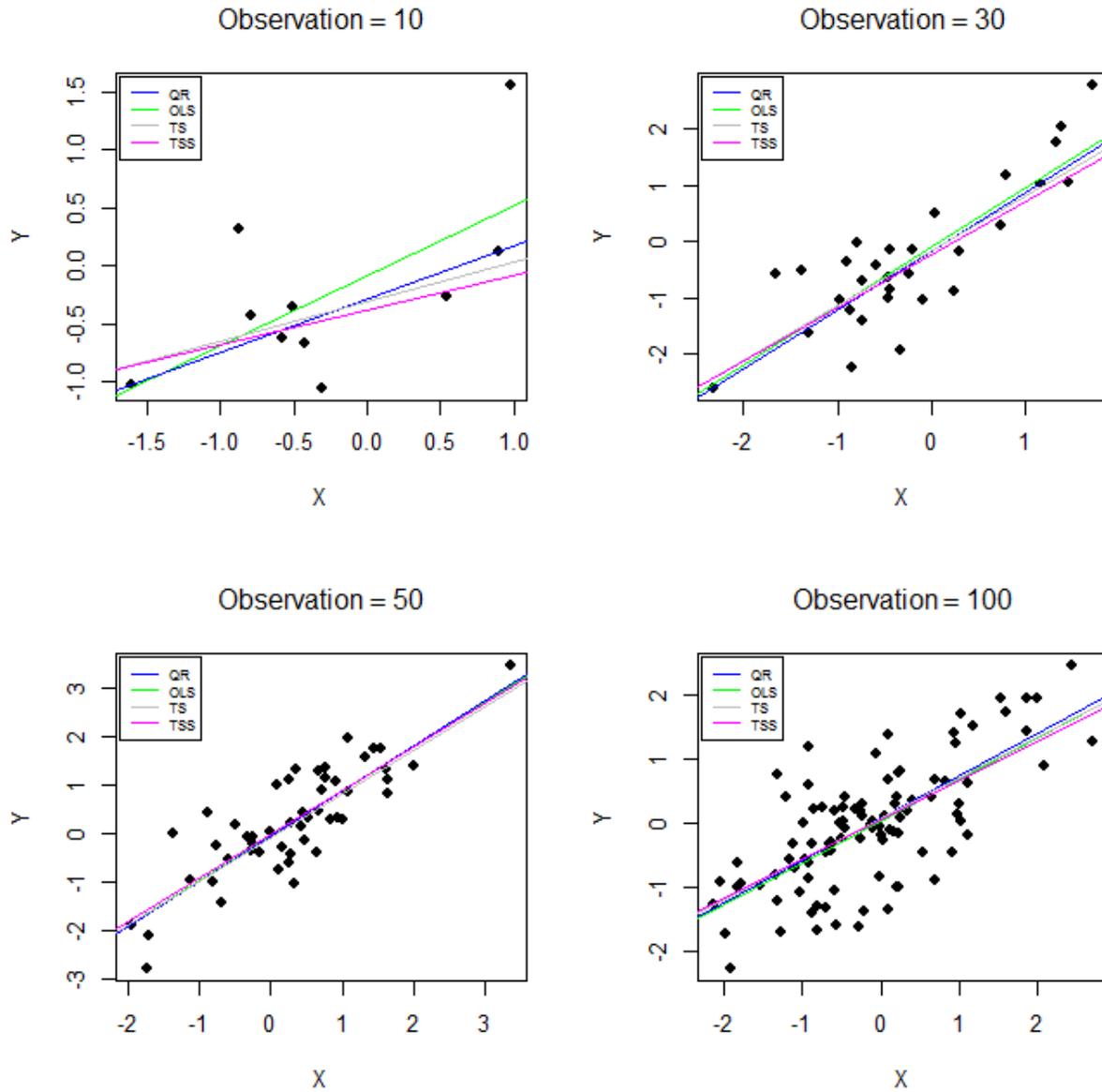
Table 2: Results from the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	RMSE	Bias
n=10	OLS	-0.08	0.60	0.25	0.043	(0.02, 1.17)	0.555	0.00
	QR	-0.29	0.46	0.32	0.193	(-0.31, 1.39)	0.590	0.17
	TS	-0.30	0.34	0.27	0.246	(-0.29, 0.97)	0.610	0.51
	TSS	-0.38	0.30	0.28	0.344	(-0.37, 0.96)	0.643	0.22
n=30	OLS	0.02	0.82	0.10	<0.001	(0.62, 1.02)	0.499	0.00
	QR	0.04	0.76	0.14	<0.001	(0.49, 1.12)	0.503	0.00
	TS	0.03	0.80	0.09	<0.001	(0.60, 1.00)	0.499	0.00
	TSS	0.02	0.77	0.09	<0.001	(0.57, 0.97)	0.502	0.02
n=50	OLS	0.11	0.85	0.11	<0.001	(0.64, 1.07)	0.659	0.00
	QR	0.15	0.79	0.17	<0.001	(0.49, 1.15)	0.662	0.05
	TS	0.15	0.81	0.11	<0.001	(0.60, 1.03)	0.661	0.05
	TSS	0.05	0.81	0.11	<0.001	(0.59, 1.03)	0.662	0.05
n=100	OLS	0.06	0.81	0.06	<0.001	(0.70, 0.92)	0.538	0.00
	QR	0.02	0.83	0.08	<0.001	(0.64, 0.96)	0.539	0.04
	TS	0.01	0.80	0.06	<0.001	(0.69, 0.91)	0.541	0.06
	TSS	0.03	0.82	0.06	<0.001	(0.71, 0.93)	0.539	0.04

Table 3: Results of Relative Root Mean Square Error of the four regression procedures at $n=10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$

	Relative Root Mean Square Error	Value
n=10	OLS vs QR	-0.063
	OLS vs TS	-0.099
	OLS vs TSS	-0.158
	QR vs TS	-0.034
	QR vs TSS	-0.089
	TS vs TSS	-0.054
n=30	OLS vs QR	-0.007
	OLS vs TS	-0.001
	OLS vs TSS	-0.006
	QR vs TS	0.006
	QR vs TSS	0.001
	TS vs TSS	-0.005
n=50	OLS vs QR	-0.005
	OLS vs TS	-0.004
	OLS vs TSS	-0.005
	QR vs TS	0.002
	QR vs TSS	0.001
	TS vs TSS	-0.001
n=100	OLS vs QR	-0.004
	OLS vs TS	-0.007
	OLS vs TSS	-0.003
	QR vs TS	-0.003
	QR vs TSS	0.001
	TS vs TSS	0.004

Figure 1: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression Model with slop and intercept under the Normality Assumption with no Outliers:

If the errors (e_i) were independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a uniform distribution with $\min=0$ and $\max=1$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$.

Table 4: Descriptive Statistics of (X, Y) in regression procedures with $n= 10, 30, 50, 100$,

$Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $e \sim Normal(n, 0, 2)$, and $Y = 2 + 3 * X + e$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	-0.24	5.68	2.89	2.98	2.01	2.82
X	10	0.10	0.89	0.39	0.33	0.25	0.17
Y	30	-0.28	9.93	3.39	3.53	2.34	2.95
X	30	0.02	0.90	0.48	0.51	0.29	0.49
Y	50	-0.84	6.14	3.26	3.29	1.71	2.15
X	50	0.00	0.95	0.43	0.41	0.29	0.54
Y	100	-1.58	8.26	3.80	3.59	2.16	3.15
X	100	0.00	0.99	0.52	0.50	0.29	0.53

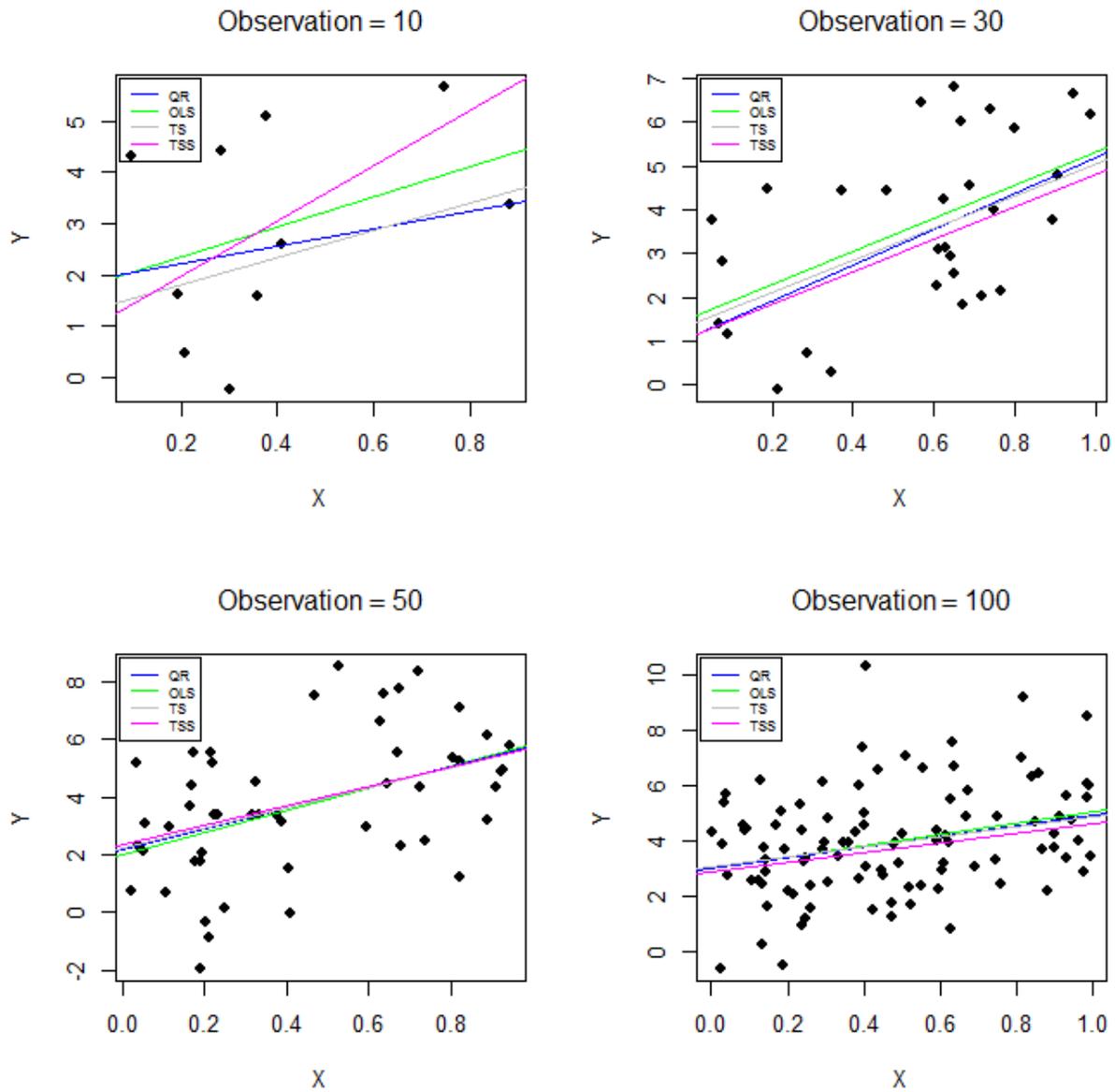
Table 5: Results of the four regression procedures with $n = 10, 30, 50, 100$ $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $e \sim Normal(n, 0, 2)$, $Y = 2 + 3 * X + e$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	RMSE	Bias
n=10	OLS	1.75	2.94	2.67	0.30	(-3.24, 9.12)	1.78	0.00
	QR	1.89	1.68	5.75	0.77	(-5.72, 9.68)	1.83	0.35
	TS	1.29	2.66	2.81	0.37	(-3.84, 9.15)	1.86	0.57
	TSS	0.91	5.38	2.82	0.09	(-1.12, 11.88)	1.87	-0.10
	OLS	2.55	1.74	1.48	0.25	(-1.30, 4.78)	2.242	0.00
n=30	QR	2.64	1.47	1.47	0.33	(-1.36, 4.37)	2.243	0.04
	TS	2.54	1.68	1.49	0.38	(-1.37, 4.73)	2.242	0.04
	TSS	2.59	1.41	1.49	0.35	(1.64, 4.47)	2.246	0.11
	OLS	2.34	2.16	0.78	<0.01	(0.58, 3.75)	1.569	0.00
n=50	QR	2.00	2.65	1.20	0.032	(0.01, 3.94)	1.580	0.12
	TS	2.25	2.01	0.79	0.014	(0.42, 3.60)	1.577	0.15
	TSS	2.18	2.20	0.79	<0.01	(0.62, 3.79)	1.575	0.13
	OLS	1.84	3.77	0.65	<0.001	(2.48, 5.06)	1.859	0.00
n=100	QR	1.99	3.49	0.81	<0.001	(2.44, 5.60)	1.860	0.00
	TS	1.78	4.00	0.65	<0.001	(2.71, 5.30)	1.861	-0.06
	TSS	1.93	4.01	0.66	<0.001	(2.71, 5.31)	1.872	-0.21
	OLS	1.84	3.77	0.65	<0.001	(2.48, 5.06)	1.859	0.00

Table 6: Results of Relative Root Mean Square Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $e \sim Normal(n, 0, 2)$, and $Y = 2 + 3 * X + e$:

	Relative Root Mean Square Error	Value
n=10	OLS vs QR	-0.063
	OLS vs TS	-0.099
	OLS vs TSS	-0.158
	QR vs TS	-0.034
	QR vs TSS	-0.089
	TS vs TSS	-0.054
n=30	OLS vs QR	-0.001
	OLS vs TS	-0.001
	OLS vs TSS	-0.002
	QR vs TS	0.001
	QR vs TSS	-0.001
	TS vs TSS	-0.002
n=50	OLS vs QR	-0.007
	OLS vs TS	-0.005
	OLS vs TSS	-0.004
	QR vs TS	0.002
	QR vs TSS	0.003
	TS vs TSS	0.001
n=100	OLS vs QR	-0.001
	OLS vs TS	-0.001
	OLS vs TSS	-0.007
	QR vs TS	-0.000
	QR vs TSS	-0.006
	TS vs TSS	-0.006

Figure 2: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim \text{Unif}(n, 0, 1)$, $e \sim \text{Normal}(n, 0, 2)$, and $Y = 2 + 3 * X + e$:



Regression Model with slop, intercept and dichotomous predictor variable with no Outliers:

If the errors (e_i) were independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a binomial distribution with a single trial and $p = 0.5$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$.

Table 7: Descriptive Statistics of Y variable in regression procedures with n= 10, 30, 50, 100, Nsim = 1000, $X \sim \text{Binomial}(n, 1, 0.5)$, $e \sim \text{Normal}(n, 0, 2)$, and $Y = 2 + 3 * X + e$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	-1.15	6.43	2.33	2.46	2.36	3.27
Y	30	-0.87	10.99	3.55	3.10	2.54	2.29
Y	50	-1.32	7.42	3.29	3.58	2.03	2.82
Y	100	-2.35	8.74	3.74	3.44	2.55	3.72

Table 8 Frequency distribution of X variable in regression procedures with n= 10, 30, 50, 100, Nsim = 1000, $X \sim \text{Binomial}(n, 1, 0.5)$, $e \sim \text{Normal}(n, 0, 2)$, and $Y = 2 + 3 * X + e$:

n=10	X	Frequency	Percentage	n=30	X	Frequency	Percentage
	0	8	80%		0	14	47%
	1	2	20%	1	16	53%	
n=50	X	Frequency	Percentage	n=100	X	Frequency	Percentage
	0	28	56%		0	50	50%
	1	22	44%	1	50	50%	

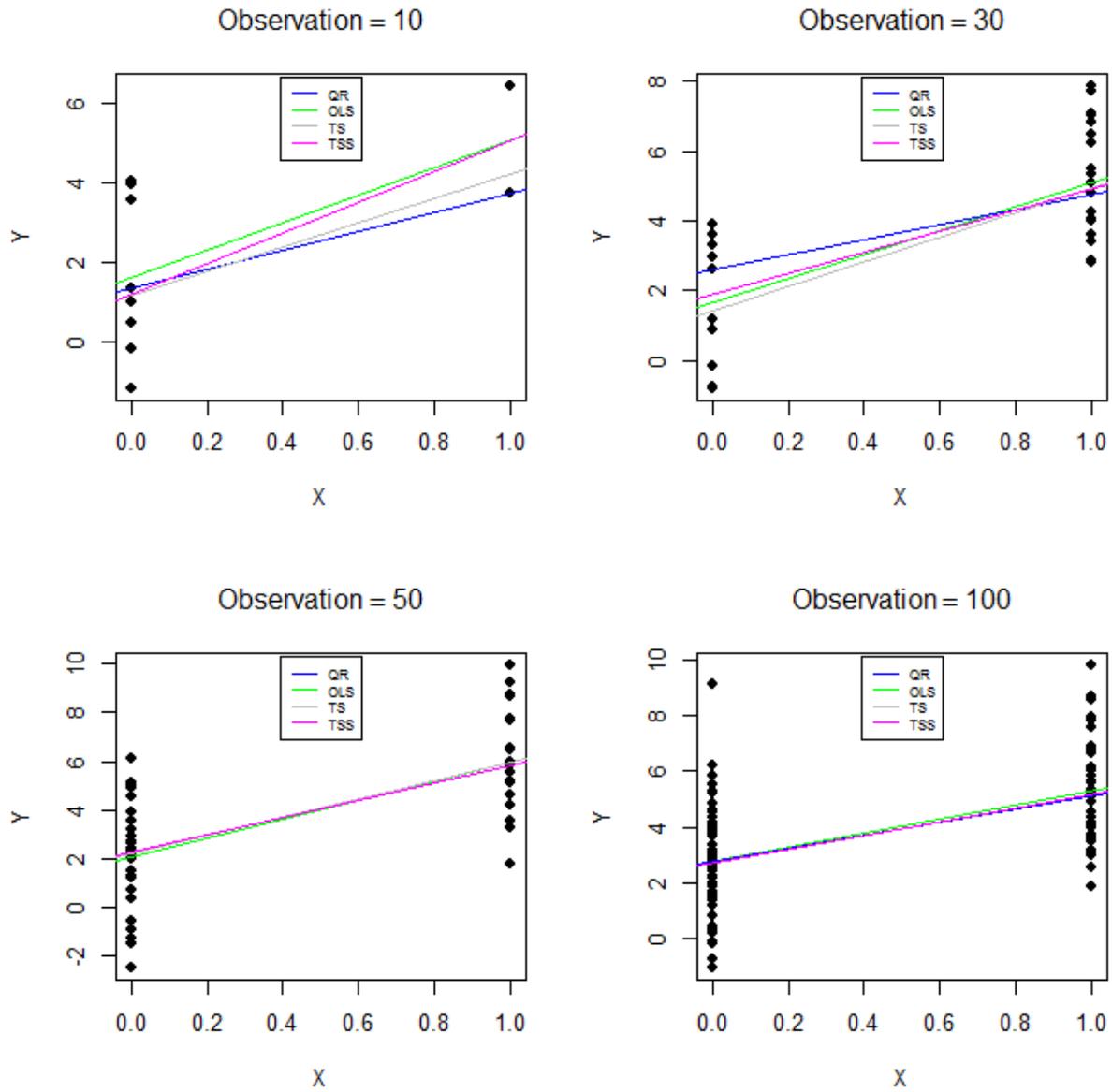
Table 9: Results of the four regression procedures with $n = 10, 30, 50, 100$ $Nsim = 1000$, $X \sim \text{Binomial}(n, 1, 0.5)$, $e \sim \text{Normal}(n, 0, 2)$, and $Y = 2 + 3 * X + e$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	RMSE	Bias
n=10	OLS	1.64	3.43	1.56	0.06	(-0.17, 7.04)	1.760	0.00
	QR	1.34	2.37	1.95	0.26	(0.00, 5.93)	1.889	0.50
	TS	1.18	3.05	1.64	0.10	(-0.73, 6.83)	1.855	0.53
	TSS	1.18	3.89	1.61	0.04	(0.19, 7.60)	1.816	0.36
n=30	OLS	2.35	2.55	0.84	0.01	(0.52, 3.99)	2.239	0.00
	QR	2.59	1.85	0.90	0.05	(0.11, 3.93)	2.249	-0.03
	TS	2.50	2.05	0.85	0.02	(0.31, 3.79)	2.242	-0.04
	TSS	2.55	1.95	1.85	0.03	(0.21, 3.69)	2.245	-0.03
n=50	OLS	2.18	2.52	0.45	<0.01	(1.61, 3.44)	1.569	0.00
	QR	2.25	2.27	0.67	<0.01	(0.98, 3.41)	1.574	0.04
	TS	2.15	2.47	0.45	<0.01	(1.55, 3.39)	1.570	0.05
	TSS	2.18	2.20	0.79	<0.01	(0.62, 3.79)	1.575	0.13
n=100	OLS	2.01	3.46	0.37	<0.01	(2.72, 4.21)	1.857	0.00
	QR	2.05	3.36	0.47	<0.01	(2.46, 4.48)	1.858	0.01
	TS	1.84	3.58	0.38	<0.01	(2.84, 4.33)	1.862	0.01
	TSS	1.93	3.49	0.38	<0.01	(2.75, 4.24)	1.858	0.07

Table 10: Results of Relative Root Mean Square Error of the four regression procedures with $n = 10, 30, 50, 100, N_{sim} = 1000, X \sim \text{Binomial}(n, 1, 0.5), e \sim \text{Normal}(n, 0, 2)$ and $Y = 2 + 3 * X + e$.

	Relative Root Mean Square Error	Value
n=10	OLS vs QR	-0.067
	OLS vs TS	-0.048
	OLS vs TSS	-0.026
	QR vs TS	0.018
	QR vs TSS	0.038
	TS vs TSS	0.021
n=30	OLS vs QR	-0.004
	OLS vs TS	-0.001
	OLS vs TSS	-0.002
	QR vs TS	0.003
	QR vs TSS	0.002
	TS vs TSS	-0.001
n=50	OLS vs QR	-0.004
	OLS vs TS	-0.008
	OLS vs TSS	-0.001
	QR vs TS	0.003
	QR vs TSS	0.002
	TS vs TSS	-0.004
n=100	OLS vs QR	-0.001
	OLS vs TS	-0.002
	OLS vs TSS	-0.001
	QR vs TS	- 0.002
	QR vs TSS	-0.038
	TS vs TSS	0.002

Figure 3: Four regression lines are shown in each plot with $n=10, 30, 50, 100, Nsim = 1000, X \sim \text{Binomial}(n, 1, 0.5), e \sim \text{Normal}(n, 0, 2), \text{ and } Y = 2 + 3 * X + e$:



Regression Model with Outliers in both X and Y direction:

If the errors (e_i) were independent and normally distributed, then a random sample of size n was generated from a bivariate normal distribution with mean $(0, 0)$ and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 20%, 30%, and 50% of n were introduced in both X and Y variables from a bivariate normal distribution with means $(2, 6)$ with variances $0.1 \times$ variance of the above bivariate normal distribution, i.e. the variances $(0.1, 0.1)$.

Regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in both X and Y variables:

Table 11: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in both X and Y variables:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	11	-1.86	5.79	0.25	0.16	1.97	0.86
	X	11	-1.68	2.41	-0.002	-0.19	1.18	1.31
20%	Y	12	-1.24	5.81	1.21	0.48	2.23	1.03
	X	12	-1.74	1.85	0.45	0.26	1.09	1.55
30%	Y	13	-1.23	5.81	1.22	0.48	2.24	1.04
	X	13	-1.74	1.85	0.45	0.26	1.09	1.55
50%	Y	15	-1.31	6.30	1.91	0.45	2.95	5.94
	X	15	-1.11	2.06	0.84	0.82	0.99	1.63

Table 12: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

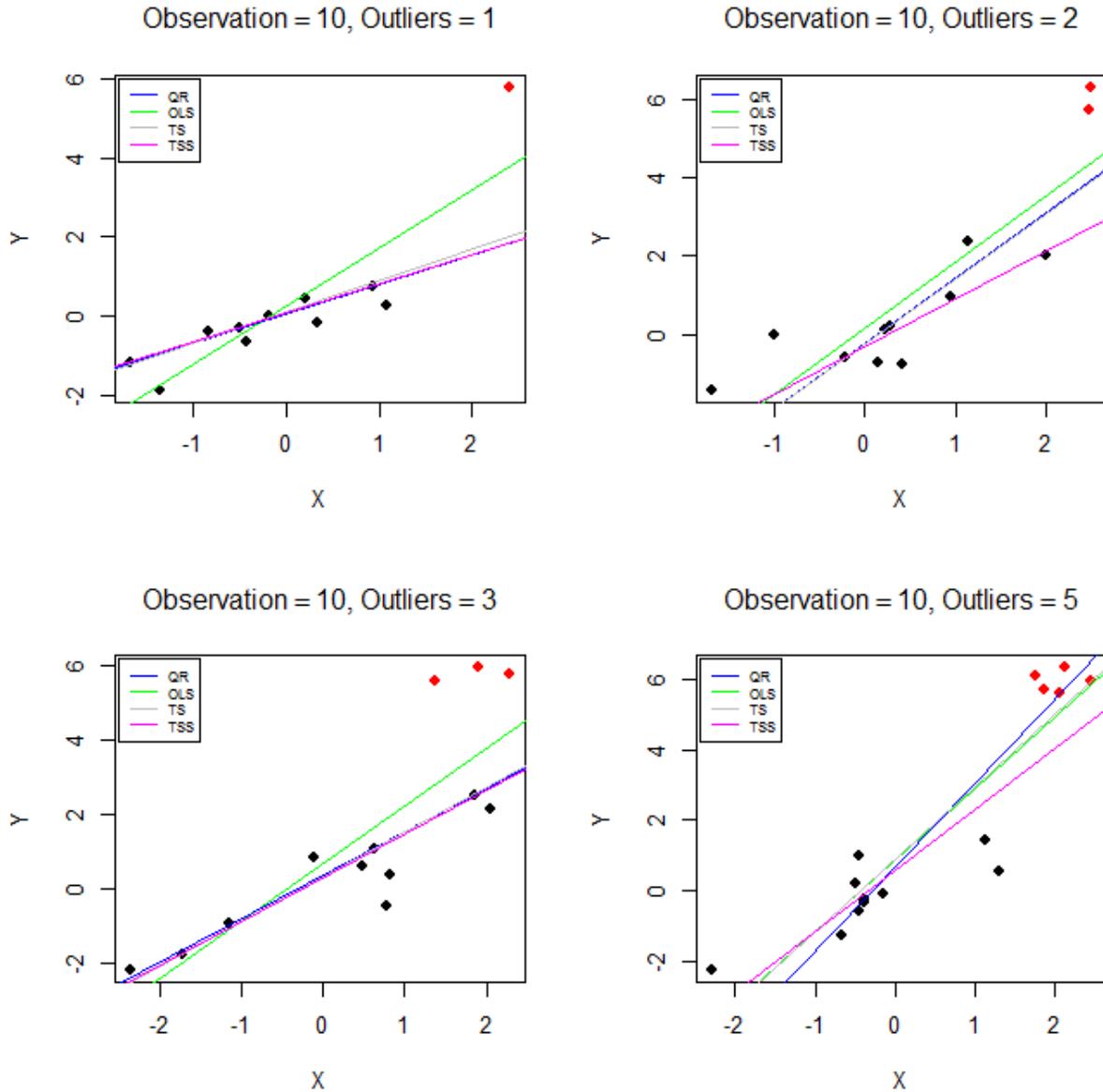
Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.25	1.46	0.278	<0.001	(0.83, 2.09)	0.706	-0.12
	QR	0.09	0.74	0.194	0.004	(0.36, 2.27)	0.198	0.00
	TS	0.12	0.79	0.784 (MAD)	0.055	(0.65, 1.25)	0.163	0.00
	TSS	0.08	0.74	0.204 (MAD)	<0.001	(0.64, 1.39)	0.205	0.00
20%	OLS	0.54	1.50	0.441	<0.01	(0.52, 2.48)	0.718	-0.36
	QR	0.28	0.87	0.201	0.01	(0.46, 2.36)	0.298	0.00
	TS	0.28	0.80	0.753 (MAD)	<0.001	(0.68, 1.69)	0.280	0.00
	TSS	0.28	0.87	0.427 (MAD)	<0.01	(0.65, 2.06)	0.298	0.00
30%	OLS	0.14	1.99	0.526	<0.01	(0.83, 3.15)	1.267	-0.00
	QR	-0.08	1.60	0.430	<0.01	(0.75, 2.44)	0.902	0.50
	TS	-0.16	1.70	2.460 (MAD)	<0.001	(1.49, 3.02)	0.944	0.51
	TSS	-0.38	1.11	0.448 (MAD)	<0.001	(0.85, 2.80)	0.523	1.15
50%	OLS	-0.20	2.52	0.444	<0.001	(1.56, 3.47)	1.309	0.46
	QR	-0.05	2.84	0.614	<0.001	(1.63, 4.04)	0.957	0.00
	TS	0.28	2.42	2.69 (MAD)	<0.001	(1.89, 3.04)	1.337	0.00
	TSS	-0.51	1.90	1.197 (MAD)	<0.01	(1.51, 2.65)	1.477	0.82

Table 13: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, $N_{sim} = 1000$, $X \sim Normal(n, 0, 1)$,

$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.719
	OLS vs TS	0.769
	OLS vs TSS	0.710
	QR vs TS	0.175
	QR vs TSS	-0.034
	TS vs TSS	-0.253
20%	OLS vs QR	0.585
	OLS vs TS	0.609
	OLS vs TSS	0.585
	QR vs TS	0.061
	QR vs TSS	0.000
	TS vs TSS	-0.001
30%	OLS vs QR	0.288
	OLS vs TS	0.255
	OLS vs TSS	0.587
	QR vs TS	-0.046
	QR vs TSS	0.421
	TS vs TSS	-0.004
50%	OLS vs QR	0.268
	OLS vs TS	-0.022
	OLS vs TSS	-0.128
	QR vs TS	-0.297
	QR vs TSS	-0.543
	TS vs TSS	-0.1.5

Figure 4: Four regression lines are shown in each plot with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 30$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in both X and Y variables:

Table 14: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in both X and Y variables:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	33	-1.57	6.11	0.80	0.36	1.80	0.95
	X	33	-1.09	2.54	0.30	0.32	0.94	1.20
20%	Y	36	-1.29	6.49	1.25	0.47	2.32	1.75
	X	36	-1.70	2.43	0.45	0.39	1.12	1.60
30%	Y	39	-2.19	6.50	1.65	1.00	2.74	3.02
	X	39	-1.43	2.38	0.61	0.60	1.21	1.94
50%	Y	45	-2.42	6.47	1.90	0.96	3.18	6.72
	X	45	-1.86	2.58	0.54	0.69	1.37	2.17

Table 15: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, $N_{sim} = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) =$

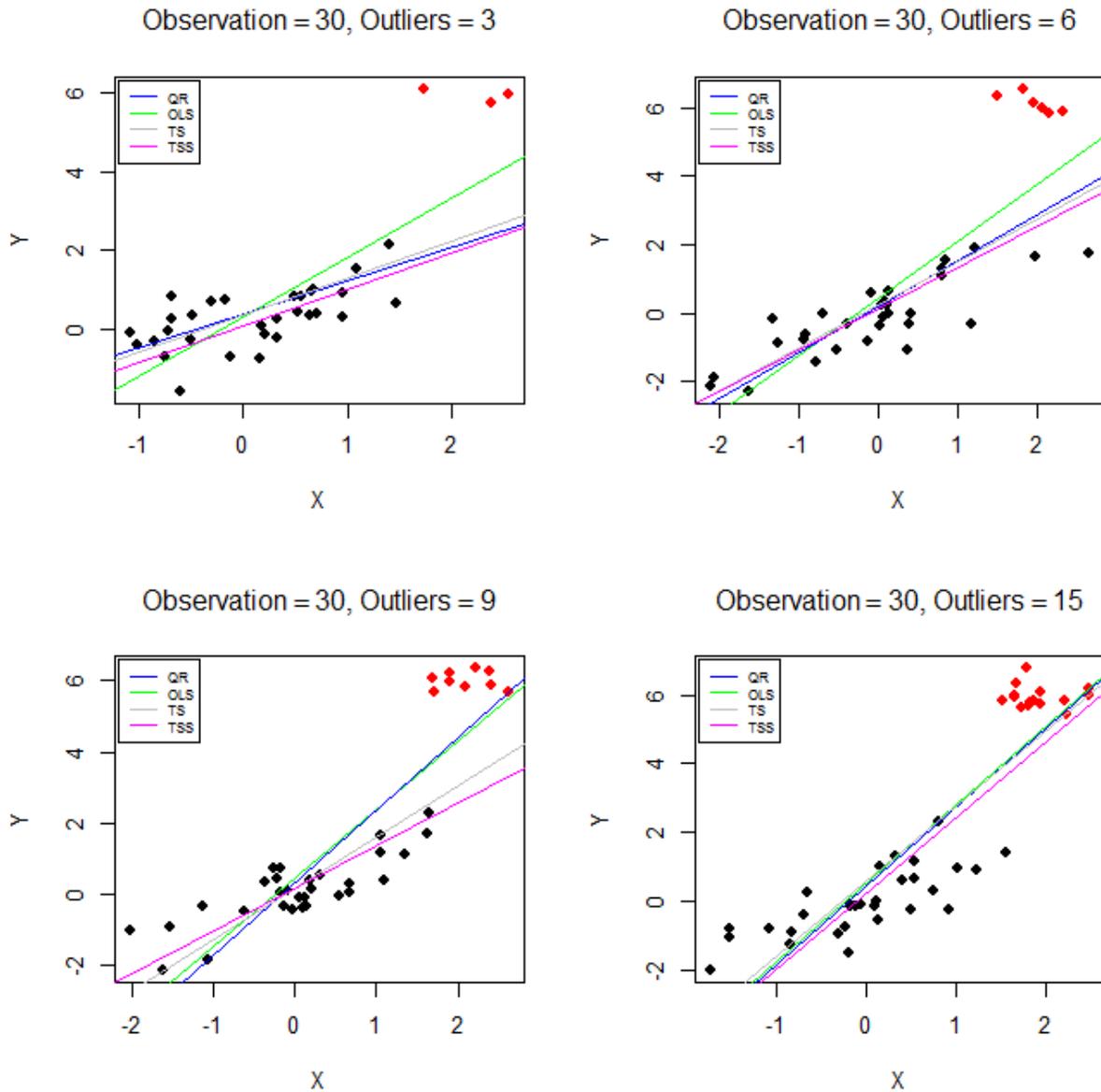
0.80:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.25	1.46	0.278	<0.001	(0.83, 2.09)	0.706	-0.12
	QR	0.09	0.74	0.194	0.004	(0.36, 2.27)	0.198	0.00
	TS	0.12	0.79	0.784 (MAD)	0.055	(0.65, 1.25)	0.163	0.00
	TSS	0.08	0.74	0.204 (MAD)	<0.001	(0.64, 1.39)	0.205	0.00
20%	OLS	0.78	1.99	0.183	<0.01	(1.62, 2.35)	1.325	0.00
	QR	0.89	2.07	0.296	<0.01	(1.49, 2.65)	1.156	-0.15
	TS	1.23	1.90	1.890 (MAD)	<0.01	(1.67, 1.94)	1.291	-0.40
	TSS	0.52	2.05	0.620 (MAD)	<0.01	(1.70, 2.14)	1.226	0.24
30%	OLS	0.46	1.94	0.194	<0.01	(1.44, 2.34)	1.211	0.00
	QR	0.31	1.78	0.334	<0.01	(1.13, 2.44)	1.025	0.25
	TS	0.37	1.82	1.540 (MAD)	<0.01	(1.78, 2.09)	1.109	0.16
	TSS	-0.06	1.52	0.670 (MAD)	<0.01	(1.39, 2.03)	0.920	0.77
50%	OLS	0.78	1.99	0.183	<0.01	(1.62, 2.35)	1.325	0.00
	QR	0.89	2.07	0.296	<0.01	(1.49, 2.65)	1.156	-0.15
	TS	1.23	1.90	1.890 (MAD)	<0.01	(1.67, 1.94)	1.291	-0.40
	TSS	0.52	2.05	0.620 (MAD)	<0.01	(1.39, 2.03)	0.920	0.77

Table 16: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.301
	OLS vs TS	0.195
	OLS vs TSS	0.209
	QR vs TS	-0.151
	QR vs TSS	-0.131
	TS vs TSS	0.018
20%	OLS vs QR	0.547
	OLS vs TS	0.505
	OLS vs TSS	0.511
	QR vs TS	-0.092
	QR vs TSS	-0.079
	TS vs TSS	0.012
30%	OLS vs QR	0.154
	OLS vs TS	0.084
	OLS vs TSS	0.242
	QR vs TS	-0.082
	QR vs TSS	0.104
	TS vs TSS	0.171
50%	OLS vs QR	0.128
	OLS vs TS	0.025
	OLS vs TSS	0.074
	QR vs TS	-0.117
	QR vs TSS	-0.061
	TS vs TSS	0.050

Figure 5: Four regression lines are shown in each plot with $n = 30$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 50$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in both X and Y variables:

Table 17: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in both X and Y variables:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	55	-2.88	6.54	0.66	0.30	1.98	1.40
	X	55	-3.14	2.44	0.32	0.42	1.23	1.61
20%	Y	60	-2.53	6.59	0.88	0.09	2.42	1.40
	X	60	-1.66	2.88	0.20	0.14	1.17	1.71
30%	Y	65	-2.37	6.49	1.17	0.23	2.76	1.69
	X	65	-3.50	2.56	0.25	0.14	1.26	1.59
50%	Y	75	-2.92	6.55	2.03	0.82	2.97	6.06
	X	75	-1.84	2.69	0.62	0.45	1.23	2.09

Table 18: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) =$

0.80:

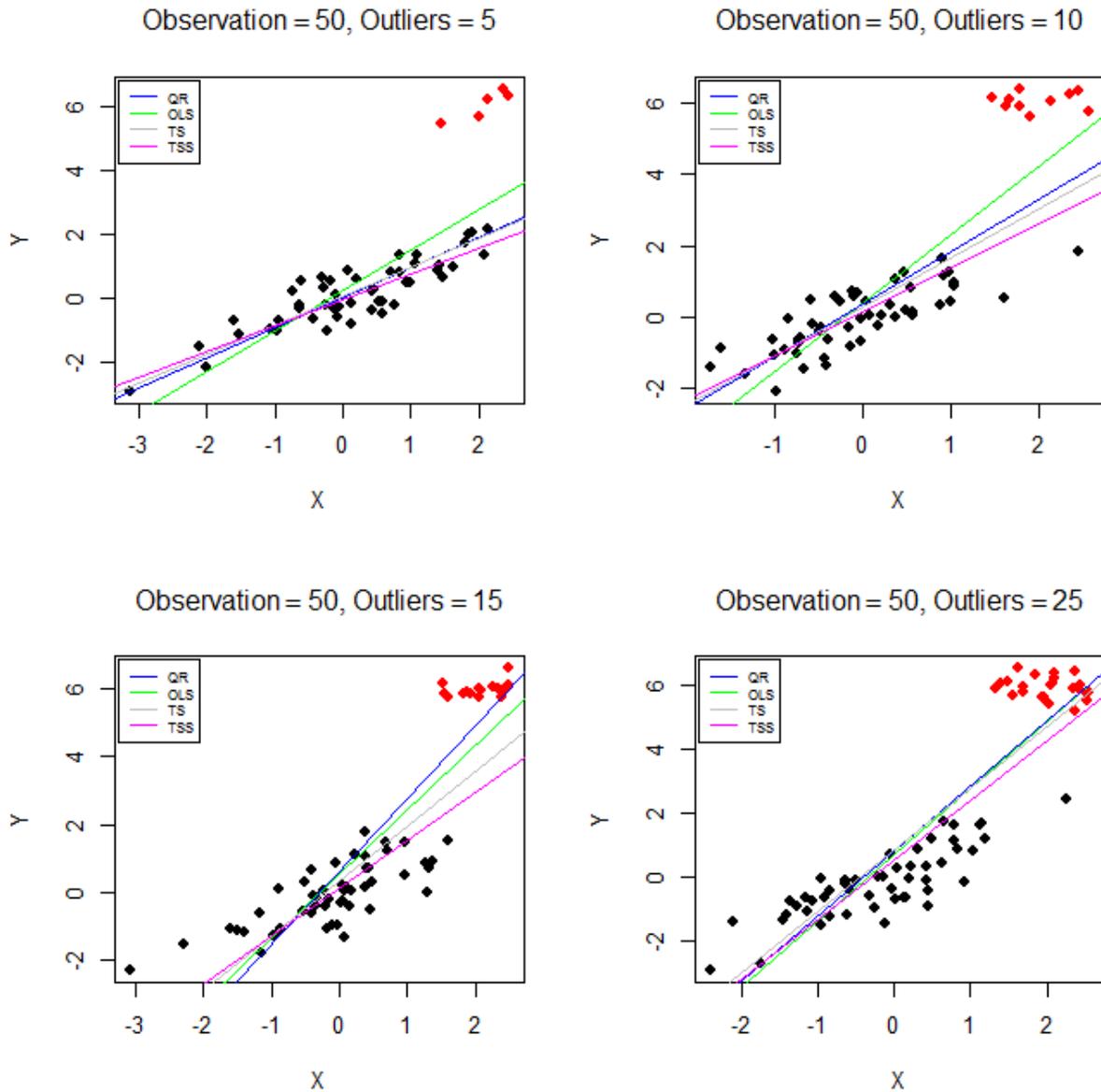
Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.26	1.26	0.140	<0.01	(0.98, 1.53)	0.660	0.00
	QR	0.06	0.94	0.103	<0.01	(0.73, 1.14)	0.425	0.31
	TS	0.08	0.90	1.00 (MAD)	<0.01	(0.95, 1.09)	0.416	0.30
	TSS	-0.03	0.80	0.217 (MAD)	<0.01	(0.77, 0.92)	0.418	0.44
20%	OLS	0.55	1.69	0.155	<0.01	(1.38, 2.00)	0.944	0.00
	QR	0.18	1.12	0.143	<0.01	(0.84, 1.40)	0.575	0.48
	TS	0.16	1.06	1.53 (MAD)	<0.01	(1.27, 1.44)	0.558	0.52
	TSS	0.11	0.91	0.412 (MAD)	<0.01	(0.87, 1.34)	0.516	0.59
30%	OLS	0.69	1.89	0.140	<0.01	(1.16, 2.17)	0.960	0.00
	QR	0.44	2.10	0.167	<0.01	(1.77, 2.43)	1.050	0.19
	TS	0.42	1.56	1.740 (MAD)	<0.01	(1.55, 1.71)	0.951	0.35
	TSS	0.19	1.24	0.723 (MAD)	<0.01	(1.20, 1.72)	0.743	0.66
50%	OLS	0.71	2.13	0.133	<0.01	(1.87, 2.40)	1.081	0.00
	QR	0.76	2.36	0.174	<0.01	(2.02, 2.70)	0.968	-0.19
	TS	0.89	1.94	1.784 (MAD)	<0.01	(1.78, 1.93)	1.204	-0.06
	TSS	0.58	1.80	0.669 (MAD)	<0.01	(1.20, 1.72)	0.743	0.66

Table 19: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, $N_{sim} = 1000$, $X \sim Normal(n, 0, 1)$,

$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.355
	OLS vs TS	0.369
	OLS vs TSS	0.366
	QR vs TS	0.021
	QR vs TSS	0.016
	TS vs TSS	-0.005
20%	OLS vs QR	0.391
	OLS vs TS	0.409
	OLS vs TSS	0.454
	QR vs TS	0.031
	QR vs TSS	0.103
	TS vs TSS	0.074
30%	OLS vs QR	-0.092
	OLS vs TS	0.009
	OLS vs TSS	0.227
	QR vs TS	0.093
	QR vs TSS	0.291
	TS vs TSS	0.219
50%	OLS vs QR	0.105
	OLS vs TS	-0.113
	OLS vs TSS	-0.187
	QR vs TS	-0.244
	QR vs TSS	-0.327
	TS vs TSS	-0.067

Figure 6: Four regression lines are shown in each plot with $n = 50$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in both X and Y variables:

Table 20: Descriptive Statistics of (X, Y) in regression analysis with $n = 100$ and outliers of 10%, 20%, 30%, and 50% of $n = 100$ in both X and Y variables:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	110	-2.57	6.80	0.59	0.21	2.01	1.94
	X	110	-2.35	2.80	0.20	0.18	1.13	1.60
20%	Y	120	-3.18	6.78	1.00	0.22	2.43	1.60
	X	120	-2.81	2.93	0.31	0.35	1.30	1.80
30%	Y	130	-2.72	6.84	1.12	0.15	2.80	2.38
	X	130	-3.01	2.86	0.29	0.12	1.39	2.38
50%	Y	150	-2.21	6.41	2.07	0.74	2.84	5.80
	X	150	-1.95	3.31	0.77	0.77	1.17	1.98

