AN EVALUATION OF SOME ROBUST ESTIMATORS OF REGRESSION COEFFICIENTS

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science

In

Statistics

University of Regina

By

Jie Ding

Regina, Saskatchewan

December 2015

UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Jie Ding, candidate for the degree of Master of Science in Statistics**,** has presented a thesis titled, ***An Evaluation of Some Robust Estimators of Regression Coefficients,*** in an oral examination held on November 30, 2015. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner:	Dr. Yiyu Yao, Department of Computer Science
Co-Supervisor:	Dr. Andrei Volodin, Department of Mathematics & Statistics
Co-Supervisor:	*Dr. Christine Chan, Software Systems Engineering
Committee Member:	Dr. Dianliang Deng, Department of Mathematics & Statistics
Chair of Defense:	Dr. Paul Laforge, Faculty of Engineering & Applied Science

*Not present at defense

Abstract

In the theory of regression analysis, the method of least squares is most commonly used because of its mathematical beauty and computational simplicity. However, this method is now criticized more and more because it often has very poor performance when there are outliers in the data. In this connection a variety of robust statistics are developed for that they are not unduly affected by outliers. In this thesis comparison studies have been made for several robust statistics to see which performs better than the others. Monte Carlo simulation has been used to carry out the comparison of these statistics, including the least absolute deviations estimator and the least median of squares estimator, least trimmed squares estimator and some M-estimators.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Andrei Volodin for his continuous help to my Master study and related research, and for his patience and motivation. Also I am grateful to my co-supervisor, Dr. Christine Chan, for her enlightening advice. Besides, I want to thank the Faculty of Graduate Studies and Research for offering me Teaching Assistant jobs and allowing me to work out of campus as a co-op student so I can get the financial support and gain precious working experience.

Last but not the least, I would like to thank my family for all the support they give me.

Jie Ding

Regina, Canada, December 2015

Contents

List of Tables

List of Figures

Chapter 1

Intruduction

1.1 Background

In classical statistical theory, a lot of statistical estimation methods are based on the model with certain assumptions, like the variables are following normal distribution and mutually independent. However, this is not often met in practice. It is hard to describe most of random phenomena with simple statistical models, and furthermore, the optimal estimators could be gotten via these methods only if all the assumptions are fulfilled. In a word, we would find that these statistical models are just approximations to the actual events at a certain degree and when the actual data departure from the model assumptions, the estimators would no longer be the best. A contemporary and important statistical topic name Robust statistics was developed under this background.

The statistical use of the word "Robust" was first proposed by George Box in 1953. But robustness was accepted by most of people as a subdiscipline in statistics was primarily the creation of John W. Turkey and Peter J. Huber in 1960s. In 1964, Peter Huber published his article,"Robust Estimation of a Location Parameter",

introducing the M-estimator, which was recognised as an astonishing accomplishment in the literature on robustness. Over the next few decades, Hampel proposed two important statistical concepts, the breakdown point and influence curve, and with R-estimate and L-estimates joining Huber's M-estimates, the three main classes of estimates have been formed. Stigler (2010) wrote an article called "The Changing History of Robustness" well summarizing the development of the robust statistics. Note that there other non-parametric methods of estimation, such as jackknife and bootstrap, that are out of the scope of thesis.

These robust methods seek to provide ways with good performance when the fundamental assumptions for data sets are not fully fulfilled. According to previous studies, we could conclude that a good robust statistical method should satisfy the following theoretical requirements:

1. When the actual observations are consistent with the model assumptions, the method is near-optimal;

2. When the actual observations have a small departure from the model assumptions, the robust method still has a good performance;

3. When the actual observations have a serious departure from the model assumptions, the corresponding method is still working;

So my study with respect to the comparison of the various robust statistics will take these properties into account.

1.2 Object of Study

One of the main motivations to develop statistical methods is the existence of influential data points, also well-known as outliers. They have negative affect on the

performance of the regression methods but they could not be avoid. In this thesis, I am going to compare some frequently used robust methods which are used to calculate the regression coefficients in a linear regression model to see which of them are more outlier-resistant than the others. Below robust methods will be adopted

1. Detecting the outlying data points using Studentized residuals
2. Detecting the outlying data points using Cook's distance
3. M-Estimator: Huber M-estimator and MM-estimator
4. Least Absolute Deviation Estimator
5. Least Median of Squares Estimator
6. Least Trimmed Squares Estimator

1.3 Outline

The thesis contains introduction, three chapters about the comparison, conclusion, references and one appendix. After introduction and literature review in the first two chapters, in chapter three, the construction of the comparison is illustrated. Chapter four is devoted to the procedures and the simulation studies. Chapter 5 is the conclusion. In this last chapter, the summarize and review the main results of the work are presented. In Appendix, computer simulation codes for R is presented.

Chapter 2

Literature Review

2.1 Outlier, Leverage and High Influential Data

At the beginning, it is important to distinguish three important kinds data in robustness - outlier, leverage and high influential data and explain their difference and connections.

2.1.1 Outliers

Outliers in statistics don't have a rigorous definition; determining whether or not a data point is an outlier is ultimately a subjective exercise. Virtually, how people make the choice of the data model will decide the assumption about the "normal" behaviour of the data. Given this, I reviewed how the "outliers" are characterized by statisticians from some different perspectives.

Hawkins (1980) defined "outlier" as "the outlying observation appears to deviate markedly from other members of the sample in which it occurs" but this is deemed not enough because it is susceptible by people's anticipation. Some data may act as outliers under this distribution but not outliers any more under another distribution.

Later, Barnett and Lewis (1984) gave outliers an more broad but detail definition. They distinguished the extreme observations, outliers and contaminants and revised the definition of outliers to be the observations that appear inconsistent with the rest of the data. Extreme observations could be the smallest and largest value of a data set. Contaminants are data points from another distribution "slipped" upward relative to the original distribution. Both extreme observations and contaminants may or may not be outliers and vice versa. Long after that, Aggarwal (2013) divided outliers, from another point of view, into the noise and anomalies based on analyst interest.

There are three main purposes regarding outliers are clearly distinguished by Iglewicz and Hoaglin (1993) :

1. Outlier labelling - Flag potential outliers for further investigation
2. Outlier accommodation - Use robust statistical techniques to accommodate the outlying observations
3. Outlier identification - Use formal test to identify whether the observations are outliers

Note sometimes, we cannot remove outliers from the dataset, we should pay attention to them.

2.1.2 Hat Matrix and Leverage

When talking about outliers and robust statistics, an understanding of the behaviour of the hat matrix $\boldsymbol{H}$ is important. Many scholars have pointed out that the $\boldsymbol{H}$ and the leverage h_{ii} are of great importance when performing an analysis on the classical linear regression model.

Hat matrix is also called influence matrix or projection matrix. In the classical linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1.1}$$

where $\boldsymbol{y}$ is $n \times 1$ vector of the response variable, $\boldsymbol{X}$ is a $n \times p$ full rank matrix of the explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector (where p is the number of number of parameter of the model, for example, for simple linear regression, p=2 because there are two parameters α, β) the unknown regression coefficients, and $\boldsymbol{\epsilon}$ is the error vector. When the weights for each observation are identical and the errors are uncorrelated, the estimated parameters are $\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$, so the fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$. Therefore the hat matrix is given by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}. \tag{2.1.2}$$

It is also the reason the matrix has been dubbed the 'hat' matrix: it mapped $\boldsymbol{y}$ to $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.