Table 21: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$,

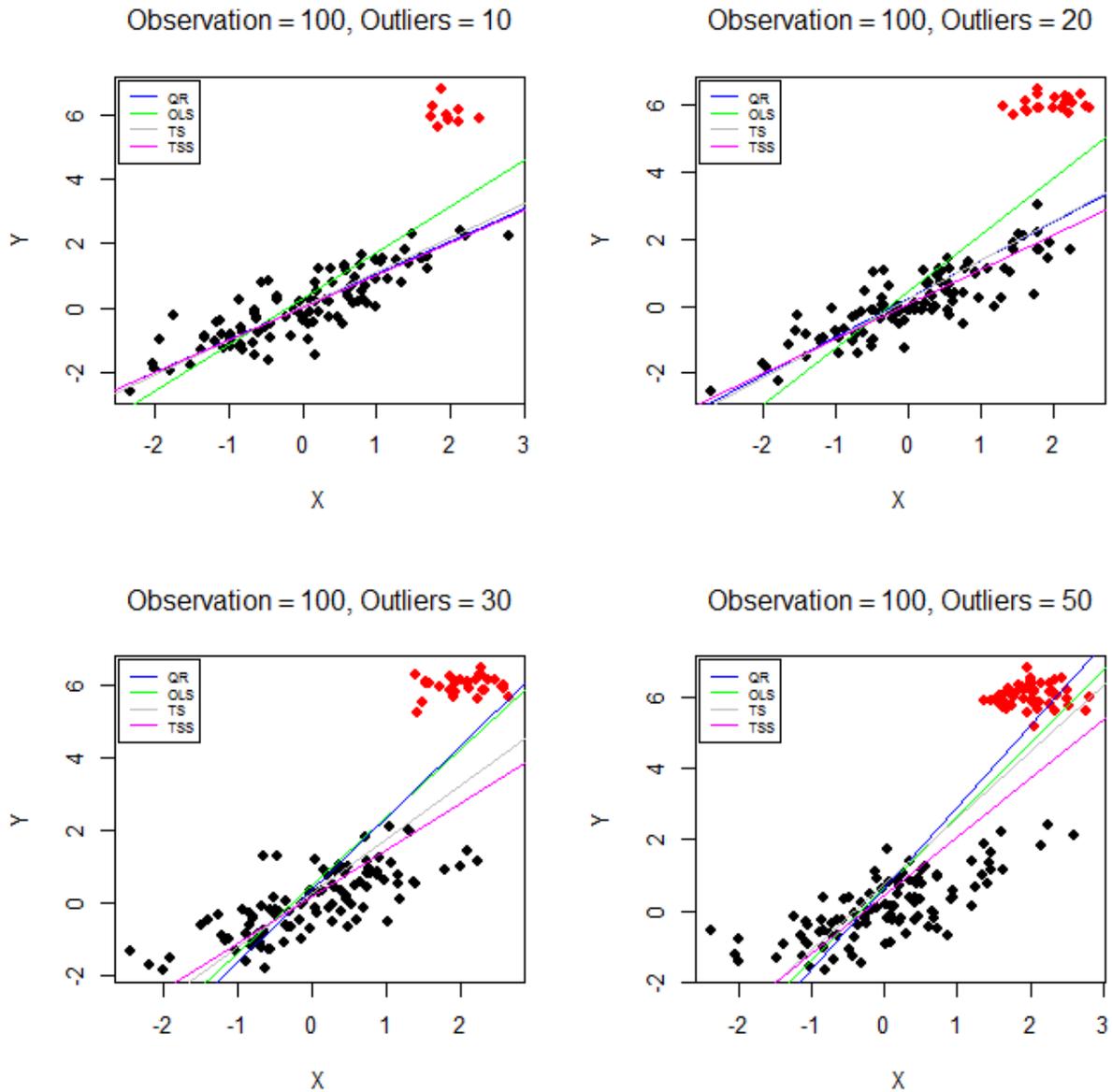
and $Cor(X, Y) = 0.80$:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.31	1.01	0.101	<0.001	(1.23, 1.63)	0.638	-0.17
	QR	0.04	2.26	0.067	<0.001	(0.88, 1.14)	0.428	0.00
	TS	0.06	1.07	1.13 (MAD)	<0.001	(1.15, 1.23)	0.417	0.00
	TSS	0.03	1.00	0.241 (MAD)	<0.001	(0.98, 1.11)	0.439	0.01
20%	OLS	0.53	1.56	0.090	<0.001	(1.37, 1.75)	0.991	-0.09
	QR	0.27	1.16	0.072	<0.001	(1.01, 1.30)	0.651	0.00
	TS	0.25	1.11	1.348 (MAD)	<0.001	(1.28, 1.36)	0.643	0.00
	TSS	0.23	0.95	0.367 (MAD)	<0.001	(0.92, 1.11)	0.600	0.35
30%	OLS	0.63	1.73	0.091	<0.001	(1.55, 1.91)	1.083	-0.14
	QR	0.44	1.66	0.155	<0.001	(1.36, 1.97)	1.088	0.00
	TS	0.32	1.37	1.383 (MAD)	<0.001	(1.48, 1.55)	0.935	0.00
	TSS	0.09	1.10	0.442 (MAD)	<0.001	(1.12, 1.50)	0.715	0.16
50%	OLS	0.48	2.08	0.101	<0.001	(1.88, 2.28)	1.092	0.04
	QR	0.50	2.33	0.132	<0.001	(2.07, 2.59)	0.994	0.00
	TS	0.52	1.93	1.977 (MAD)	<0.001	(1.86, 1.95)	1.188	0.00
	TSS	0.33	1.74	0.844 (MAD)	<0.001	(1.73, 1.97)	1.303	0.16

Table 22: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.328
	OLS vs TS	0.346
	OLS vs TSS	0.311
	QR vs TS	0.027
	QR vs TSS	-0.026
	TS vs TSS	-0.044
20%	OLS vs QR	0.343
	OLS vs TS	0.351
	OLS vs TSS	0.394
	QR vs TS	0.011
	QR vs TSS	0.077
	TS vs TSS	0.074
30%	OLS vs QR	-0.005
	OLS vs TS	0.137
	OLS vs TSS	0.339
	QR vs TS	0.141
	QR vs TSS	0.343
	TS vs TSS	0.235
50%	OLS vs QR	0.089
	OLS vs TS	-0.087
	OLS vs TSS	-0.194
	QR vs TS	-0.195
	QR vs TSS	-0.311
	TS vs TSS	-0.097

Figure 7: Four regression lines are shown in each plot with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in both X and Y with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression Model with Outliers in Y direction only:

If the errors (e_i) were independent and normally distributed, then a random sample of size n was generated from a bivariate normal distribution with mean $(0, 0)$ and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 30% 50% and 100% of n were introduced in Y variable only from a bivariate normal distribution with means $(0, 6)$ and variances $0.1 \times$ variance of the above bivariate normal distribution, i.e. the variances $(0.1, 0.1)$.

Regression analysis with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in Y variable only:

Table 23: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 10$ in Y variable only:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	11	-1.86	5.79	0.25	0.16	1.97	0.86
	X	11	-1.68	1.07	-0.18	-0.19	0.88	1.05
20%	Y	12	-1.24	5.81	1.21	0.48	2.23	1.03
	X	12	-1.74	1.85	0.11	0.12	0.95	0.67
30%	Y	13	-1.23	5.81	1.22	0.48	2.24	1.04
	X	13	-1.27	1.26	0.25	0.33	0.80	1.34
50%	Y	15	-1.31	6.30	1.91	0.45	2.95	5.94
	X	15	-1.11	1.55	0.17	0.05	0.64	0.56

Table 24: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,

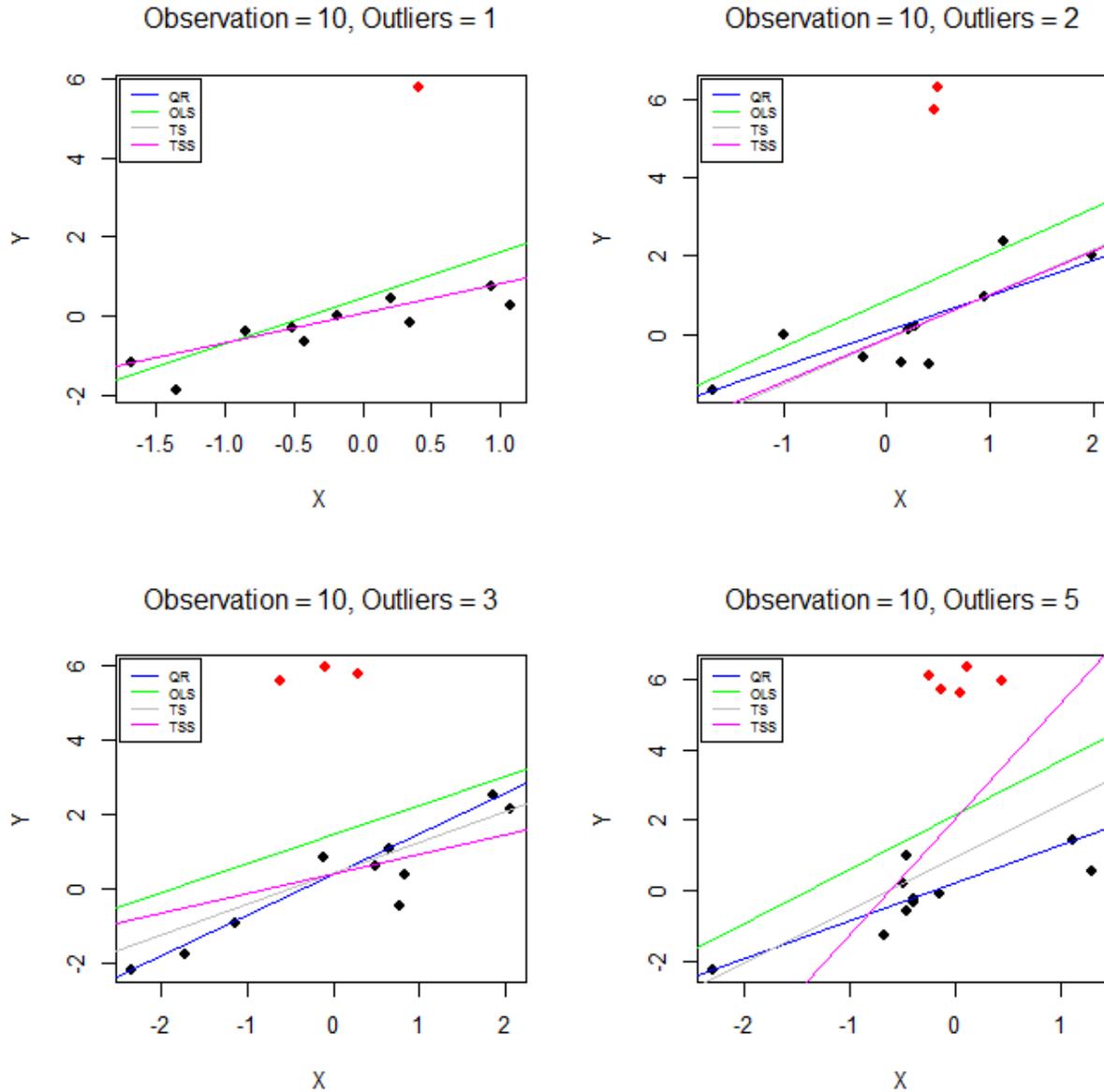
$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.46	1.17	0.636	0.100	(-0.27, 2.61)	0.480	0.27
	QR	0.07	0.74	0.25	0.017	(0.24, 1.23)	0.211	0.00
	TS	0.08	0.74	0.80 (MAD)	<0.001	(0.57, 1.24)	0.198	0.00
	TSS	0.08	0.74	0.204 (MAD)	<0.001	(0.62, 3.07)	0.205	0.00
20%	OLS	1.20	0.13	0.740	0.860	(-1.52, 1.78)	0.409	-0.75
	QR	0.33	0.48	0.171	0.018	(0.15, 0.82)	0.151	0.00
	TS	0.33	0.49	0.800 (MAD)	0.403	(-0.99, 0.65)	0.152	0.00
	TSS	0.36	0.46	0.226 (MAD)	0.224	(-1.32, 0.67)	0.145	-0.01
30%	OLS	1.68	-0.50	0.93	0.603	(-2.55, 1.55)	2.034	-0.28
	QR	0.43	0.43	0.57	0.459	(-0.69, 1.56)	0.418	0.00
	TS	0.88	0.02	2.99 (MAD)	0.601	(-1.38, 0.51)	0.988	0.00
	TSS	0.82	0.46	0.226 (MAD)	0.224	(-1.62, 0.44)	1.141	0.51
50%	OLS	2.05	-0.87	1.28	0.499	(-3.56, 1.83)	1.729	-0.89
	QR	0.19	0.31	0.56	0.584	(-0.79, 1.41)	0.588	0.00
	TS	0.46	-0.02	5.47 (MAD)	0.284	(-2.94, 0.44)	1.125	0.00
	TSS	0.52	-0.18	0.815 (MAD)	0.359	(-1.73, 0.25)	1.289	0.07

Table 25: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 10$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.560
	OLS vs TS	0.588
	OLS vs TSS	0.574
	R vs TS	0.063
	QR vs TSS	0.032
	TS vs TSS	-0.034
20%	OLS vs QR	0.632
	OLS vs TS	0.628
	OLS vs TSS	0.646
	R vs TS	-0.011
	QR vs TSS	0.038
	TS vs TSS	0.049
30%	OLS vs QR	0.794
	OLS vs TS	0.514
	OLS vs TSS	0.439
	R vs TS	-1.362
	QR vs TSS	-1.172
	TS vs TSS	-0.156
50%	OLS vs QR	0.660
	OLS vs TS	0.349
	OLS vs TSS	0.254
	R vs TS	-0.915
	QR vs TSS	-1.195
	TS vs TSS	-0.146

Figure 8: Four regression lines are shown in each plot with $n = 10$ and outliers of 10%, 20%, 30% and 50% of $n = 10$ in Y only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 30$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in Y variable only:

Table 26: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 30$ in Y variable only:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	33	-1.57	6.11	0.80	0.36	1.80	0.95
	X	33	-1.09	1.46	0.12	0.21	0.70	1.14
20%	Y	36	-1.70	2.04	0.12	0.09	0.85	1.10
	X	36	-1.70	2.43	0.45	0.39	1.12	1.60
30%	Y	39	-2.19	6.50	1.65	1.00	2.74	3.02
	X	39	-1.43	1.88	0.15	0.18	0.86	0.77
50%	Y	45	-2.42	6.47	1.90	0.96	3.18	6.72
	X	45	-1.86	2.58	-0.13	-0.19	0.96	1.00

Table 27: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,

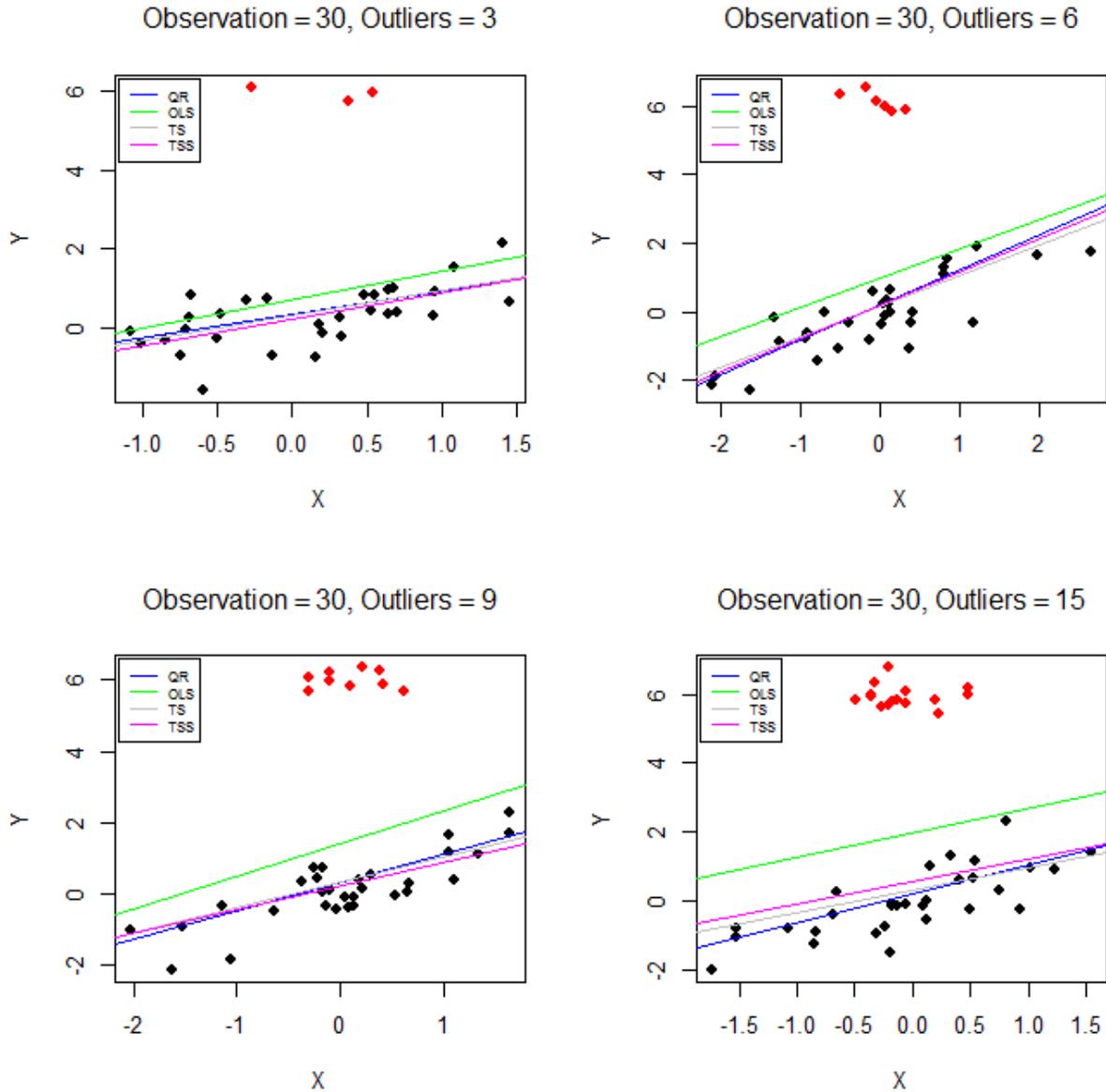
$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.71	0.72	0.440	0.109	(-0.17, 1.62)	0.443	-0.37
	QR	0.36	0.60	0.251	0.024	(0.10, 1.09)	0.395	0.00
	TS	0.31	0.65	1.660 (MAD)	<0.001	(0.59, 0.96)	0.198	0.00
	TSS	0.23	0.68	0.491 (MAD)	<0.001	(0.53, 0.94)	0.434	0.07
20%	OLS	1.17	0.63	0.454	0.171	(-0.29, 1.56)	0.475	-0.73
	QR	0.35	0.70	0.199	<0.01	(0.31, 1.09)	0.434	0.00
	TS	0.36	0.73	1.660 (MAD)	<0.001	(0.51, 0.84)	0.423	0.00
	TSS	0.35	0.72	0.280 (MAD)	<0.001	(0.55, 0.84)	0.434	0.01
30%	OLS	1.51	0.89	0.503	0.084	(-0.13, 1.91)	0.668	-1.18
	QR	0.31	0.98	0.120	<0.01	(0.74, 1.21)	0.665	0.00
	TS	0.32	0.90	2.891 (MAD)	<0.001	(0.61, 1.07)	0.667	0.00
	TSS	0.38	0.89	0.452 (MAD)	<0.001	(0.53, 1.02)	0.667	-0.05
50%	OLS	1.98	0.94	0.481	0.057	(-0.03, 1.92)	1.021	-1.18
	QR	0.39	0.92	0.293	<0.01	(0.34, 1.49)	0.946	0.00
	TS	0.35	0.82	4.149 (MAD)	<0.001	(0.70, 1.23)	0.929	0.00
	TSS	0.68	0.83	1.091 (MAD)	<0.001	(0.43, 1.04)	0.944	-0.31

Table 28: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 30$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.108
	OLS vs TS	0.105
	OLS vs TSS	0.019
	R vs TS	-0.003
	QR vs TSS	-0.099
	TS vs TSS	-0.096
20%	OLS vs QR	0.086
	OLS vs TS	0.108
	OLS vs TSS	0.085
	R vs TS	0.024
	QR vs TSS	-0.002
	TS vs TSS	-0.025
30%	OLS vs QR	0.003
	OLS vs TS	0.001
	OLS vs TSS	0.001
	R vs TS	-0.002
	QR vs TSS	-0.002
	TS vs TSS	-0.000
50%	OLS vs QR	0.073
	OLS vs TS	0.089
	OLS vs TSS	0.075
	R vs TS	0.018
	QR vs TSS	0.003
	TS vs TSS	-0.015

Figure 9: Four regression lines are shown in each plot with $n = 30$ and outliers of 10%, 20%, 30% and 50% of $n = 30$ in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 50$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in Y variable only:

Table 29: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 50$ in Y variable only:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	55	-2.88	6.54	0.66	0.30	1.98	1.40
	X	55	-3.14	2.14	0.14	0.13	1.10	1.34
20%	Y	60	-2.53	6.59	0.88	0.09	2.42	1.40
	X	60	-1.66	2.88	-0.13	-0.25	0.88	1.00
30%	Y	65	-2.37	6.49	1.17	0.23	2.76	1.69
	X	65	-3.50	1.56	-0.21	-0.09	0.83	0.77
50%	Y	75	-2.92	6.55	2.03	0.82	2.97	6.06
	X	75	-1.84	2.44	-0.05	-0.07	0.77	0.72