The ij element of the hat matrix is

$$h_{ij} = \boldsymbol{x_i^T}(\boldsymbol{X^T X})^{\boldsymbol{T}}\boldsymbol{x_j}, \quad i, j = 1, 2, \ldots, n$$

where x_i^T is the ith observation and $x_i^T = (1, x_1, \ldots, x_n)$. Specially, the diagonal elements, which are so called leverage h_{ii}, is

$$h_{ii} = \boldsymbol{x_i^T}(\boldsymbol{X^T X})^{\boldsymbol{T}}\boldsymbol{x_i}, \quad i = 1, 2, \ldots, n.$$

Some basic facts of hat matrix and leverage corresponding to a linear regression model are summarized as follows:

1. $\boldsymbol{H}$ is symmetric and idempotent, that is, $\boldsymbol{H^T = H}$ and $\boldsymbol{H^2 = H}$ separately

2. trace($\boldsymbol{H}$)=rank($\boldsymbol{H}$)=p, where p is the number of the independent parameters

3. $\forall\ i, j = 1, 2, \ldots, n$, $0 \leq h_{ii} \leq 1$, $-0.5 \leq h_{ij} \leq 0.5$ when $i \neq j$, $h_{ii} \geq n^{(-1)}$ when there is a constant column in $\boldsymbol{X}$.

Hat matrix and the leverage also play important roles in the regression theory. Since $Var(\hat{y}_i) = \sigma^2 h_{ii}$ and $Var(e_i) = \sigma^2(1 - h_{ii})$, where σ^2 is variance, the constant factor, and e_i is the residual, $e_i = y_i - \hat{y}_i$ (see section (2.2.1) for further discussion):

1. The relationship between the residual terms e_i and e_j could be determined by the elements in $\boldsymbol{H}$.

   The covariance of e_i and e_j:

$$\begin{aligned} cov(e_i, e_j) &= cov[\boldsymbol{u_i(I - H)y, u_j(I - H)y}] \\ &= \boldsymbol{u_i} cov[\boldsymbol{(I - H)Y, (I - H)Y}]\boldsymbol{u_j^T} \\ &= \boldsymbol{u_i(I - H)}\sigma^2\boldsymbol{u_j^T} \\ &= -\sigma^2 h_{ij} \end{aligned}$$

   where $\boldsymbol{u_i}$ is $1 \times n$ vector such that $\boldsymbol{u_i} = (0, \ldots, 0, 1, 0, \ldots, 0), \quad i = 1, \ldots n.$

   The correlation coefficient of e_i and e_j:

$$\begin{aligned} r_{e_i, e_j} &= \frac{cov(e_i, e_j)}{\sqrt{D(e_i)D(e_j)}} \\ &= \frac{-h_{ij}}{\sqrt{1 - h_{ii}}\sqrt{1 - h_{jj}}} \end{aligned}$$

2. The relationship between the residual terms e_i and the error term ϵ_i could be determined by the leverage, which would be shown in the later section.

3. The leverage score is also known as the observation self-sensitivity or self-influence (Park and Xu, 2013). It describes the influence each response value has on the fitted value for the same observation and it is as shown below:
   from 2.1.2,we can deduce

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_j + \sum_{i \neq j} p_{ij} y_j, \quad , i = 1, 2, \ldots, n$$

   then

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii} \quad i = 1, 2, \ldots, n$$

It clearly demonstrates how y_i affect $\hat{y}_i$. Same, h_{ij} describes the influence the how y_j affect $\hat{y}_i$.

Besides the properties mentioned above, the leverage, in particular, can be shown that as a measurement of the identifying outlying x observations (Kutner et al., 2004). In the univariate linear regression model, the leverage could be expanded like

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \tag{2.1.3}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{2} x_i$. It is evident that the value of h_i depends on the distance between the ith observation x_i and the center $\bar{x}$. A large h_i could badly distort the regression coefficient, so we also call the point with $h_i \approx 1$ high leverage point and this kind of points could be regard as outlier in x-space.

2.1.3 High Influential Data and Summary

The term "high influential data" refers to those data points if removing them, the estimate of the regression coefficients will be substantially changed. They can be considered to be the product of leverage and outliers.

Basically, the outlying data could be classified into three categories:

1. Outlying in y-space. These data points are close to the center of sample, but their y-values are larger than others. They could be outliers and high influential data, but they are not leverage.

2. Outlying in x-space. These data are leverage with normal y-value.

3. Outlying in both y-space and x-space.

All the three kind of outlying data are illustrated in the Figure (2.1). The main cluster shows the normal data. Point A is an outlier on $y-axis$ but inlier on $x-axis$, it is influential as it might change the slope or intercept of the line, but it is close to the center so it's not a leverage. Point B is an outlier on $x-axis$ but an inlier on $y-axis$, it is leverage and it's high influential as well. Point C is outlying on both $x-axis$ and $y-axis$, it is far from the center but it doesn't affect both the parameters, so although it's a leverage, it's not a high influential data.

Figure 2.1: Outliers, High Influential points and Leverages

2.2 Diagnostic Measurements of Outliers

The detection of influential data points is conducted using either informal graphics (like Figure 1.1) to show the general characters of the data or formal tests to provide a concise summary of the data points. The most commonly used diagnostic statistics for detecting outliers are Studentized residual and Cook's distance. Hossain and Naik (1991) did a comparison on different methods regarding how they performed on the detection of the influential observations in linear regression models and the results shows that the Studentized residual is appropriate for detection of outliers with abnormal high value for the variance $\boldsymbol{\epsilon}$ and the Cook's distance are appropriate for the detection of influential observations.

2.2.1 Studentized Residual

In the regression analysis, the difference between the fitted value and the true value, the residuals, carries important informations concerning the appropriateness of

assumptions. Considering the linear regression model (2.1.1), for linear least squares, the vector of ordinary residuals $\boldsymbol{e}$ is given by

$$\boldsymbol{e} = (e_i) = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$$

where H is the hat matrix and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}$ and then $var(e_i) = (1 - h_{ii})\sigma^2$. We can see that the variance of the ordinary residuals does not equal to σ^2 but affected by the value of the leverage h_{ii}. From previous section about the h_{ii} expression (2.1.3), we know the closer x_i is to $\bar{x}$, the larger is the variance of the residuals, the futher x_i is from $\bar{x}$, the smaller is the variance of the residuals. So different location of the observation results in different variance of the ordinary residuals. It is better to use a set of residuals which has the consistent variance and the Studentization of residuals are shown to be independent of scale parameters (Cook and Weisberg, 1982).

The Studentized residual are defined by

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}$$

where m is the number of the parameters, $\hat{\sigma}^2 = \frac{1}{n-m}\sum_{j=1}^{n} e_j^2$ is the unbiased estimate of the variance of the error term then furthermore, $\hat{\sigma}^2(1 - h_{ii})$ is the unbiased estimate of the variane of e_i.This transform of the residuals is useful for judging how far away the data point is from the other observations in $y - direction$.

2.2.2 Cook's Distance

Another effective way to detect outliers is the famous Cooks Distance. Its used commonly ever since it was introduced by American statistician Cook (1977). Data points with large Cooks Distance are those who need to be investigated further because they could either be data containing important information or be contaminations which should be gotten rid of.

This distance is defined to be

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' X' X((\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2}; \quad i = 1, 2, \ldots, n$$

where $\hat{\beta}$ denotes the least squares estimate of β, $\hat{\beta}_{(-i)}$ is the LS estimate of β with the ith point deleted, p is the number of the parameters and s^2 is the sample variance.

We can derive the below expression of the Cook's distance in a few steps (see Cook, 1977)

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] = \frac{r_i^2}{p}\left[\frac{h_{ii}}{(1-h_{ii})}\right]$$

where obviously r_i is the Studentized residual. It combines the information from both leverage and residual: the larger either r_i or $h_i i$ is, the lager D_i is.