Table 30: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,

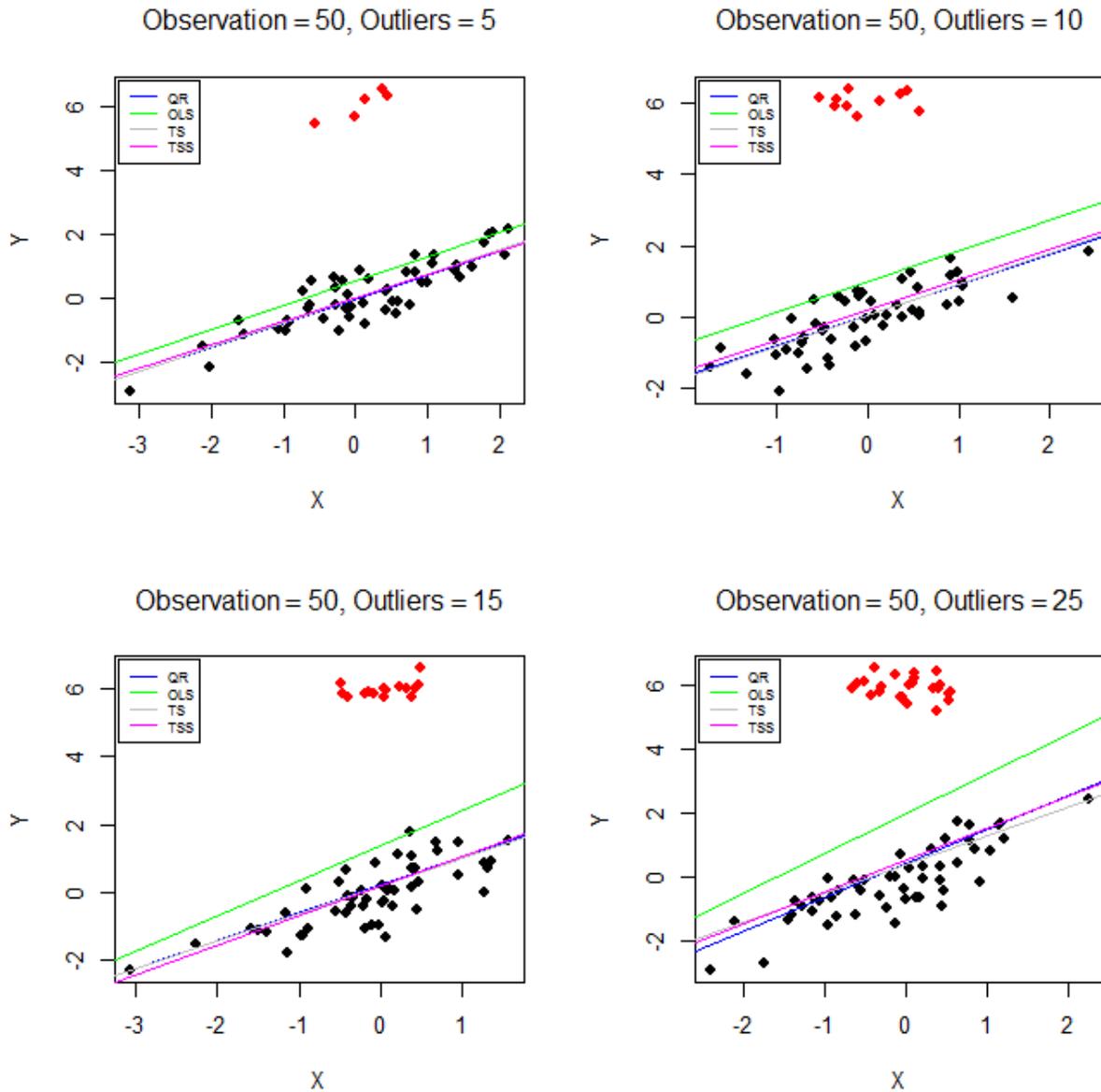
$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.56	0.76	0.224	<0.01	(0.31, 1.21)	0.467	-0.54
	QR	0.00	0.74	0.110	<0.01	(0.53, 0.96)	0.486	0.00
	TS	0.01	0.76	0.947 (MAD)	<0.001	(0.68, 0.80)	0.471	0.00
	TSS	0.04	0.73	0.305 (MAD)	<0.001	(0.67, 0.82)	0.493	-0.04
20%	OLS	0.99	0.80	0.347	0.025	(0.10, 1.49)	0.452	-0.89
	QR	0.10	0.70	0.117	<0.01	(0.47, 0.93)	0.438	0.00
	TS	0.10	0.73	1.625 (MAD)	<0.001	(0.70, 0.90)	0.424	0.00
	TSS	0.15	0.75	0.400 (MAD)	<0.001	(0.66, 1.02)	0.420	-0.04
30%	OLS	1.40	1.13	0.393	<0.01	(0.35, 1.92)	0.693	-1.17
	QR	0.25	0.72	0.122	<0.001	(0.04, 0.45)	0.539	0.00
	TS	0.31	0.81	3.158 (MAD)	<0.001	(0.79, 1.08)	0.594	0.00
	TSS	0.27	0.85	0.551 (MAD)	<0.001	(0.74, 1.11)	0.589	-0.04
50%	OLS	2.07	0.86	0.441	0.054	(-0.01, 1.74)	0.856	-1.66
	QR	0.49	0.73	0.251	<0.01	(0.25, 1.22)	0.915	0.00
	TS	0.39	0.84	4.801 (MAD)	<0.001	(0.75, 1.05)	0.852	0.00
	TSS	0.65	0.72	0.711 (MAD)	<0.001	(0.60, 0.98)	0.920	-0.15

Table 31: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 50$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	-0.042
	OLS vs TS	-0.009
	OLS vs TSS	-0.054
	R vs TS	0.031
	QR vs TSS	-0.012
	TS vs TSS	-0.044
20%	OLS vs QR	0.033
	OLS vs TS	0.064
	OLS vs TSS	0.071
	R vs TS	0.032
	QR vs TSS	0.039
	TS vs TSS	0.008
30%	OLS vs QR	0.223
	OLS vs TS	0.141
	OLS vs TSS	0.148
	R vs TS	-0.105
	QR vs TSS	-0.095
	TS vs TSS	0.008
50%	OLS vs QR	-0.069
	OLS vs TS	0.004
	OLS vs TSS	-0.075
	R vs TS	0.069
	QR vs TSS	-0.005
	TS vs TSS	-0.080

Figure 10: Four regression lines are shown in each plot with $n = 50$ and outliers of 10%, 20%, 30% and 50% of $n = 50$ in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression analysis with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in Y variable only:

Table 32: Descriptive Statistics of (X, Y) in regression analysis with $n = 10$ and outliers of 10%, 20%, 30%, and 50% of $n = 100$ in Y variable only:

Outliers	Variables	n	Min	Max	Mean	Median	SD	IQR
10%	Y	110	-2.57	6.80	0.59	0.21	2.01	1.94
	X	110	-2.35	2.80	0.02	0.07	0.98	1.29
20%	Y	120	-3.18	6.78	1.00	0.22	2.43	1.60
	X	120	-2.81	2.93	-0.03	0.06	1.00	1.31
30%	Y	130	-2.72	6.84	1.12	0.15	2.80	2.38
	X	130	-3.01	2.02	-0.17	-0.05	0.98	1.06
50%	Y	150	-2.21	6.41	2.07	0.74	2.84	5.80
	X	150	-1.95	3.31	0.10	0.11	0.77	0.83

Table 33: Results from the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$,

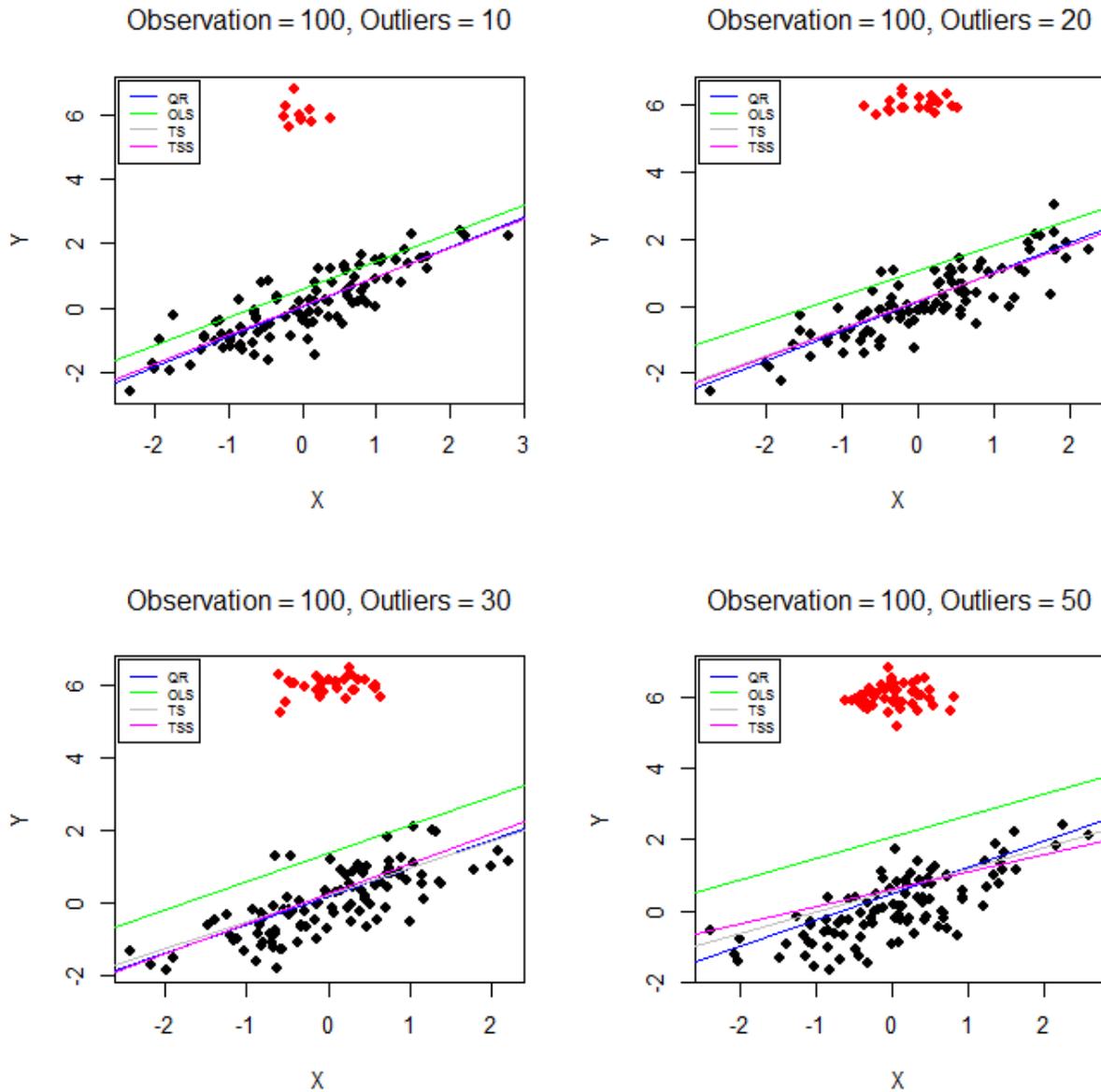
$Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
10%	OLS	0.58	0.87	0.177	<0.001	(0.52, 1.23)	0.458	-0.53
	QR	0.03	0.93	0.069	<0.001	(0.80, 1.07)	0.409	0.00
	TS	0.07	0.89	1.11 (MAD)	<0.001	(0.83, 0.91)	0.469	0.00
	TSS	0.05	0.90	0.247 (MAD)	<0.001	(0.80, 0.92)	0.469	0.02
20%	OLS	1.03	0.96	0.205	<0.001	(0.55, 1.36)	0.593	-0.81
	QR	0.23	0.75	0.088	<0.001	(0.58, 0.93)	0.517	0.00
	TS	0.23	0.75	1.52 (MAD)	<0.001	(0.73, 0.82)	0.516	0.00
	TSS	0.27	0.76	0.406 (MAD)	<0.001	(0.80, 0.92)	0.524	-0.03
30%	OLS	1.32	1.12	0.230	<0.001	(0.69, 1.60)	0.735	-1.01
	QR	0.15	0.79	0.106	<0.001	(0.58, 1.00)	0.574	0.00
	TS	0.10	0.86	2.19 (MAD)	<0.001	(0.95, 1.07)	0.587	0.00
	TSS	0.14	0.76	0.897 (MAD)	<0.001	(0.87, 1.13)	0.615	-0.03
50%	OLS	2.02	0.49	0.301	0.104	(-0.10, 1.09)	0.891	-1.47
	QR	0.48	0.71	0.153	<0.001	(0.41, 1.01)	0.885	0.00
	TS	0.50	0.60	4.677 (MAD)	<0.001	(0.35, 0.51)	0.899	-0.00
	TSS	0.57	0.44	0.732 (MAD)	<0.001	(0.23, 0.52)	0.960	0.01

Table 34: Results of Relative Median Absolute Error of the four regression procedures with outliers of 10%, 20%, 30% and 50% of $n = 100$, in Y direction only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:

Outliers	Relative Median Absolute Error	Value
10%	OLS vs QR	0.106
	OLS vs TS	-0.026
	OLS vs TSS	-0.026
	R vs TS	-0.148
	QR vs TSS	-0.147
	TS vs TSS	0.000
20%	OLS vs QR	0.128
	OLS vs TS	0.128
	OLS vs TSS	0.116
	R vs TS	0.001
	QR vs TSS	-0.014
	TS vs TSS	-0.014
30%	OLS vs QR	0.218
	OLS vs TS	0.200
	OLS vs TSS	0.163
	R vs TS	-0.023
	QR vs TSS	-0.071
	TS vs TSS	-0.047
50%	OLS vs QR	0.007
	OLS vs TS	-0.009
	OLS vs TSS	-0.078
	R vs TS	-0.017
	QR vs TSS	-0.086
	TS vs TSS	-0.067

Figure 11: Four regression lines are shown in each plot with $n = 100$ and outliers of 10%, 20%, 30% and 50% of $n = 100$ in Y only with $Nsim = 1000$, $X \sim Normal(n, 0, 1)$, $Y \sim Normal(n, 0, 1)$, and $Cor(X, Y) = 0.80$:



Regression Model under the Non-Normality Assumption:

An outcome variable Y was generated by using the log link function so for a predictor variable X , was assumed $Y \sim \text{Poisson}(\lambda)$, and $\log(\lambda) = 1 + 0.2 * X$. It was assumed X is uniformly distributed with $\min=0$ and $\max=1$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 0, 1)$, $Y \sim \text{pois}(n, \lambda = \lambda)$:

Table 35: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 0, 1)$, and $Y \sim \text{pois}(n, \lambda = \lambda)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	1.00	5.00	2.40	2.00	1.35	1.74
X	10	0.16	0.90	0.53	0.51	0.26	0.42
Y	30	0.00	6.00	2.53	2.00	1.38	1.75
X	30	0.03	0.89	0.49	0.57	0.25	0.39
Y	50	0.00	6.00	2.53	2.00	1.38	1.75
X	50	0.03	0.89	0.49	0.57	0.25	0.39
Y	100	0.00	7.00	3.11	3.00	1.64	2.00
X	100	0.01	0.99	0.50	0.52	0.28	0.49

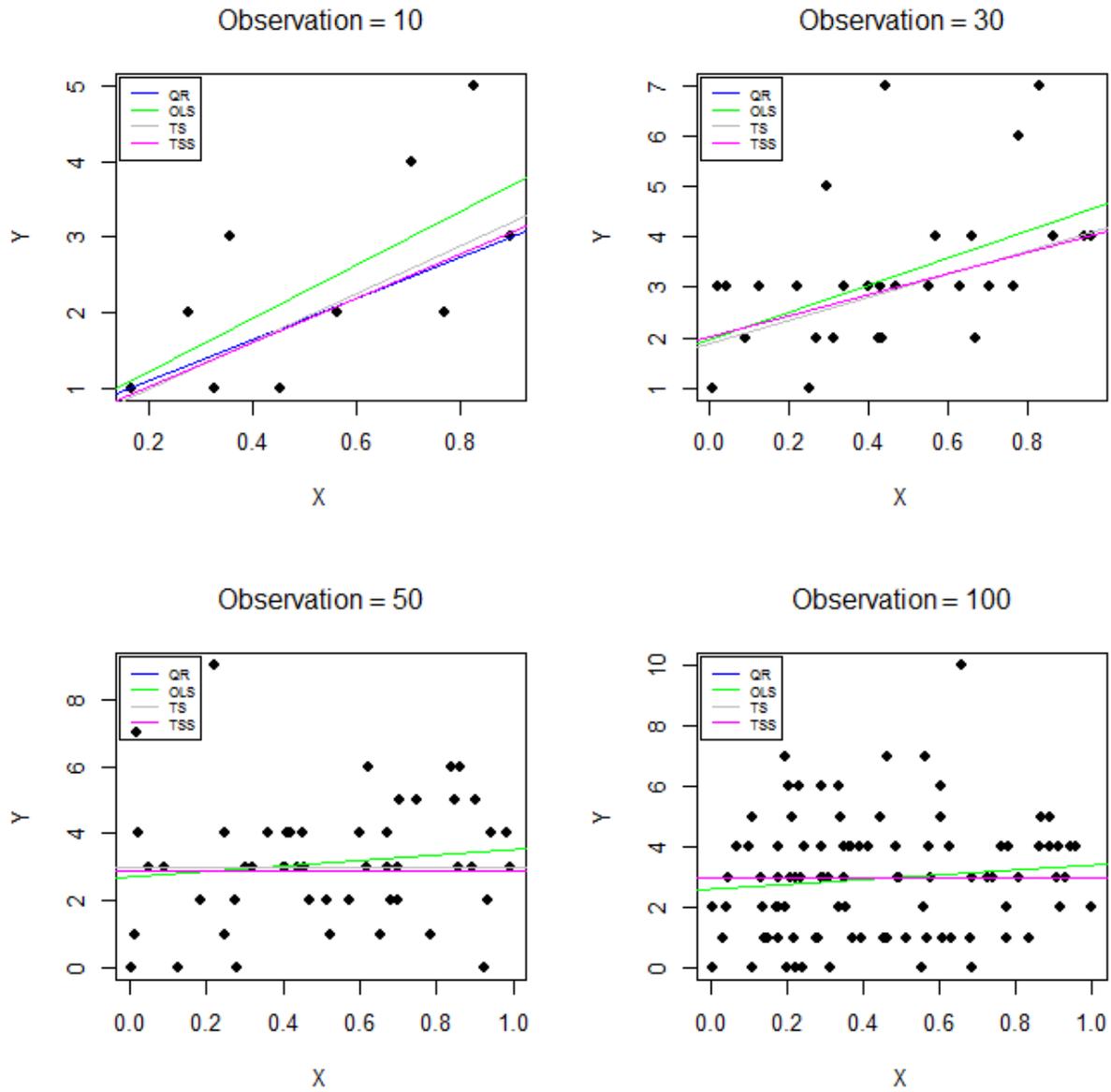
Table 36: Results from the four regression procedures with $n = 10, 30, 50,$ and $100,$ $N_{sim} = 1000,$ $X \sim Unif(n, 0, 1),$ and $Y \sim pois(n, lambda = \lambda):$

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	0.52	3.51	1.40	0.036	(0.29, 6.73)	0.808	-0.30
	QR	0.55	2.72	2.39	0.289	(0.00, 8.30)	0.671	0.00
	TS	0.36	3.15	4.66 (MAD)	<0.01	(0.21, 6.08)	0.776	0.00
	TSS	0.44	2.92	0.63 (MAD)	<0.01	(0.01, 5.91)	0.720	0.00
n=30	OLS	2.35	0.38	1.03	0.710	(-1.74, 2.49)	1.021	-0.42
	QR	2.00	0.00	1.28	1.000	(-2.46, 3.06)	1.000	0.00
	TS	2.00	0.00	7.93 (MAD)	0.080	(-2.18, 2.18)	1.000	0.00
	TSS	2.00	0.00	2.10 (MAD)	0.587	(-2.18, 2.18)	1.000	0.00
n=50	OLS	2.45	1.25	0.78	0.120	(-0.34, 2.83)	0.854	-0.22
	QR	3.00	0.00	1.05	1.000	(-4.31, 4.31)	1.000	0.00
	TS	2.00	0.00	7.99 (MAD)	0.095	(-1.59, 1.59)	1.000	0.00
	TSS	2.79	0.00	2.36 (MAD)	0.050	(-1.63, 1.63)	1.000	0.28
n=100	OLS	2.61	0.99	0.58	0.093	(-0.17, 2.15)	1.068	-0.05
	QR	3.00	0.00	1.00	1.000	(-4.05, 4.05)	1.000	0.00
	TS	3.00	0.00	8.07 (MAD)	0.072	(-1.16, 1.16)	1.000	0.00
	TSS	3.00	0.00	2.72 (MAD)	0.063	(-1.16, 1.16)	1.000	0.00

Table 37: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100, Nsim = 1000, X \sim Unif(n, 0, 1), Y \sim pois(n, lambda = \lambda)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.618
	OLS vs TS	0.039
	OLS vs TSS	0.109
	QR vs TS	0.155
	QR vs TSS	0.071
	TS vs TSS	0.072
n=30	OLS vs QR	0.021
	OLS vs TS	0.021
	OLS vs TSS	0.021
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000
n=50	OLS vs QR	-0.172
	OLS vs TS	-0.172
	OLS vs TSS	-0.172
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000
n=100	OLS vs QR	-0.063
	OLS vs TS	-0.063
	OLS vs TSS	-0.063
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000

Figure 12: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 0, 1)$, $Y \sim pois(n, \lambda = \lambda)$:



Regression Model under the Micceri distributions:

Eight Micceri distributions were used to generate a random sample for an outcome variable Y, with a uniform distribution of predictor variable X with min=0 and max=1.

1. Smooth Symmetric distribution:

An outcome variable Y was generated from a Smooth Symmetric distribution and a predictor variable X was generated from uniform distributed with min=1 and max=10. A Monte Carlo Simulation was conducted with Nsim=1000 repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Smooth\ Symmetric(n)$:

Table 38: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth\ Symmetric(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	6.00	18.00	13.80	14.50	3.46	3.25
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	4.00	20.00	13.13	14.00	3.97	5.75
X	30	1.26	8.97	5.42	6.10	2.27	3.54
Y	50	4.00	23.00	14.08	14.00	4.75	7.00
X	50	1.11	9.97	5.81	6.06	2.55	4.01
Y	100	3.00	25.00	13.10	13.00	5.02	8.00
X	100	1.09	9.96	5.25	5.69	2.52	4.40

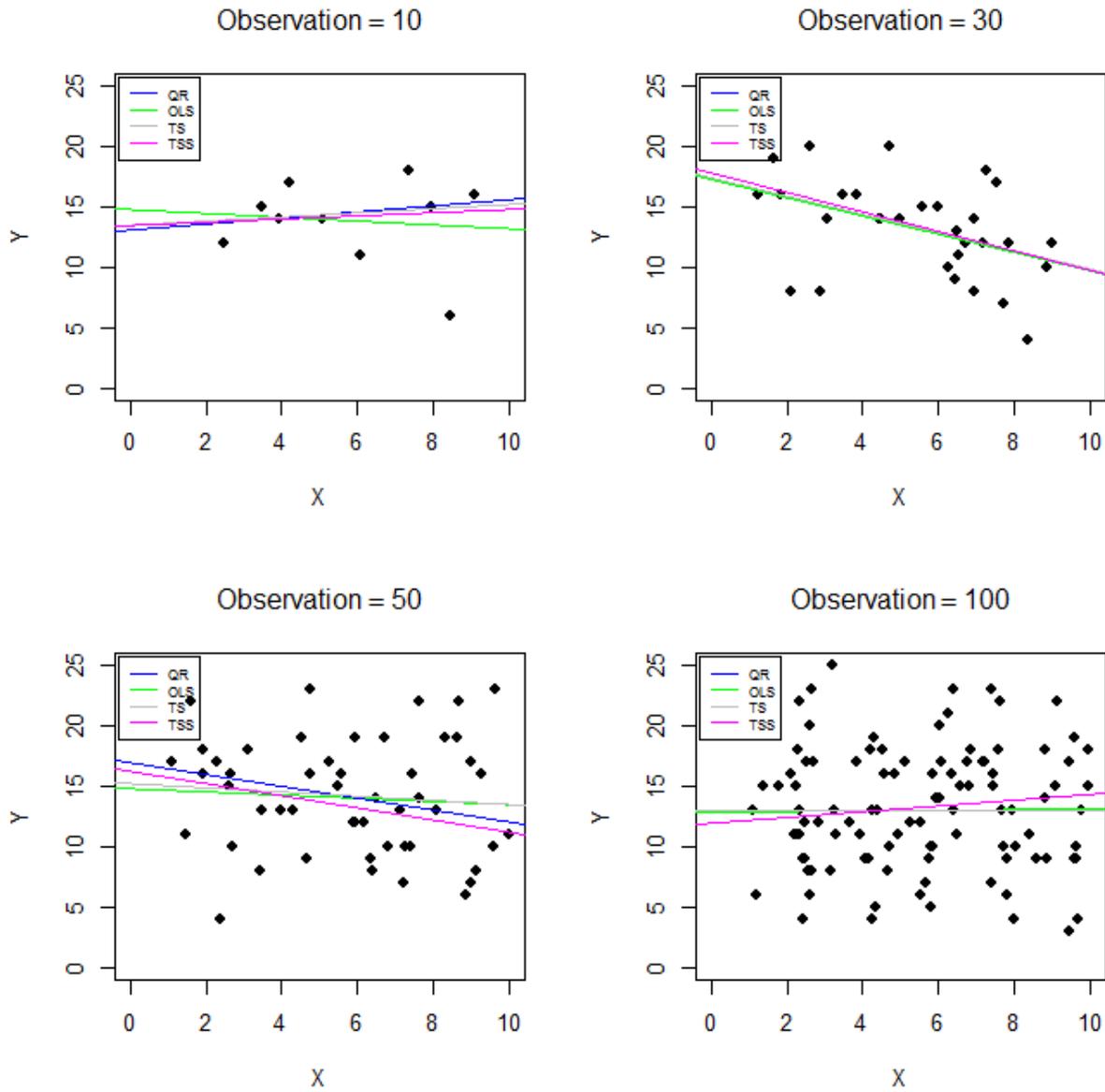
Table 39: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth\ Symmetric(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	14.70	-0.15	0.529	0.777	(-1.37, 1.06)	2.366	0.46
	QR	13.00	0.25	0.988	0.806	(-2.38, 1.23)	1.373	0.00
	TS	13.44	0.18	2.186 (MAD)	0.748	(-0.90, 1.26)	1.409	0.00
	TSS	13.50	0.13	0.521 (MAD)	0.833	(-0.94, 1.19)	1.586	0.25
n=30	OLS	17.18	-0.75	0.298	0.019	(-1.36, -0.13)	1.651	0.17
	QR	17.38	-0.75	0.337	0.034	(-1.63, -0.39)	1.653	0.00
	TS	17.41	-0.76	2.159 (MAD)	<0.001	(-1.34, -0.17)	1.654	0.00
	TSS	17.78	-0.81	0.511 (MAD)	<0.001	(-1.40, -0.23)	1.684	-0.04
n=50	OLS	14.89	-0.14	0.268	0.602	(-0.68, 0.40)	3.542	0.09
	QR	16.94	-0.49	0.390	0.217	(-1.24, 0.26)	3.376	0.00
	TS	15.19	-0.17	2.683 (MAD)	0.165	(-0.70, 0.36)	3.449	0.00
	TSS	16.30	-0.50	0.744 (MAD)	0.099	(-1.04, 0.04)	3.344	0.71
n=100	OLS	12.91	0.03	0.195	0.863	(-0.35, 0.42)	3.915	-0.06
	QR	13.00	0.00	0.504	1.000	(-0.99, 0.99)	4.000	0.00
	TS	13.00	0.00	2.471 (MAD)	0.251	(0.00, 0.14)	4.000	0.00
	TSS	11.96	0.24	0.828 (MAD)	0.135	(-0.16, 0.63)	3.691	-0.26

Table 40: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Smooth\ Symmetric(n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.419
	OLS vs TS	0.404
	OLS vs TSS	0.329
	QR vs TS	-0.027
	QR vs TSS	-0.155
	TS vs TSS	-0.125
n=30	OLS vs QR	0.000
	OLS vs TS	-0.001
	OLS vs TSS	-0.019
	QR vs TS	0.000
	QR vs TSS	-0.019
	TS vs TSS	-0.018
n=50	OLS vs QR	0.047
	OLS vs TS	0.026
	OLS vs TSS	0.055
	QR vs TS	-0.022
	QR vs TSS	0.009
	TS vs TSS	0.031
n=100	OLS vs QR	-0.012
	OLS vs TS	-0.012
	OLS vs TSS	0.067
	QR vs TS	0.000
	QR vs TSS	0.077
	TS vs TSS	0.077

Figure 13: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Smooth\ Symmetric(n)$:



2. Extreme Asymmetric distribution:

An outcome variable Y was generated from an Extreme Asymmetric distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, $Y \sim \text{Extreme Asymmetric}(n)$:

Table 41: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetric}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	17.00	30.00	27.00	28.50	3.50	4.03
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	14.00	30.00	26.13	27.00	4.00	5.00
X	30	1.26	8.98	5.43	6.10	2.27	3.55
Y	50	11.00	30.00	24.54	26.00	5.23	5.00
X	50	1.11	9.97	5.81	6.06	2.56	4.00
Y	100	8.00	30.00	24.21	26.00	5.74	5.25
X	100	1.09	9.96	5.52	5.69	2.52	4.40

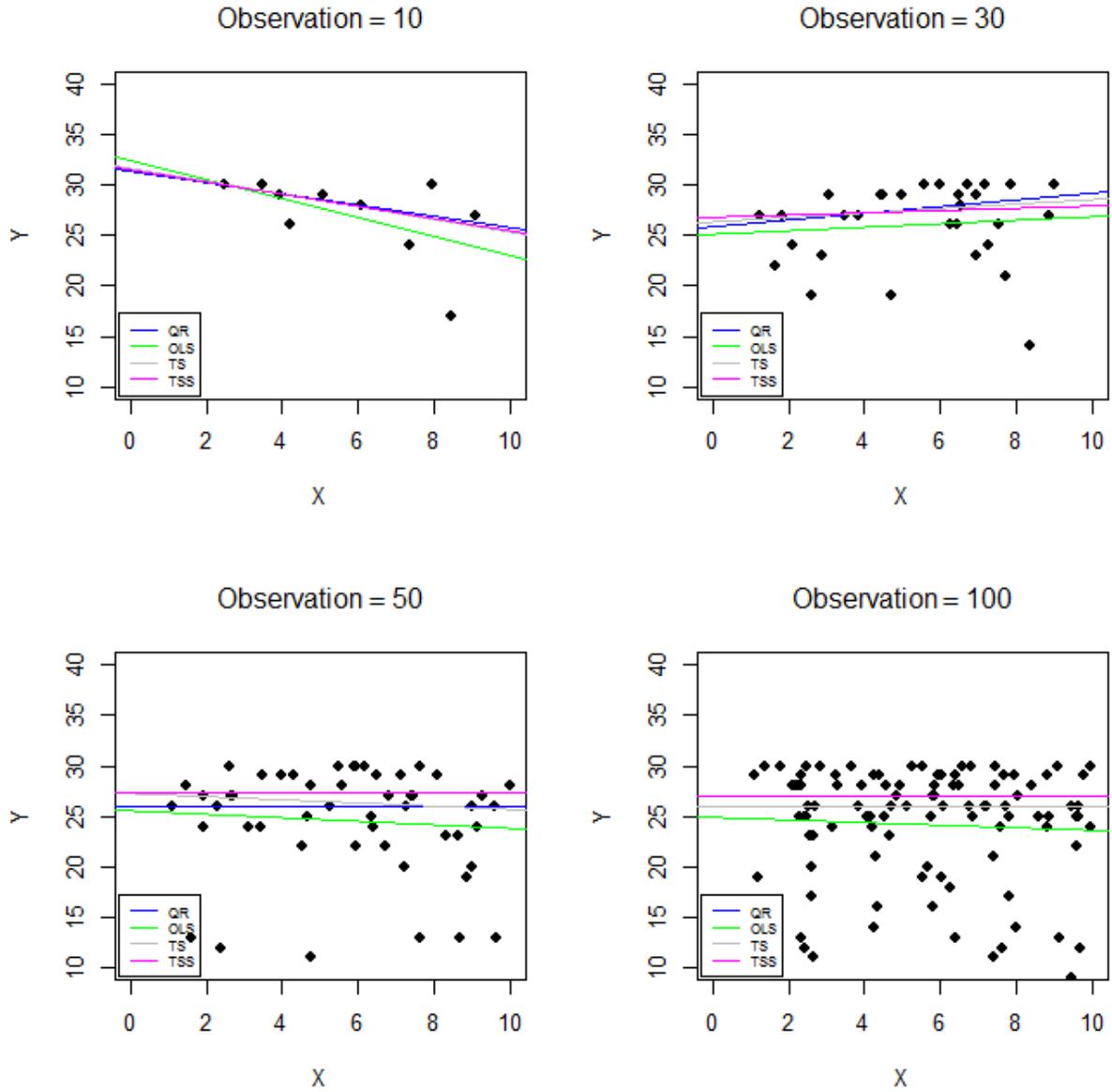
Table 42: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	32.46	-0.94	0.522	0.109	(-2.14, 0.26)	1.428	0.53
	QR	31.38	-0.56	0.752	0.479	(-2.67, 0.25)	0.627	0.00
	TS	31.71	-0.63	1.457 (MAD)	0.002	(-1.73, -0.01)	0.780	0.00
	TSS	31.46	-0.59	0.208 (MAD)	0.008	(-1.69, -0.01)	0.706	0.09
n=30	OLS	25.19	0.17	0.332	0.604	(-0.51, 0.85)	1.914	1.35
	QR	25.87	0.33	0.391	0.412	(-0.34, 1.09)	1.926	0.00
	TS	26.35	0.23	2.111 (MAD)	0.305	(-0.47, 0.92)	1.896	0.00
	TSS	26.79	0.11	0.518 (MAD)	0.173	(-0.58, 0.79)	2.143	-0.07
n=50	OLS	25.61	-0.18	0.298	0.539	(-0.78, 0.42)	2.737	1.73
	QR	26.00	0.00	0.243	1.000	(-0.78, 0.28)	3.000	0.00
	TS	27.22	-1.57	2.224 (MAD)	0.002	(-0.78, -0.01)	2.709	0.00
	TSS	27.38	0.00	0.566 (MAD)	0.033	(-0.67, -0.01)	3.000	-1.38
n=100	OLS	24.95	-1.35	0.229	0.557	(-0.59, 0.32)	2.716	1.70
	QR	26.00	0.00	0.220	1.000	(-0.51, 0.21)	2,667	0.00
	TS	26.00	0.00	2.224 (MAD)	0.101	(-0.47, -0.47)	3.000	0.00
	TSS	26.96	0.00	0.070 (MAD)	0.183	(-0.50, 0.50)	3.000	-0.96

Table 43: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.560
	OLS vs TS	0.454
	OLS vs TSS	0.505
	QR vs TS	-0.243
	QR vs TSS	-0.126
	TS vs TSS	0.095
n=30	OLS vs QR	-0.006
	OLS vs TS	-0.009
	OLS vs TSS	-0.119
	QR vs TS	0.016
	QR vs TSS	-0.011
	TS vs TSS	-0.130
n=50	OLS vs QR	-0.096
	OLS vs TS	0.009
	OLS vs TSS	-0.096
	QR vs TS	0.096
	QR vs TSS	0.000
	TS vs TSS	-0.107
n=100	OLS vs QR	-0.104
	OLS vs TS	-0.104
	OLS vs TSS	-0.104
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000

Figure 14: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme Asymmetric(n)$:



3. Extreme Bimodal distribution:

An outcome variable Y was generated from an Extreme Bimodal distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, $Y \sim \text{Extreme Bimodal}(n)$:

Table 44: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Extreme Bimodal}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	1.00	5.00	2.60	1.50	1.90	3.75
X	10	2.48	9.09	5.81	5.57	2.30	3.77
Y	30	1.00	5.00	3.03	4.00	1.80	4.00
X	30	1.26	8.97	5.42	6.10	2.27	3.55
Y	50	1.00	5.00	3.50	4.00	1.58	2.75
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	1.00	5.00	3.19	4.00	1.80	4.00
X	100	1.09	9.96	5.23	5.69	2.52	4.40

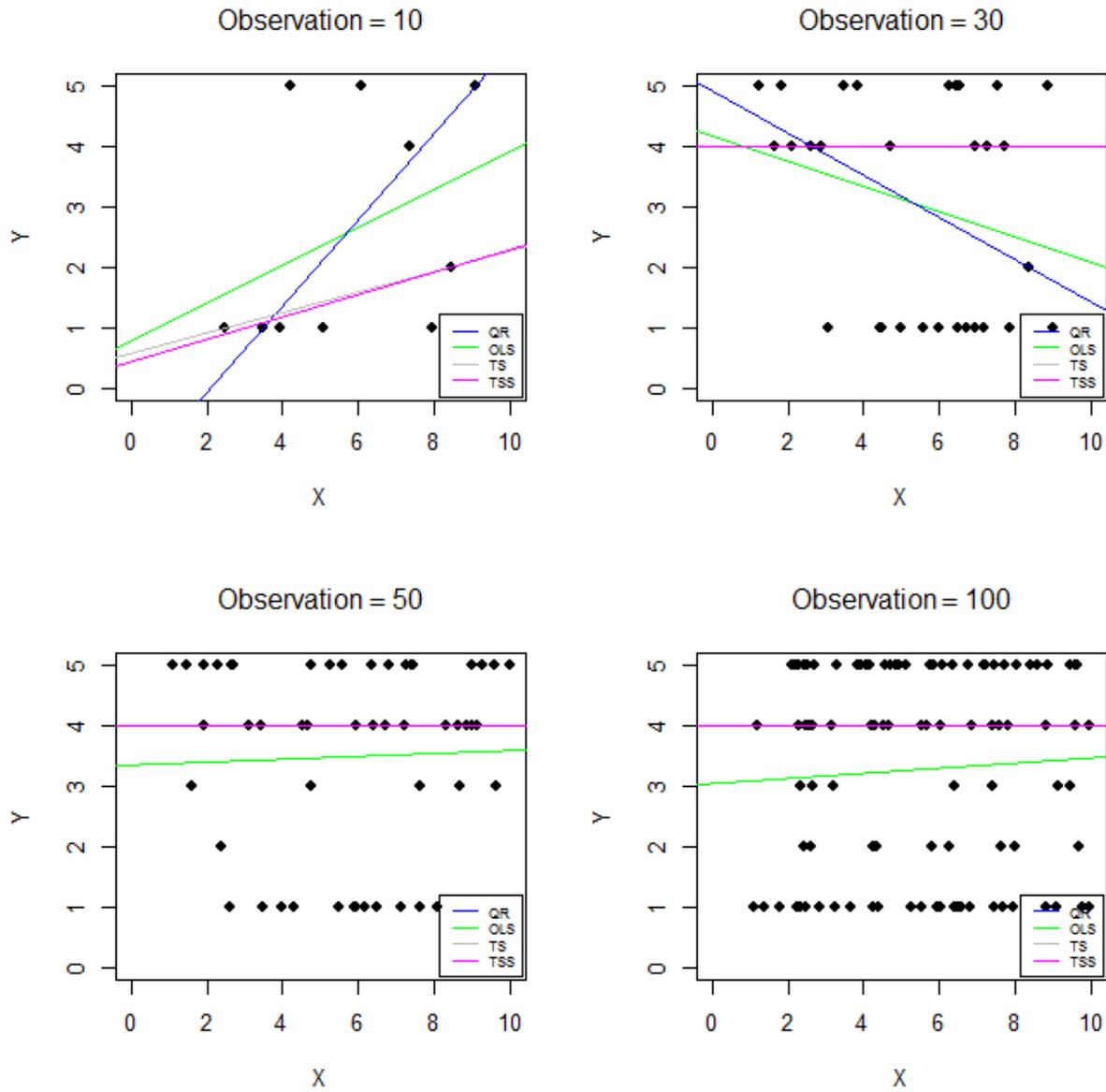
Table 45: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme\ Bimodal(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	0.77	0.32	0.260	0.276	(-0.31, 0.94)	1.138	-0.71
	QR	-1.51	0.71	0.403	0.114	(-0.71, 0.78)	0.928	0.00
	TS	0.58	0.17	0.904 (MAD)	0.060	(-0.46, 0.80)	0.676	0.00
	TSS	0.44	0.19	0.001 (MAD)	0.035	(-0.45, 0.00)	0.693	0.05
n=30	OLS	4.16	-0.21	0.143	0.158	(-0.50, 0.09)	1.830	0.33
	QR	4.91	-0.35	0.255	0.183	(-0.66, 0.15)	1.538	0.00
	TS	4.00	0.00	1.027 (MAD)	0.005	(-0.33, 0.00)	1.000	0.00
	TSS	4.00	0.00	0.001 (MAD)	0.025	(-0.33, 0.00)	1.000	0.00
n=50	OLS	3.36	0.02	0.089	0.801	(-0.16, 0.20)	1.013	0.48
	QR	4.00	0.00	0.088	1.000	(-0.19, 0.25)	1.000	0.00
	TS	4.00	0.00	0.860 (MAD)	0.036	(-0.18, 0.00)	1.000	0.00
	TSS	4.00	0.00	0.001 (MAD)	0.175	(-0.18, 0.18)	1.000	0.00
n=100	OLS	3.05	0.04	0.066	0.525	(-0.09, 0.17)	1.100	0.68
	QR	4.00	0.00	0.108	1.000	(-0.48, 0.48)	1.000	0.00
	TS	4.00	0.00	0.860 (MAD)	0.041	(-0.14, 0.00)	1.000	0.00
	TSS	4.00	0.00	0.001 (MAD)	0.233	(-0.14, 0.14)	1.000	0.00

Table 46: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Extreme\ Bimodal(n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.185
	OLS vs TS	0.406
	OLS vs TSS	0.390
	QR vs TS	0.272
	QR vs TSS	0.252
	TS vs TSS	-0.026
n=30	OLS vs QR	0.015
	OLS vs TS	0.453
	OLS vs TSS	0.453
	QR vs TS	0.349
	QR vs TSS	0.349
	TS vs TSS	0.000
n=50	OLS vs QR	0.014
	OLS vs TS	0.014
	OLS vs TSS	0.014
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000
n=100	OLS vs QR	0.088
	OLS vs TS	0.088
	OLS vs TSS	0.088
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000

Figure 15: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim \text{Extreme Bimodal}(n)$:



4. Mass at Zero distribution:

An outcome variable Y was generated from Mass at Zero distribution and a predictor variable X were generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, $Y \sim \text{Mass at Zero}(n)$:

Table 47: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Mass at Zero}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	9.00	20.00	12.70	11.50	3.40	2.50
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	6.00	20.00	13.03	12.50	3.43	4.00
X	30	2.48	9.09	5.81	5.57	2.30	3.78
Y	50	0.00	21.00	12.82	14.00	4.85	5.00
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	0.00	23.00	13.12	13.00	4.54	4.40
X	100	1.09	9.96	5.53	5.69	2.52	4.40

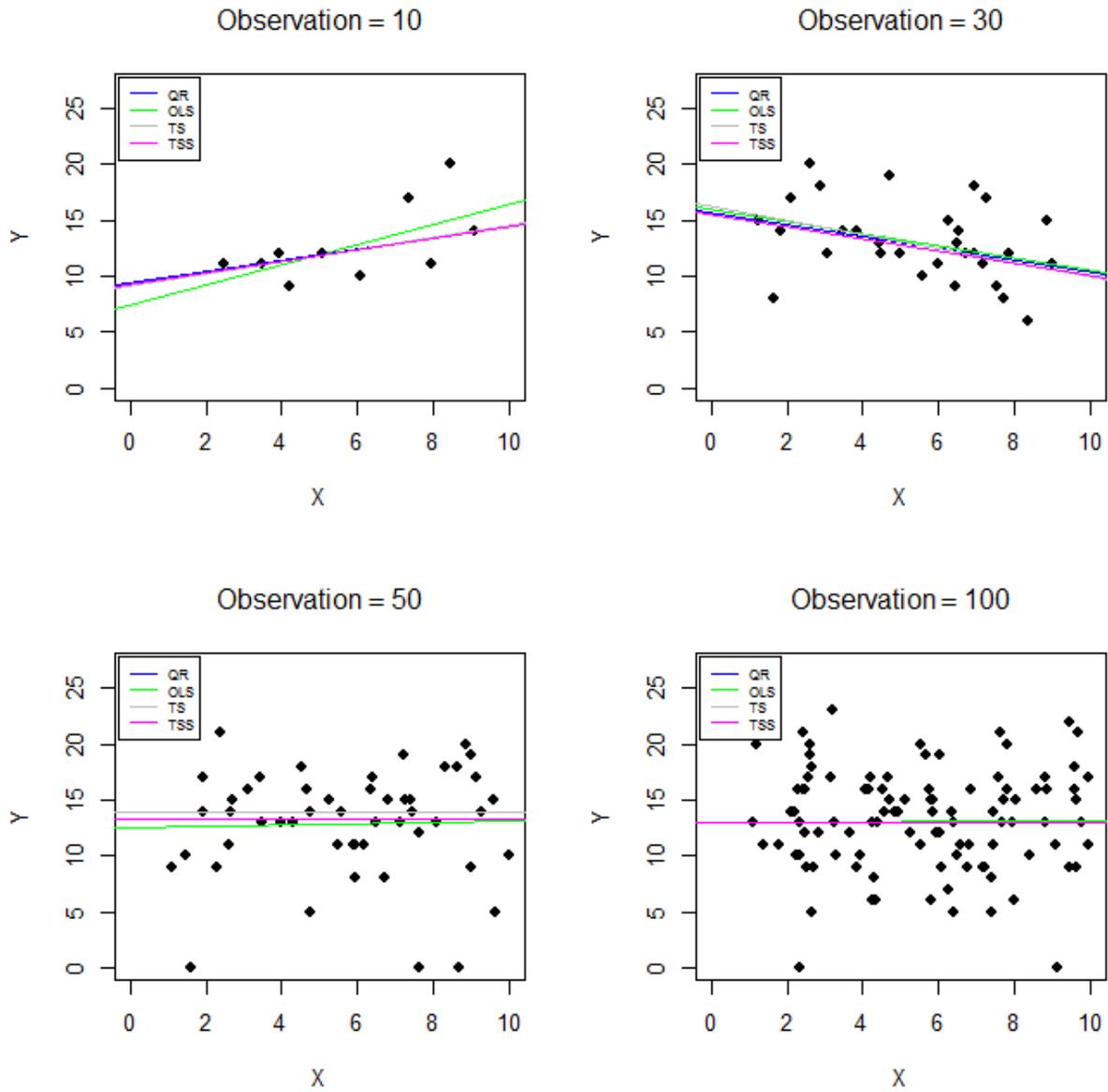
Table 48: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass\ at\ Zero(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	7.44	0.91	0.413	0.060	(-0.05, 1.86)	2.140	0.18
	QR	9.47	0.50	0.363	0.206	(-0.21, 1.21)	1.493	0.00
	TS	9.19	0.54	1.444 (MAD)	0.015	(0.32, 1.39)	1.570	0.00
	TSS	9.29	0.52	4.488 (MAD)	0.009	(0.35, 1.38)	1.530	0.05
n=30	OLS	15.94	-0.54	0.266	0.054	(-1.08, 0.01)	1.930	-0.24
	QR	15.66	-0.53	0.363	0.031	(-0.98, -0.07)	1.919	0.00
	TS	16.16	-0.62	1.950 (MAD)	<0.001	(-1.15, -0.10)	2.200	0.00
	TSS	15.55	-0.55	0.373 (MAD)	0.002	(-1.08, -0.02)	1.961	0.22
n=50	OLS	12.49	0.06	0.274	0.835	(-0.49, 0.61)	2.862	1.04
	QR	14.00	0.00	0.363	1.000	(-0.72, 0.72)	3.000	0.00
	TS	14.00	0.00	2.406 (MAD)	0.521	(-0.55, 0.55)	3.000	0.00
	TSS	13.24	0.00	0.659 (MAD)	0.617	(-0.54, 0.54)	3.000	-0.05
n=100	OLS	13.05	0.01	0.182	0.943	(-0.35, 0.37)	2.994	-0.08
	QR	13.00	0.00	0.363	1.000	(-0.72, 0.72)	3.000	0.00
	TS	13.00	0.00	2.540 (MAD)	0.995	(-0.36, 0.36)	3.000	0.00
	TSS	13.00	0.00	0.850 (MAD)	0.967	(-0.36, 0.36)	3.000	-0.05