A conservative approach for deleting highly influential observations relies on the fact that Cook's distance is a F-like statistic with p and $n-p$ as the degree of freedom, so we can use $F(p, n-p, 1-\alpha)$ as a critical yardstick. Note that the usual confidence level F-distribution, say α=95%, is no more appropriate here. The value of α, say α=50%, could be interpreted as deletion of the ith point moves the LS estimate from the center to the boundary of the 50% confidence interval. The following graph for Cook's Distance is quite straightforward:

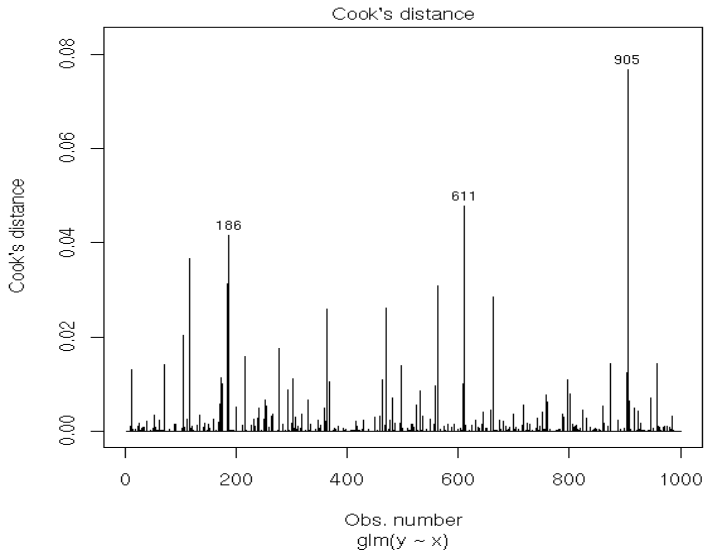


Figure 2.2: Cook's Distance

2.3 Robust Statistical Methods

After a brief review of influential data, in this section, I introduce some more commonly used robust statistical methods. Ordinary Least Squares Method would first be illustrated in section 2.3.1. Although it is not robust, it is quite flexible and could be used with other methods. M-estimator method is illustrated in the section 2.3.2. In sections 2.3.3 and 2.3.4 are about least absolute deviation method and least median squares method. The last section, 2.3.5, is about least trimmed squares estimator.

2.3.1 Ordinary Least Squared Estimator(OLS)

Ordinary least squares method is also called linear least squares method; it is least squared method used in a linear regression model for estimating the unknown parameters. Although in the present days, it is blamed more and more because even a small

contamination could have a large negative influence on the estimation result, least squares method still remain King for its magic - the perfection of the normal distribution theory and the available ANOVA analysis method provides a unified approach for testing the linear hypothesis (Stigler, 2010). As OLS does work overwhelms its very real shortcomings, I would like to employed this method in my research and try to combined it with Student Residuals and Cook's Distance respectively.

Recall the classical linear regression,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

the $p \times 1$ vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is unknown, β_0 is the constant and β_1,. . . ,β_{p-1} are the the regression coefficients which we are looking for. Regarding the error term $\boldsymbol{\epsilon}$, each error e_i follows $e_i \sim N(0, \sigma^2)$, which means

$$(e_1, \ldots, e_n) \sim \frac{1}{(\sqrt{2\pi}\sigma)^n} exp - \frac{1}{2\sigma^2} \sum_{i=1}^{n} e_i^2$$

And when below equation is satisfied, we could get the OLS estimator for β. Our goal is to minimize sum of squares for residuals that is

$$\sum_{i=1}^{n} e_i^2 \to \min \tag{2.3.1}$$

The matrix form of the expression of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= arg\min_{\beta}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \end{aligned}$$

2.3.2 M-Estimator

M-estimators are a large class of estimators which are obtained from a given data set by minimizing the sum of certain functions of the data. Basically, an M-estimation is derived from statistical function equals to zero, like finding the value of a parameter allowing the maximum-likelihood function equals to zero. This property makes the M-estimator could be widely used.