Table 49: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass \text{ at Zero } (n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.302
	OLS vs TS	0.267
	OLS vs TSS	0.285
	QR vs TS	-0.049
	QR vs TSS	-0.024
	TS vs TSS	0.024
n=30	OLS vs QR	0.006
	OLS vs TS	-0.047
	OLS vs TSS	-0.016
	QR vs TS	-0.054
	QR vs TSS	-0.022
	TS vs TSS	0.030
n=50	OLS vs QR	-0.048
	OLS vs TS	-0.048
	OLS vs TSS	-0.048
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000
n=100	OLS vs QR	-0.002
	OLS vs TS	-0.002
	OLS vs TSS	-0.002
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000

Figure 16: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim \text{Mass at Zero}(n)$:



5. Mass at Zero with Gap distribution:

An outcome variable Y was generated from Mass at Zero with Gap distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, $Y \sim \text{Mass at Zero with Gap}(n)$:

Table 50: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Mass at Zero with Gap}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	0.00	9.00	0.90	0.00	2.85	0.00
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	0.00	10.00	0.93	0.00	2.85	0.00
X	30	1.26	8.98	5.43	6.10	2.27	3.55
Y	50	0.00	9.00	1.14	0.00	2.85	0.00
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	0.00	11.00	1.92	0.00	3.77	0.00
X	100	1.09	9.96	5.53	5.69	2.52	4.40

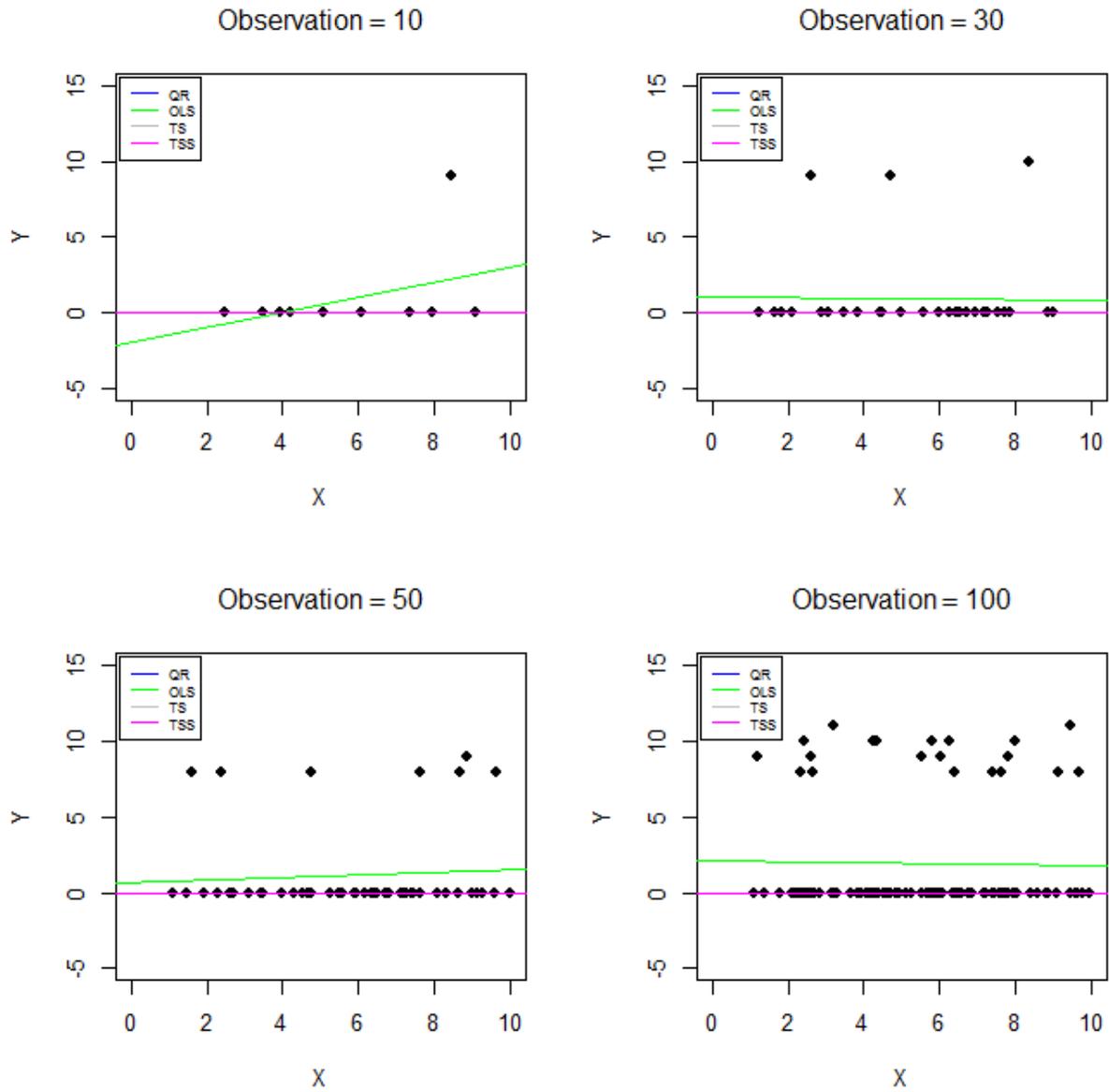
Table 51: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass \text{ at Zero with Gap } (n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	-1.99	0.50	0.401	0.250	(-0.43, 1.42)	0.889	-0.32
	QR	0.00	0.00	NA	NA	NA	0.000	0.00
	TS	0.00	0.00	0.000 (MAD)	0.097	(-0.90, 0.09)	0.000	0.00
	TSS	0.00	0.00	0.000 (MAD)	1.000	(-0.90, 0.09)	0.000	0.000
n=30	OLS	1.03	-0.02	0.240	0.941	(-0.50, 0.47)	0.030	-0.91
	QR	0.00	0.00	NA	NA	NA	0.000	0.00
	TS	0.00	0.00	0.000 (MAD)	0.691	(-0.49, 0.49)	0.000	0.00
	TSS	0.00	0.00	0.000 (MAD)	1.000	(-0.49, 0.49)	0.000	0.00
n=50	OLS	0.76	0.08	0.161	0.621	(-0.24, 0.40)	0.182	-1.17
	QR	0.00	0.00	NA	NA	NA	0.000	0.00
	TS	0.00	0.00	0.000 (MAD)	0.691	(-0.34, 0.34)	0.000	0.00
	TSS	0.00	0.00	0.000 (MAD)	0.205	(-0.34, 0.34)	0.000	0.000
n=100	OLS	2.08	-0.03	0.151	0.851	(-0.33, 0.27)	0.082	-1.89
	QR	0.00	0.00	NA	NA	NA	0.000	0.00
	TS	0.00	0.00	0.000 (MAD)	0.789	(-0.34, 0.34)	0.000	0.00
	TSS	0.00	0.00	0.000 (MAD)	0.562	(-0.33, 0.33)	0.000	0.000

Table 52: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Mass$ at Zerowith Gap (n):

	Relative Median Absolute Error	Value
n=10	OLS vs QR	1.000
	OLS vs TS	1.000
	OLS vs TSS	1.000
	QR vs TS	NA
	QR vs TSS	NA
	TS vs TSS	NA
n=30	OLS vs QR	1.000
	OLS vs TS	1.000
	OLS vs TSS	1.000
	QR vs TS	NA
	QR vs TSS	NA
	TS vs TSS	NA
n=50	OLS vs QR	1.000
	OLS vs TS	1.000
	OLS vs TSS	1.000
	QR vs TS	NA
	QR vs TSS	NA
	TS vs TSS	NA
n=100	OLS vs QR	1.000
	OLS vs TS	1.000
	OLS vs TSS	1.000
	QR vs TS	NA
	QR vs TSS	NA
	TS vs TSS	NA

Figure 17: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim \text{Mass at Zero with Gap}(n)$:



6. Multimodal Lumpy distribution:

An outcome variable Y was generated from Multimodal Lumpy distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, $Y \sim \text{Multimodal Lumpy}(n)$:

Table 53: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Multimodal Lumpy}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	1.00	40.00	19.90	16.50	12.88	17.75
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	5.00	41.00	18.10	15.50	9.77	9.50
X	30	1.26	8.98	5.43	6.10	2.27	3.55
Y	50	4.00	42.00	23.54	21.50	12.10	23.25
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	0.00	43.00	22.60	23.00	11.80	20.00
X	100	1.09	9.96	5.53	5.69	2.52	4.40

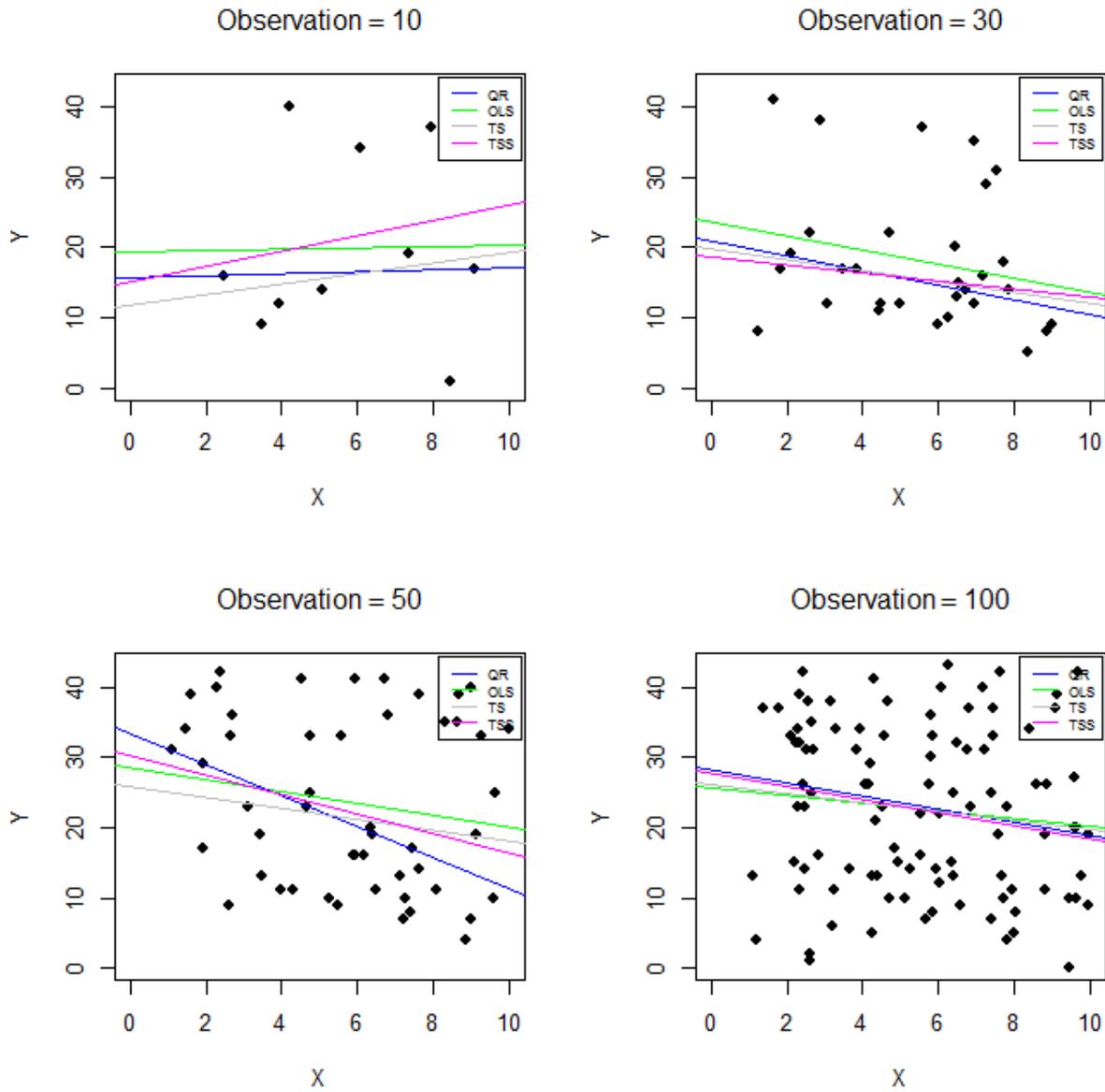
Table 54: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Multimodal\ Lumpy(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	19.35	0.09	1.978	0.963	(-4.47, 4.66)	5.806	-3.40
	QR	15.62	0.51	3.385	0.965	(-6.48, 6.79)	5.688	0.00
	TS	11.85	0.75	8.100 (MAD)	0.647	(-3.33, 4.83)	4.138	0.00
	TSS	15.05	1.11	1.420 (MAD)	0.359	(-2.87, 5.08)	4.064	-5.44
n=30	OLS	23.56	-1.01	0.789	0.213	(-2.62, 0.61)	4.400	-2.77
	QR	21.00	-1.04	0.763	0.182	(-2.54, 0.45)	4.390	0.00
	TS	19.79	-0.76	5.662 (MAD)	0.024	(-2.37, 0.84)	4.690	0.00
	TSS	18.70	-0.58	1.195 (MAD)	0.035	(-2.19, 1.03)	4.868	-0.007
n=50	OLS	28.37	-0.83	0.674	0.223	(-2.19, 0.52)	11.178	-2.28
	QR	33.24	-2.19	0.931	0.023	(-4.01, -0.37)	9.573	0.00
	TS	25.77	-0.78	6.702 (MAD)	0.003	(-2.13, 0.57)	11.167	0.00
	TSS	30.27	-1.39	2.422 (MAD)	0.004	(-2.73, -0.05)	10.512	-1.12
n=100	OLS	25.60	-0.54	0.469	0.251	(-1.47, 0.39)	10.060	-0.23
	QR	28.25	-0.93	0.825	0.263	(-2.55, 0.69)	9.555	0.00
	TS	26.11	-0.64	6.857 (MAD)	<0.001	(-1.56, 0.28)	10.009	0.00
	TSS	27.79	-0.93	2.278 (MAD)	0.002	(-1.85, -0.01)	9.555	0.47

Table 55: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Multimodal\ Lumpy(n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.020
	OLS vs TS	0.287
	OLS vs TSS	0.300
	QR vs TS	0.272
	QR vs TSS	0.286
	TS vs TSS	0.018
n=30	OLS vs QR	0.001
	OLS vs TS	-0.066
	OLS vs TSS	-0.107
	QR vs TS	-0.067
	QR vs TSS	-0.108
	TS vs TSS	-0.037
n=50	OLS vs QR	0.143
	OLS vs TS	0.001
	OLS vs TSS	0.059
	QR vs TS	-0.166
	QR vs TSS	-0.098
	TS vs TSS	0.059
n=100	OLS vs QR	0.052
	OLS vs TS	0.005
	OLS vs TSS	0.050
	QR vs TS	-0.047
	QR vs TSS	-0.000
	TS vs TSS	0.045

Figure 18: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Multimodal Lumpy(n)$:



7. Extreme Asymmetry – Decay distribution:

An outcome variable Y was generated from Extreme Asymmetry – Decay distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$,

$Y \sim \text{Extreme Asymmetry - Decay}(n)$:

Table 56: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$,

$N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry - Decay}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	10.00	30.00	12.80	10.00	6.42	0.75
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	10.00	30.00	12.93	10.00	5.67	2.75
X	30	1.26	8.97	5.43	6.10	2.27	3.55
Y	50	10.00	30.00	13.88	11.00	6.09	4.00
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	10.00	30.00	14.07	11.00	6.05	6.25
X	100	1.09	9.96	5.53	5.69	2.52	4.40

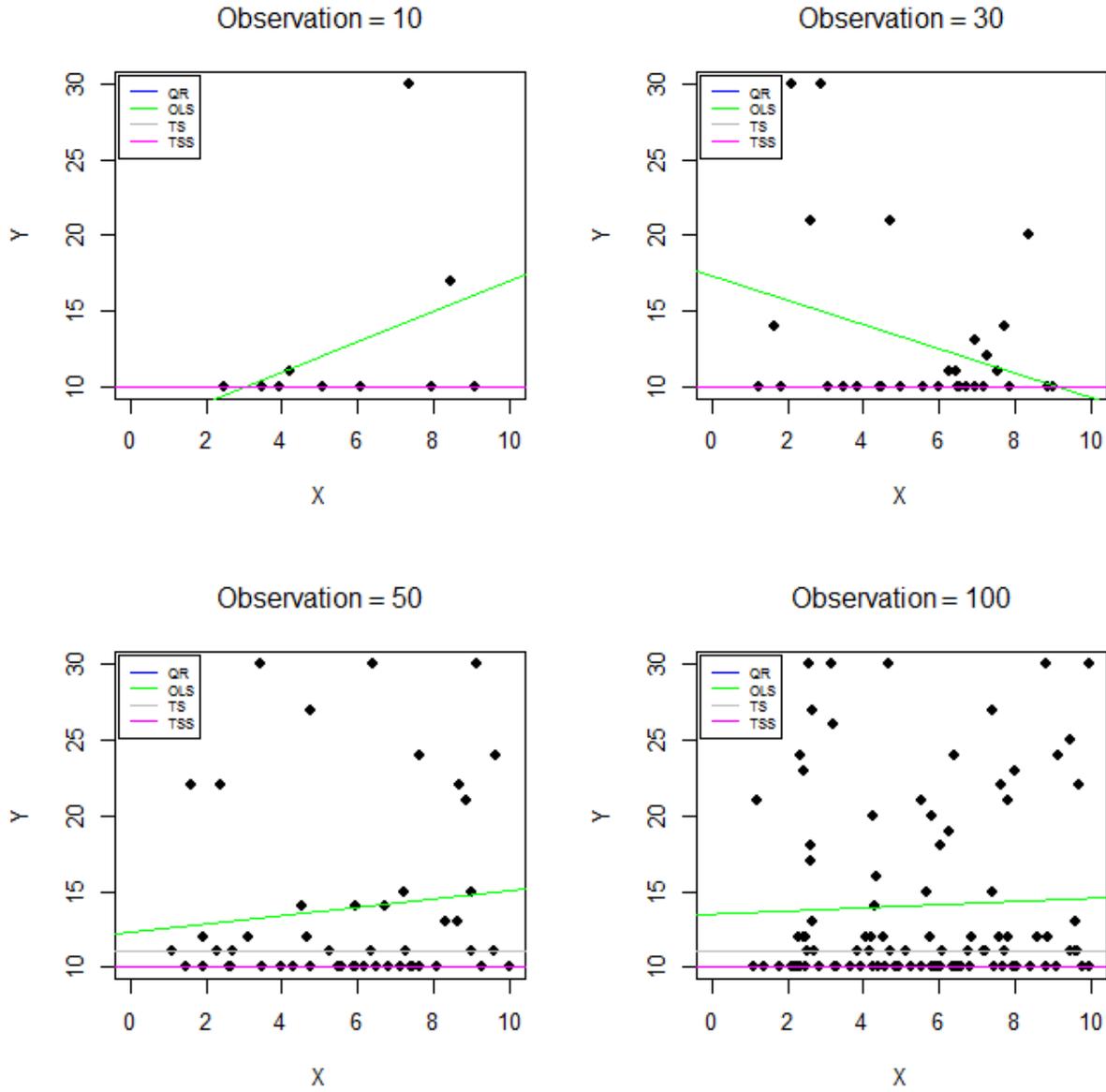
Table 57: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry} - \text{Decay}(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	6.98	1.00	0.922	0.309	(-1.12, 3.13)	1.817	-0.71
	QR	10.00	0.00	1.096	1.000	(-0.20, 5.18)	0.000	0.00
	TS	10.00	0.00	0.397 (MAD)	0.119	(-2.13, 2.13)	0.000	0.00
	TSS	10.00	0.00	0.000 (MAD)	0.181	(-2.13, 2.13)	0.000	0.05
n=30	OLS	17.33	-0.81	0.446	0.080	(-1.72, 0.10)	1.817	-1.61
	QR	10.00	0.00	0.461	1.000	(-1.38, 0.23)	0.000	0.00
	TS	10.00	0.00	1.189 (MAD)	0.011	(-1.04, 0.00)	0.000	0.00
	TSS	10.00	0.00	0.000 (MAD)	0.415	(-1.04, 1.04)	0.000	0.05
n=50	OLS	12.29	0.27	0.340	0.429	(-0.42, 0.96)	1.334	-2.77
	QR	10.00	0.00	0.231	1.000	(-0.18, 0.63)	1.000	0.00
	TS	11.00	0.00	1.876 (MAD)	<0.001	(-0.75, -0.01)	1.000	0.00
	TSS	10.00	0.00	0.000 (MAD)	0.207	(-0.80, 0.80)	1.000	1.00
n=100	OLS	13.48	0.11	0.242	0.662	(-0.37, 0.59)	1.012	-3.16
	QR	11.00	0.00	0.155	1.000	(-0.18, 0.31)	1.000	0.00
	TS	11.00	0.00	2.083 (MAD)	<0.01	(-0.53, -0.01)	1.000	0.00
	TSS	10.00	0.00	0.000 (MAD)	0.398	(-0.57, 0.57)	1.000	1.00

Table 58: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim \text{Extreme Asymmetry} - \text{Decay}(n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	1.000
	OLS vs TS	1.000
	OLS vs TSS	1.000
	QR vs TS	NA
	QR vs TSS	NA
	TS vs TSS	NA
n=30	OLS vs QR	0.001
	OLS vs TS	-0.066
	OLS vs TSS	-0.107
	QR vs TS	-0.067
	QR vs TSS	-0.108
	TS vs TSS	-0.037
n=50	OLS vs QR	0.251
	OLS vs TS	0.251
	OLS vs TSS	0.251
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000
n=100	OLS vs QR	0.013
	OLS vs TS	0.013
	OLS vs TSS	0.013
	QR vs TS	0.000
	QR vs TSS	0.000
	TS vs TSS	0.000

Figure 19: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim \text{Extreme Asymmetry - Decay}(n)$:



8. Digit Preference distribution:

An outcome variable Y was generated from Digit Preference distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions.

Regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$,

$Y \sim \text{Digit Preference}(n)$:

Table 59: Descriptive Statistics of (X, Y) in regression analysis with $n = 10, 30, 50, 100$, $N_{\text{sim}} = 1000$, $X \sim \text{Unif}(n, 1, 10)$, and $Y \sim \text{Digit Preference}(n)$:

Variables	n	Min	Max	Mean	Median	SD	IQR
Y	10	495.0	595.0	540.5	545.0	30.95	37.5
X	10	2.48	9.09	5.81	5.57	2.30	3.78
Y	30	470.0	590.0	539.7	547.5	32.32	53.75
X	30	1.26	8.98	5.43	6.10	2.27	3.55
Y	50	460.0	610.0	534.7	532.5	36.82	53.75
X	50	1.11	9.97	5.81	6.06	2.55	4.00
Y	100	450.0	615.0	528.5	530.0	37.15	56.25
X	100	1.09	9.96	5.53	5.69	2.52	4.40

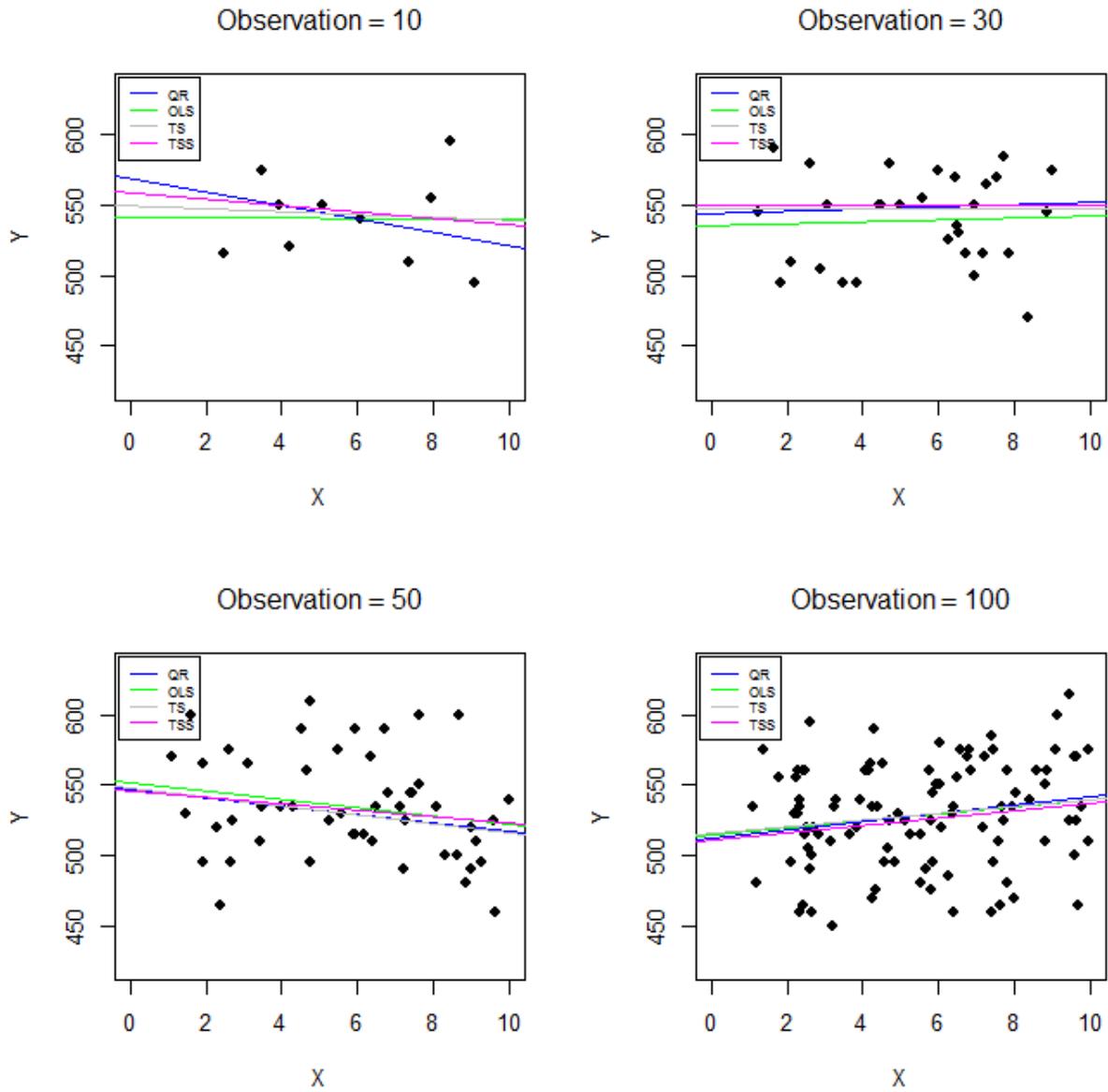
Table 60: Results of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Digit Preference(n)$:

	Regression	$\hat{\beta}_0$	$\hat{\beta}_1$	S.E($\hat{\beta}_1$)	P-value	95% CI for β_1	MEDAE	Median Bias
n=10	OLS	541.0	-0.19	4.76	0.968	(-11.0, 10.8)	27.43	4.34
	QR	568.7	-4.72	9.45	0.631	(-14.0, 13.4)	23.87	0.00
	TS	550.1	-1.03	18.79 (MAD)	0.695	(-10.4, 8.4)	27.13	0.00
	TSS	559.3	-2.36	11.84 (MAD)	0.555	(11.9, 7.2)	26.66	-2.50
n=30	OLS	551.8	-2.95	2.04	0.154	(-7.05, 1.15)	23.41	-5.17
	QR	546.8	-2.98	2.98	0.321	(-9.5, 1.98)	23.38	0.00
	TS	548.4	-3.19	19.97 (MAD)	<0.001	(-7.23, 0.00)	23.08	0.00
	TSS	545.6	-2.25	7.43 (MAD)	0.011	(-6.27, 0.00)	24.37	-2.14
n=50	OLS	12.29	0.27	0.340	0.429	(-0.42, 0.96)	1.334	-2.77
	QR	10.00	0.00	0.231	1.000	(-0.18, 0.63)	1.000	0.00
	TS	11.00	0.00	1.876 (MAD)	<0.001	(-0.75, -0.01)	1.000	0.00
	TSS	10.00	0.00	0.000 (MAD)	0.207	(-0.80, 0.80)	1.000	1.00
n=100	OLS	514.9	2.46	1.47	0.097	(-0.45, 5.37)	28.79	0.85
	QR	512.5	2.93	2.46	0.237	(-1.24, 7.22)	28.84	0.00
	TS	516.2	2.41	21.45 (MAD)	0.687	(-0.46, 5.29)	28.72	0.00
	TSS	511.4	2.57	6.39 (MAD)	0.083	(-0.31, 5.45)	28.96	3.54

Table 61: Results of Relative Median Absolute Error of the four regression procedures with $n = 10, 30, 50, 100$, $N_{sim} = 1000$, $X \sim Unif(n, 1, 10)$, and $Y \sim Digit Preference (n)$:

	Relative Median Absolute Error	Value
n=10	OLS vs QR	0.130
	OLS vs TS	0.011
	OLS vs TSS	0.028
	QR vs TS	-0.136
	QR vs TSS	-0.116
	TS vs TSS	0.017
n=30	OLS vs QR	0.002
	OLS vs TS	-0.092
	OLS vs TSS	-0.092
	QR vs TS	-0.095
	QR vs TSS	-0.095
	TS vs TSS	0.000
n=50	OLS vs QR	0.001
	OLS vs TS	0.014
	OLS vs TSS	-0.041
	QR vs TS	-0.013
	QR vs TSS	-0.042
	TS vs TSS	-0.055
n=100	OLS vs QR	-0.002
	OLS vs TS	0.002
	OLS vs TSS	-0.006
	QR vs TS	0.004
	QR vs TSS	-0.004
	TS vs TSS	-0.009

Figure 20: Four regression lines are shown in each plot with $n = 10, 30, 50, 100$. $Nsim = 1000$, $X \sim Unif(n, 1, 10)$, $Y \sim Digit\ Preference(n)$:



CHAPTER 5 DISCUSSION

Overview:

Monte Carlo techniques were used to estimate the regression coefficients, standard errors, p-values, confidence intervals and median absolute deviation based on Ordinary Least Square Regression, the Quantile Regression, the Theil Sen Regression, and the Theil Sen Siegel Regression procedures. Visual as well as numerical comparisons were made using these four regression procedures. For visual comparison scatter plots with fitted regression lines using all four regression procedures were used. For numerical comparison, standard errors (in case of normal data with no outliers), median absolute deviation (in case of non-normal data or outliers), confidence intervals, mean bias (in case of normal data with no outliers), median bias (in case of non-normal data or outliers), root mean square error (RMSE), Relative Root Mean Square Error (in case of normal data with no outliers), median absolute error (MEDAE), and Relative Median Absolute Error (in case of non-normal data or outliers) were used.

The results from the simulation study, compiled above were indicated in the Tables 1 to 61 and based on $n = 10, 30, 50, 100$ and 1,000 simulations. Provided in these tables were descriptive statistics for both X and Y variables in different regression procedures and the estimates of regression coefficients i.e., β_0 and β_1 along with the standard error, p-value, 95% confidence interval, RMSE, MEDAE, and biasness of β_1 in Ordinary Least Square Regression, the Quantile Regression, the Theil Sen Regression, and the Theil Sen Siegel Regression procedures. Tables also provide the estimates of Relative Root Mean Square Error to measure the relative performance of OLS, QR, TS, and TSS. A negative value of Relative Root Mean Square Error (RRMSE) refer to the proportional increase in RMSE of first vs second regression procedure, on the other hand positive value of

Relative Root Mean Square Error (RRMSE) indicates a proportional decrease in RMSE of first vs second regression procedure.

The performance of Ordinary Least Square Regression, the Quantile Regression, the Theil Sen Regression, and the Theil Sen Siegel Regression lines are also presented for visual comparison with scatter plots and fitted regression lines.

Regression Model passing through origin under the Normality Assumption with no Outliers:

If the errors (e_i) were independent and normally distributed from a double exponential distribution then a random sample of size n for both predictor variable X , and outcome variable y from a bivariate normal distribution were generated with mean $(0, 0)$, variances equal to 1, and a correlation coefficient equal to 0.80. The results from simulation study are indicated from Table 1, to Table 3 with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 1 were the descriptive statistics for both X and Y variables and it can be seen as sample size increase all descriptive statistics become stable.

Provided in Table 2 were the estimates, of regression coefficients i.e., β_0 and β_1 along with the standard error, p-value, 95% confidence interval, RMSE, and biasness of β_1 with Ordinary Least Square Regression, the Quantile Regression, the Theil Sen Regression, and the Theil Sen Siegel Regression procedures. It can be seen; when sample size is small i.e. $n = 10$ the estimates of Y-intercept $\hat{\beta}_0$ are all negative and the estimates of slopes $\hat{\beta}_1$ are all positive with more or less same standard errors except the Regression procedure. The predictor X is significant only in OLS with p-value=0.043. The regression coefficient $\hat{\beta}_1$ is the only unbiased estimate of population regression coefficient β_1 with minimum RMSE in OLS as compared to other regression procedures. It can also be seen that QR with low RMSE and bias, perform better as compared to TS and TSS. When sample size $n = 30, 50, \text{and } 100$ all the estimates of Y-intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_1$ are all positive

with decreasing standard errors. The regression coefficient $\hat{\beta}_1$ is again the only unbiased estimate of population regression coefficient β_1 with minimum RMSE in OLS as compared to other regression procedures. It is to be noted that the biasness of QR, TS, and TSS are not quite stable with sample size, but still approaches to zero with an increase in n .

Provided in Table 3 were the estimates of Relative Root Mean Square Error to measure the relative performance of OLS, QR, TS, and TSS. It can be seen; when sample size is small i.e. $n = 10$ there is a proportional increase in RMSE of QR, TS, and TSS as compared to OLS. When sample size $n = 30$ and 50 OLS is still better followed by TS and TSS. When sample size $n = 100$ OLS is still better followed by QR and TS.

The four regression lines can also be seen at each plot in Figure 1. It can be seen, when sample size is small i.e. $n = 10$ more or less all four regression lines fit well. When sample size $n = 30, 50, 100$ all four regression line have same performance.

Regression Model with slope and intercept under the Normality Assumption with no Outliers:

If the errors (e_i) were independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a uniform distribution with $\min=0$ and $\max=1$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$. The results from simulation study were indicated in Table 4 to Table 6 with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 4 were the descriptive statistics for both X and Y variables and it can be seen as sample size increase all descriptive statistics become stable.

As indicated in the Table 5, with $n = 10, 30, 50, \text{ and } 100$ all the estimates of Y-intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_1$ were all positive with decreasing standard errors. The regression coefficient $\hat{\beta}_1$ is again the only unbiased estimate of population regression coefficient β_1 with minimum RMSE in OLS as compared to other regression procedures. When n is start getting large, the predictor X is

also start getting significant with $p < 0.001$. Again the biasness of QR, TS, and TSS were not quite stable with sample size, but still approaches to zero with an increase in n . In Table6, the relative performance of OLS is still better, followed by QR, TS, and TSS.

The four regression lines can also be seen at each plot in Figure 2. When the sample size is small i.e. $n = 10$ more or less all four regression lines fit well. When sample size $n = 30, 50, 100$, all four regression fit the same.

Regression Model with slop, intercept and dichotomous predictor variable with no Outliers:

If the errors (e_i) were independent and normally distributed with zero mean and 2 standard deviation then a random sample of size n for a predictor variable X was generated from a binomial distribution with a single trial and $p = 0.5$, and an outcome variable Y was defined as $Y = 2 + 3 * X + e$. The results from simulation study were indicated from Table 7 to Table 10 with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 7 and Table 8 were the descriptive statistics for both X and Y variables and it can be seen as sample size increase all descriptive statistics become stable.

As indicated in the Table 9, with $n = 10, 30, 50, \text{ and } 100$ all the estimates of Y-intercept $\hat{\beta}_0$ and slops $\hat{\beta}_1$ were all positive with decreasing standard errors. The regression coefficient $\hat{\beta}_1$ is again the only unbiased estimate of population regression coefficient β_1 with minimum RMSE in OLS as compared to other regression procedures. Again, when n is start getting large, the predictor X is also start getting significant with $p < 0.001$. Again, it is to be noted that the biasness of QR, TS, and TSS were not quite stable with sample size, but still approaches to zero with an increase in n . In Table6, the relative performance of OLS is still better, followed by TS, TSS and QR.

The four regression lines can also be seen at each plot in Figure 3. It can be seen, when sample size is small i.e. $n = 10$ more or less all four regression lines fit well. When sample size $n = 30, 50, 100$ all four regression fit the same.

Regression Model with Outliers in both X and Y direction:

If the errors (e_i) were independent and normally, then a random sample of size n was generated from a bivariate normal distribution with mean $(0, 0)$ and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 30% 50% and 100% of n were introduced in both X and Y variables from a bivariate normal distribution with means $(2, 6)$ and variances $0.1 \times$ variance of the above bivariate normal distribution, i.e. the variances $(0.1, 0.1)$. The results from simulation study were indicated from Table 11 to Table 22, with $n = 10, 30, 50, 100$ and 1000 simulations. Provide from Table 11, Table 14, and Table 17 were the descriptive statistics for both X and Y variables with 10%, 20%, 30% and 50% of n . In Table 12, when $n = 10$ with 10% and 20% outliers, all the estimates of Y-intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_1$ were all positive with increasing standard errors. The regression coefficient $\hat{\beta}_1$ is now the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The predictor X is statistically significant in all four procedures. The relative performance with median absolute deviation of TS is better, followed by QR and TSS. With 30% and 50% outliers, all the estimates of Y-intercept $\hat{\beta}_0$ were unstable with still positive slopes $\hat{\beta}_1$ with increasing standard errors. The regression coefficient $\hat{\beta}_1$ is not a stable median unbiased estimate of population regression coefficient β_1 as well as MEDAE. The predictor X is still statistically significant in all four procedures. The relative performance with median absolute deviation in Table 13 is not very consistent.

The four regression lines with 10%, 20%, 30% and 50% of $n = 10$ can also be seen at each plot in Figure 4. It can be seen, when sample size is small i.e. $n = 10$ with 10% outlier the QR,

TS, and TSS were more robust regression lines as compared to OLS. Moreover, QR and TSS were slightly more robust than TS. As the percentage of outliers increases TSS still robust regression line as compared to QR and TSS. It can be seen TSS is about to tolerate 50% of the outliers with $n = 10$.

In Table 14, Table 18, and, Table 21, when $n = 30, 50, \text{ and } 100$ with 10%, 20%, 30%, and 50% outliers in n , all the estimates of Y-intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_1$ were all mostly positive with increasing standard errors. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The predictor X is still statistically significant in all four procedures. The relative performance with median absolute deviation in Table 15, Table 19, and Table 22 shows TSS is more or less better, followed by TS and QR.

The four regression lines with 10%, 20%, 30% and 50% of $n = 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 4 to Figure 7. It can be seen, in all cases QR, TS, and TSS were still more robust regression lines as compared to OLS. Moreover, the TSS is clearly more robust regression procedure than TS followed by QR.

Regression Model with Outliers in Y direction only:

If the errors (e_i) were independent and normally, then a random sample of size n was generated from a bivariate normal distribution with mean $(0, 0)$ and variances equal to 1, and a correlation coefficient equal to 0.80. Outliers of 10%, 30% 50% and 100% of n were introduced in Y variable only from a bivariate normal distribution with means $(0, 6)$ and variances $0.1 \cdot \text{variance}$ of the above bivariate normal distribution, i.e. the variances $(0.1, 0.1)$. The results from simulation study were indicated from Table 23 to Table 34, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided from Table 23, Table 26, Table 29, and Table 32 were the descriptive statistics for both X and

Y variables with 10%, 20%, 30% and 50% of n in Y direction only. In Table 24, when $n = 10$ with 10%, 20%, 30%, and 50% outliers, all the estimates of Y-intercept $\hat{\beta}_0$ were positive and slopes $\hat{\beta}_1$ were all positive except the slope of OLS with 30% and 50% outliers, with increasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The predictor X is statistically significant at 10% outliers in all four procedures only. The relative performance with median absolute deviation of TS is better, followed by QR and TSS. With 30% and 50% outliers, all the estimates of slopes $\hat{\beta}_1$ were unstable with increasing standard errors in OLS and QR and median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is not a stable median unbiased estimate of population regression coefficient β_1 . The predictor X is not any more statistically significant in all four procedures. The relative performance with median absolute deviation in Table 25 is not very consistent.

The four regression lines with 10%, 20%, 30% and 50% of $n = 10$ can also be seen at each plot in Figure 5. It can be seen, when sample size is small i.e. $n = 10$ with 10% and 20 % outlier the QR, TS, and TSS were coincide and were more robust regression lines as compared to OLS. As the percentage of outliers increases QR still robust regression line as compared to TS and TSS.

In Table 27, Table 30, and, Table 33, when $n = 30, 50, \text{ and } 100$ with 10%, 20%, 30%, and 50% outliers in n , all the estimates of Y-intercept $\hat{\beta}_0$ and slopes $\hat{\beta}_1$ were all mostly positive with increasing standard errors in OLS and QR and an increasing median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The predictor X is still statistically significant in most of the all four procedures. The relative performance with

median absolute deviation were indicated in the Table 28, Table 31, and Table 34 shows that the QR is more or less better, followed by TS and TSS.

The four regression lines with 10%, 20%, 30% and 50% of $n = 30, 50, \text{and } 100$ can also be seen at each plot in Figure 6 to Figure 8. It can be seen, in all cases QR, TS, and TSS were still more robust regression lines as compared to OLS. Moreover, the QR is slightly more robust regression procedure than TS followed by TSS.

Regression Model under the Non-Normality Assumption:

An outcome variable Y was generated and uses the log link function so for a predictor variable X , was assumed $Y \sim \text{Poisson}(\lambda)$, and $\log(\lambda) = 1 + 0.2 * X$. It was assumed X is uniformly distributed with $\min=0$ and $\max=1$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. Again, the results from simulation study were indicated from Table 35 to Table 37, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 36 was the descriptive statistics for both X and Y variables. In Table 37, as the sample size increases the estimates of Y -intercept $\hat{\beta}_0$ approaches to 3.00 and slopes $\hat{\beta}_1$ approaches to zero, with decreasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The variable X is not a consistent statistically significant predictor. The relative performances with median absolute deviation of all four procedures were more or less same. The relative performance with median absolute deviation in Table38 looks consistent.

The four regression lines with $n = 10, 30, 50, \text{and } 100$ can also be seen at each plot in Figure 6. It can be seen that the QR, TS, and TSS were coincide and were more robust regression lines as compared to OLS.

Regression Model under the Micceri distributions:

Eight Micceri distributions were used to generate a random sample for an outcome variable Y , with a uniform distribution of predictor variable X with $\min=0$ and $\max=1$.

1. Smooth Symmetric distribution:

An outcome variable Y was generated from a Smooth Symmetric distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. Again, a Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. As usual the results from simulation study were indicated from Table 38 to Table 40, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 38 was the descriptive statistics for both X and Y variables. In Table 39, all the estimates of Y -intercept $\hat{\beta}_0$ were positive and slope $\hat{\beta}_1$ changes from negative to zero, with decreasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 with minimum MEDAE in QR, TS, and TSS as compared to OLS. The variable X is not a consistent statistically significant predictor. The relative performances with median absolute deviation of QR is better with $n=10$ and 30 as compared to others, similarly relative performances with median absolute deviation of TSS is better with $n=50$ and 100 as compared to other procedures. The relative performance with median absolute deviation in Table 40 looks consistent.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 7. It can be seen that the QR, TS, and TSS were almost coincide and were more robust regression lines as compared to OLS. The QR is slightly more robust line with small n and TSS is more robust regression lines with large n .

2. Extreme Asymmetric distribution:

An outcome variable Y was generated from an Extreme Asymmetric distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. Again, a Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. As usual the results from simulation study were indicated from Table 41 to Table 43, with $n = 10, 30, 50, 100$ and 1000 simulations. Again provided in Table 41 was the descriptive statistics for both X and Y variables. As indicated in the Table 42, all the estimates of Y -intercept $\hat{\beta}_0$ were positive and slope $\hat{\beta}_1$ changes from negative to zero as sample size become large, with decreasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 . It can be seen that the MEDAE is not very consistent measure in all four procedures. Again, the variable X is not a consistent statistically significant predictor. The relative performances with median absolute deviation of QR is better with $n=10$ as compared to others, similarly relative performances with median absolute deviation of TS is better with $n=30$ and 50 as compared to other procedures. The relative performance with median absolute deviation of QR is slightly better with $n=100$ as compared to others. The relative performance with median absolute deviation in Table 43 looks consistent.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 8. It can be seen that the QR, TS, and TSS were almost coincide and were more robust regression lines as compared to OLS with $n=10$ and $n=30$. The TSS is slightly more robust line as compared to others.

3. Extreme Bimodal distribution:

An outcome variable Y was generated from an Extreme Bimodal distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. The results from simulation study were indicated from Table 44 to Table 46, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in Table 44 was the descriptive statistics for both X and Y variables. As indicated in the Table 45, almost all the estimates of Y -intercept $\hat{\beta}_0$ were positive and slope $\hat{\beta}_1$ changes to zero as sample size become large, with decreasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 . It can be seen as the sample size increases MEDAE in all four procedures approaches to 1.00. Again, the variable X is not a consistent statistically significant predictor. The relative performances with median absolute deviation of TS is better with $n=10$ as compared to others, similarly relative performances with median absolute deviation of QR is better with $n=30$ as compared to other procedures. The relative performance with median absolute deviation of all procedures were same with $n=50$ and $n=100$. The relative performance with median absolute deviation in Table 46 looks consistent.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 9. It can be seen that the, TS, and TSS were almost coincide and were more robust regression lines as compared to QR and OLS.

4. Mass at Zero distribution:

An outcome variable Y was generated from Mass at Zero distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. The results from simulation study were indicated from

Table 47 to Table 49, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in the Table47 was the descriptive statistics for both X and Y variables. As indicated in the Table48, all the estimates of Y-intercept $\hat{\beta}_0$ were positive and slop $\hat{\beta}_1$ changes to zero as sample size become large, with decreasing standard errors in OLS and QR and Median absolute deviation in TS and TSS. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 in TS and TSS. The variable X is not a consistent statistically significant predictor with large n. The relative performances with median absolute deviation of QR is batter with $n=10$ and $n=30$ as compared to others, similarly relative performances with median absolute deviation of OLS is batter with $n= 50$ and $n=100$ as compared to other procedures. The relative performance with median absolute deviation in Table 49 looks consistent.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 10. It can be seen that the QR, TS, and TSS were almost coincide and were more robust regression lines as compared to OLS.

5. Mass at Zero with Gap distribution:

An outcome variable Y was generated from Mass at Zero with Gap distribution and a predictor variable X was generated from uniform distributed with $\text{min}=1$ and $\text{max}=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. The results from simulation study were indicated from Table 50 to Table 52, with $n = 10, 30, 50, 100$ and 1000 simulations. Again, provided in the Table50 was the descriptive statistics for both X and Y variables. As indicated in the Table51, all the estimates of Y-intercept $\hat{\beta}_0$ and slop $\hat{\beta}_1$ were zero except for OLS, with not available standard errors in QR, TS, and TSS. The Median absolute deviation in QR, TS and TSS were also zero. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 for QR, TS, and TSS. The variable X is not statistically significant predictor. The relative

performances with median absolute deviation of QR, TS and TSS were all zero. The relative performance with median absolute deviation in Table 52 is not available while comparing QR, TS, and TSS.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 11. It can be seen that the QR, TS, and TSS were all coincide and were more robust regression lines as compared to OLS.

6. Multimodal Lumpy distribution:

An outcome variable Y was generated from Multimodal Lumpy distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{\text{sim}}=1000$ repetitions. The results from simulation study were indicated in Table 53 to Table 55, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in the Table 53 was the descriptive statistics for both X and Y variables. As indicated in the Table 54, all the estimates of Y -intercept $\hat{\beta}_0$ were positive and slope $\hat{\beta}_1$ changes from positive to negative for $n=30, 50, \text{ and } 100$ with an inconsistent performance of standard errors. The Median absolute deviation error is not consistent also. The regression coefficient $\hat{\beta}_1$ is still the median unbiased estimate of population regression coefficient β_1 for QR and TSS. The variable X is not statistically significant predictor. The relative performance with median absolute deviation in Table 55 is not consistent again comparing QR, TS, and TSS.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 12. It is difficult to decide which one is more robust.

7. Extreme Asymmetry – Decay distribution:

An outcome variable Y was generated from Extreme Asymmetry – Decay distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte

Carlo Simulation was conducted with $N_{sim}=1000$ repetitions. The results from simulation study were indicated in Table 56 to Table 58, with $n = 10, 30, 50, 100$ and 1000 simulations. Table 56 again provided the descriptive statistics for both X and Y variables. As indicated in the Table 57, all the estimates of Y-intercept $\hat{\beta}_0$ were positive and slop $\hat{\beta}_1$ were all zero except for OLS with a decreasing value of standard errors and Median absolute deviation. The regression coefficient $\hat{\beta}_1$ is a median unbiased estimate of population regression coefficient β_1 for QR and TS. The variable X is not consistent statistically significant predictor in most of the cases. The relative performance with median absolute deviation in Table 58 is not consistent.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 13. It can be seen QR, TS, and TSS coincides with $n=10$ and 30. Also QR and TSS were still robust followed by TS and OLS.

8. Digit Preference distribution:

An outcome variable Y was generated from Digit Preference distribution and a predictor variable X was generated from uniform distributed with $\min=1$ and $\max=10$. A Monte Carlo Simulation was conducted with $N_{sim}=1000$ repetitions. The results from simulation study were indicated in Table 59 to Table 61, with $n = 10, 30, 50, 100$ and 1000 simulations. Provided in the Table 59 was the descriptive statistics for both X and Y variables. As indicated in the Table 57, all the estimates of Y-intercept $\hat{\beta}_0$ were positive and slop $\hat{\beta}_1$ changes from negative to positive for $n=50$ and 100 with an inconsistent performance of standard errors. The Median absolute deviation error is not very consistent also. The regression coefficient $\hat{\beta}_1$ is a median unbiased estimate of population regression coefficient β_1 for QR and TS. The variable X is not consistent statistically significant predictor. The relative performance with median absolute deviation in Table 61 looks consistent again comparing QR, TS, and TSS.

The four regression lines with $n = 10, 30, 50, \text{ and } 100$ can also be seen at each plot in Figure 14. It is difficult to decide which one is more robust.

Conclusion

When the regression model passing through origin under the normality assumption with no outliers, it can be seen OLS was more robust regression models with small standard errors and small RMSE as compared to QR, TS, and TSS. Among the QR, TS, and TSS models, when $n=10$ QR, was perform batter followed by TS and TSS. When $n=30$ and 50 , TS model perform well as compared to QR and TSS. Similarly, when $n=100$ TSS model looks slightly more robust with little small RMSE.

When the regression model is being used with slop and intercept under the normality assumption with no outliers, it can be seen again OLS was more robust regression models with small standard errors and small RMSE as compared to QR, TS, and TSS. Among the QR, TS, and TSS models, when $n=10$ QR, was perform batter followed by TS and TSS. When $n=30$ and 50 , TS model perform well as compared to QR and TSS. Similarly, when $n=100$ QR model looks slightly more robust with small RMSE.

When the regression model is being used with slop, intercept and a dichotomous predictor variable with no outliers, it can be seen again OLS was still more robust regression models with small standard errors and small RMSE as compared to QR, TS, and TSS.

When the regression model is being used with outliers in both X and Y directions, it can be seen QR has small standard as compared to OLS and TSS has small median absolute deviation than TS. All figures illustrated in regression model with outliers in both X and Y directions that TSS was more robust regression model followed by TS, QR, and OLS.

When the regression model is being used with outliers in Y directions only, it can be seen QR again has small standard errors as compared to OLS and TSS has small median absolute deviation than TS in some cases. All figures illustrated in regression model with outliers in Y direction only, QR was more robust regression model followed by TS, TSS, and OLS.

When the regression model is being used under the non-normality assumption, it can be seen OLS has small standard errors as compared to QR and TS has more or less same median absolute deviation than TSS. All figures illustrated in regression model under the non-normality assumption all three regression procedures QR, TS, and TSS were more robust compared to OLS. Moreover there was no much difference among the three procedures.

When the regression model is being used with Smooth Symmetric distribution under the Micceri family, it can be seen OLS has small standard errors as compared to QR and TSS has less median absolute deviation than TS. All figures illustrated in regression model with smooth symmetric distribution under the Micceri family distribution, TSS was slightly more robust.

When the regression model is being used with Extreme Asymmetric distribution under the Micceri family, it can be seen OLS again has small standard errors as compared to QR and TSS has less median absolute deviation than TS with large n. All figures again illustrated in regression model with Extreme Asymmetric distribution under the Micceri family distribution, TSS was slightly more robust.

When the regression model is being used with Extreme Bimodal distribution under the Micceri family, it can be seen OLS again has small standard errors as compared to QR and TSS has less median absolute deviation than TS with large n. All figures again illustrated in regression model with Extreme Bimodal distribution under the Micceri family distribution, TSS was slightly more robust.

When the regression model is being used with Mass at Zero distribution under the Micceri family, it can be seen OLS has small standard errors as compared to QR when $n=30, 50,$ and 100 and TSS has less median absolute deviation than TS with similar $n=30, 50,$ and 100 . All figures illustrated in regression model with Mass at Zero distribution under the Micceri family distribution, QR, TS, and TSS was slightly more robust than OLS, but they have similar performance among each other.

When the regression model is being used with Mass at Zero with Gap distribution under the Micceri family, it can be seen OLS again has small standard errors as compared to QR and TSS has less median absolute deviation than TS. All figures illustrated in regression model with Mass at Zero with Gap distribution under the Micceri family distribution, QR, TS, and TSS was slightly robust than OLS, but they have similar performance among each other.

When the regression model is being used with Multimodal Lumpy distribution under the Micceri family, it can be seen OLS has small standard errors as compared to QR and TSS has less median absolute deviation than TS. All figures illustrated in regression model with Multimodal Lumpy distribution under the Micceri family distribution, QR, TS, TSS, and OLS have similar performance.

When the regression model is being used with Extreme Asymmetry distribution under the Micceri family, it can be seen OLS has small standard errors as compared to QR and TSS has less median absolute deviation than TS. All figures illustrated in regression model with Extreme Asymmetry distribution under the Micceri family distribution, QR, TSS, were more robust as compared to TS and OLS.

When the regression model is being used with Digital Preference distribution under the Micceri family, it can be seen OLS again has small standard errors as compared to QR and TSS

has less median absolute deviation than TS. All figures illustrated in regression model with Digital Preference distribution under the Micceri family distribution QR, TS, TSS, and OLS have similar performance.

Therefore, it is recommended that, under the normality assumption with no outliers OLS should be the most suitable regression procedure followed by QR, TS and TSS. When there are outliers in both X and Y direction TSS should be the most suitable followed by QR and TS. Under the non-normality assumption QR, TS and TSS have more or less same performance. For, Micceri family distribution overall TSS might be a suitable regression procedure.

Since, the eight empirical distributions identified by Micceri (1989), were categorized into general achievement/ability tests, criterion/mastery tests, psychometric measures, pre-test measures, and post-test measures, therefore it is recommended to use TTS procedure while using these distribution in a large data set, with an exception is the smooth symmetric data set, which appeared in less than 3% of Micceri's data sets.

It is to be noted that the results presented in this dissertation are the solutions to outliers when doing regression analysis. The results equally apply to ANOVA models as well.

Although, multiple linear regression was discussed in literature review, but our main focus was to study simple linear regression. For future study, all four regression techniques can be compared with the presence of outliers and non-normality of population distribution assumption using multiple regression technique.

REFERENCES

- Brauner, J.S. (1997). Nonparametric estimation of slope Sen's method, in Gallagher, Daniel (ed.), Environmental sampling and monitoring primer, accessed on June 30, 2004, at URL <http://www.cce.vt.edu/>.
- Chaudhary, S.M. & Kamal, S. (2000). *Introduction to Statistical Theory* (Part II.): Lahore, Punjab: Ilmi katab khana.
- Dietz, E.J. (1987). A comparison of robust estimators in simple linear regression: *Communication in Statistics Simulation*, v. 16, p. 1209–1227.
- Efron, B. (1987). Better bootstrap confidence intervals: *Journal of the American Statistical Association*, 82, 171-185.
- Efron, B. and Tibshirani R.J. (1993). *An Introduction to the Bootstrap*: London: Chapman & Hall.
- Edgeworth, F.Y. (1887). On Observations Relating to Several Quantities: *Hermathena*, 6, 279-285.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.
- Helsel D.R., Hirsch, R.M. (2002). *Statistical methods in water resources-Hydrologic analysis and interpretation*: Techniques of Water-Resources Investigations of the U.S. Geological Survey, chap. A3, book 4, 510 p.
- Hao, L., and Naiman, D. Q. (2007). *Regression procedure*. Thousand Oaks: Sage Publications Inc.
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* 19, 293-325.
- Hollander, M. & Wolfe, D. (1999). *Nonparametric statistical methods*: New York: Wiley.

- Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics*, 2nd Ed. New York: Wiley. Hussain, S.S., and Sprent, P., 1983, nonparametric regression: *Journal of the Royal Statistical Society, Series A*, v. 146, p. 182–191.
- Kendall, M. G. (1938). A new measure of rank correlation: *Biometrika* 30, 81-93.
- Koenker, R., Bassett, G. W. (1978). Regression Quantiles: *Econometrica* 46:33–50.
- Koenker R., Bassett GJ. (1982). Robust tests for heteroscedasticity based on regression quantiles: *Econometrica*. 50:43–61.
- Koenker, R., Bassett, G. W. (1982b). Test for linear hypothesis and L1 estimation: *Econometrica* 50:1577–1583.
- Koenker R, Machado JAF. (1999). A goodness of fit and related inference processes for regression procedure: *Journal of the American Statistical Association*. 94:1296–1310.
- Lawson, Kevin Duwan. "Statistical Inference for a Linear Function of Medians: A Comparison of the Maritz -Jarrett and Price -Bonett Estimators." Order No. 3243056 Wayne State University, 2006. Ann Arbor: ProQuest, web. 10 Sep. 2018.
- Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures: *Psychological Bulletin*, 105, 156-166.
- Michael Greenacre & H. Ayhan, (2015). "[Identifying Inliers](#)," [Working Papers](#) 763, Barcelona Graduate School of Economics.
- Mosteller F, Tukey JW. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley Publishing Company.
- Nelson, D. (1998). *Dictionary of mathematics*: London, England: Penguin.

- Nevitt, Jonathan, and Tam, H.P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model: *Multiple linear regression viewpoints*, v. 25, p. 54–69.
- Petscher Y, Logan JAR. (2014). Regression procedure in the study of developmental sciences: *Child Development*. 85(3):861–881.
- Petscher, Y.; Logan, JAR.; & Zhou, C.(2013). Extending conditional means modeling: An introduction to regression procedure. *An applied quantitative analysis in the social sciences*. (p. 3-33). Routledge; New York.
- Rogers WH. (1992). Regression procedure standard errors: *Stata Techn Bull Reprints*. 2:133–7.
- Sawilowsky, S.S and Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality: *Psychological bulletin*. 111 (2), 352.
- Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.*, 63, 1379-1389.
- Siegel, A.F. (1982). Robust Regression Using Repeated Medians, *Biometrika*, 69, 242-244.
- Syed, H. S., Rehman, A., Rashid, T., Karim, J., Syed, & S.M., (2016): A Comparative Study of Ordinary Least Squares Regression and Theil-Sen Regression through Simulation in the Presence of Outliers: *Lasbela, U. J.Sci.Techl.*, vol.V, pp.137-142.
- Theil, H. (1950). A rank invariant method for linear and polynomial regression analysis: *Nederl. Akad. Wetensch. Proc. Ser. A* 53, 386-392 (Part I), 521-525 (Part II), 1397-1412 (Part III).
- Vogt, W. (1993). *Dictionary of statistics and methodology: a non-technical guide for the social sciences*: Newbury Park: Sage Publications.

Weiss A. (1990). Least absolute error estimation in the presence of serial correlation: *Journal of Econometric*. 44, 127-159.

Wilcox, R. R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic, *Biometrical Journal*, 3, 261-268.

Wilcox, R. R. (2012b). Introduction to Robust Estimation and Hypothesis Testing, 3rd Edition. San Diego, CA: Academic Press.

Xin D., et al. (2009). Theil-Sen Estimators in a Multiple Linear Regression Model:
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.508.3461>

ABSTRACT**A COMPARATIVE STUDY OF KENDALL-THEIL SEN, SIEGEL VS QUANTILE REGRESSION WITH OUTLIERS**

by

AHMAD FAROOQI**December 2019****Advisor:** Dr. Shlomo S. Sawilowsky**Major:** Education Evaluation & Research**Degree:** Doctor of Philosophy

Researchers in social and behavioral sciences usually interested in study the relationship between a response variable Y_i and one or more independent predictors X_i either for the purpose of explanation or prediction. Ordinary Least Square Regression is a parametric approach used to study this kind of relationship. One of the disadvantages of Ordinary Least Square is it does not fit well in the presence of outliers in the response variable Y_i or both in the response variable Y_i and the predictor variable X_i , also if the data were sampled from a non-normal distribution. Quantile Regression, Theil-Sen regression, and the modified Theil-Sen Siegel regression are non-parametric approaches that can also be used to study the relationship and are more robust methods to outliers and non-normality of the distribution.

Several comparisons are made between Ordinary Least Square Regression, Quantile Regression, Theil Sen Regression, and Theil Sen Siegel Regression, but no direct comparison is yet made between Quantile Regression, Theil Sen Regression and Theil Sen Siegel Regression in the presence of outliers. In order to investigate this claim, Monte Carlo simulation study were employed and observations were generated from three theoretical and eight Micceri family distributions. Similarly, observations for the Monte Carlo simulations will be randomly generated with dif-

ferent sample sizes in the presence of 10% and 20%, 30% and 50% outliers. A comparison based on Mean Bias, Median Bias, Standard Deviation (S.D), Standard Errors (S.E), Root Mean Square Error (RMSE), Relative Root Mean Square Error (RRMSE), Median Absolute Error (MEDAE), and Relative Median Absolute Error (RMEDAE) of the four regression procedures are used to evaluate the model fitting.

The results of the study showed, under the normality assumption with no outliers Ordinary Least Square Regression should be the most suitable regression procedure followed by Quantile Regression, Theil Sen Regression, and Theil Sen Siegel Regression. When there are outliers in both X and Y direction Theil Sen Siegel Regression should be the most suitable followed by Quantile Regression and Theil Sen Regression. Under the non-normality assumption Quantile Regression, Theil Sen Regression and Theil Sen Siegel Regression have more or less same performance. For, Micceri family distribution overall Theil Sen Siegel Regression might be a suitable regression procedure.

AUTOBIOGRAPHICAL STATEMENT

Ahmad Farooqi was born in Gujranwala, and grew up in Lahore Pakistan, where in 1989; he graduated in double Mathematics and Statistics from the Dyal Singh College, Lahore, Pakistan. Following his graduation, Farooqi spent two years completing his M.Sc. in Statistics from University of Agriculture, Faisalabad, Pakistan in 1993. Farooqi has taught Statistics to undergraduate and graduate students in Siqarah Degree College, Lahore, Pakistan in 1996 and served as a Lecturer in Statistics at Garrison Post Graduate Degree College (presently Lahore Garrison University), Lahore, Pakistan from 1997 to 2004. In 1998, Farooqi earned M.Ed. from University of Punjab, Lahore, Pakistan followed by M.Phil. Statistics from Government College University, Lahore, in 2002.

In 2004, Farooqi leave for North America as a landed immigrant where in 2007; he completed an MS in Statistics from University of Windsor, Ontario, Canada with an experience of Graduate/Research Teaching Assistant for one and a half year. Farooqi served as a lecturer in Mathematics at Hafr Al-Batin Community College affiliated with King Fahd University of Petroleum and Minerals, Dammam, Saudi Arabia. In 2012, Farooqi has also successfully completed an M.A. in Social Data Analysis from University of Windsor, Ontario, Canada with an eight-month experience of Teaching Assistantship. Farooqi has co-authored sixteen peer-reviewed articles as a statistical analyst till 2018.

Farooqi love Statistics and Mathematics, he is a SAS certificated and proficient in IBMSPSS, STATA, R, LATEX, AMOS, SMART BOARD, MINITAB, MPLUS, PASS, NCSS, and FORTRAN. Since 2015, he is working as a Biostatistician and Pediatric Education Faculty member at CRCM/Department of Pediatric, Wayne State University School of Medicine Detroit, MI, United States of America.

Farooqi began work in September 2016 on his PhD in Educational Evaluation & Research with specialization in Robust Statistics, awarded by the College of Education, Wayne State University with an expected graduation in 2019.

He has been married to Tahir Qureshi for more than 19 years. His children, Usman, Maria, Hadia, Hassan, and Afia live in Windsor, Ontario, Canada. He love travelling, playing, and watching cricket.