Least squares estimator is the prototype of M-estimator, as defined by (2.3.1). Another typical type of typical case is the maximum-likelihood estimator, which could be defined as below:

$$\hat{\theta} = arg\min_{\theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right) \quad or \quad \hat{\theta} = arg\min_{\theta} \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right)$$

The above form with ρ function is the ρ-type M-estimator, where ρ is a function and the minimization can always be calculate directly. If ρ is differentiable, instead of solving ρ, we can implement it as an iterated re-weighted least-squares one, which is another type of M-estimator, ψ-type.

Assuming the ρ-type function is

$$\min \sum_{i} \rho(x_i, \beta)$$

when β is the $p \times 1$ parameter vector, then then ψ-type function could be yielding as

$$\sum_{i} \psi(x_i) \frac{\partial x_i}{\partial \beta_j}, \quad \text{for } j = 1, 2, \ldots, p$$

where the derivative function $\psi(x) = d\rho(x)/dx$ is the influence function. Then the weight function could be defined as below

$$w(x) = \frac{\psi(x)}{x}$$

the ψ-type function becomes

$$\sum_i w(x_i) x_i \frac{\partial x_i}{\partial \beta_j} = 0, \quad \text{for } j = 1, 2, \ldots, p$$

and the object becomes to obtain the following iterated re-weighted least squares problem

$$\min \sum_i w(x_i^{(k-1)}) x_i^2$$

where k indicates the iterate number.

Below two are the M-estimators with different commonly used influence functions which would be compared with other robust estimators in this research.

Huber Estiamtor

Huber estimator, introduced by Huber (1964), is the most common general method of robust regression. It is considered nearly as effective as OLS method.

The influence function $\rho(x)$ of Huber method is

$$\rho_H(x) = \begin{cases} x^2/2 & \text{for } |x| \leq k \\ k(|x| - k/2) & \text{for } |x| \geq k \end{cases}$$

And the corresponding weight function is

$$w_H(x) = \begin{cases} 1 & \text{for } |x| \leq k \\ k/|x| & \text{for } |x| \geq k \end{cases}$$

where k is a constant.

Yohai MM Estimator

MM estimator was proposed by Yohai (1987) and it is an estimator with a high breakdown point and high efficiency under normal error (Stromberg, 1993), which is suitable for the assumption of the modes I would use.

The influence function $\rho(x)$ of MM method is a Tukey's biweight function:

$$\rho_{MM}(x) = \begin{cases} \frac{c^2}{6}(1 - [1 - (x/c)^2]^3) & \text{for } |x| \leq c \\ c^2/6 & \text{for } |x| > c \end{cases}$$

And the corresponding weight function is

$$w_{MM}(x) = \begin{cases} [1 - (\frac{x}{c})^2]^2 & \text{for } |x| \leq c \\ 0 & \text{for } |x| \geq c \end{cases}$$

where c is a constant.

2.3.3 Least Absolute Value Estimator(or L_1 Estimator)

Least Absolute Value estimator (L_1 estimator) is also known as Least Absolute Deviations(LAD), or sum of Absolute Deviations, the history of LAV is as almost as long as Least Square method (Portnoy and Koenker, 1997). It came from Edgeworth(1887), improving proposal of Boscovich. L_1 estimator is well-known to have significant robustness advantages. The main advantage of this method is that it is invariant under linear transformations which makes it coordinate free. It also allows for variances of ϵ to be different.(Chen et al., 2008)

Instead of minimizing the sum of squared errors, this estimator minimizes the sum of absolute values of errors. Strictly speaking, LAD estimator is also a M-estimator.

We will denote this estimator by β_A. The least sum of absolute values is

$$\hat{\beta}_{LAD} = arg \min_{\hat{\beta}} \sum_{i=1}^{n} |r_i|$$

Unfortunately, there is no close form solution for $\hat{\beta}_A$ and the method is not often adopted for a long time because of its computationally highly demanding. The situation has changed with the emergency of the simplex algorithm linear programming and the development of computer science. A brief historical introduction to L_1 estimator computation can be found in Dielman (2005).

2.3.4 Least Median of Squares Estimator

This estimator was proposed by Rousseuw in 1984. The idea is to minimize the median, instead of the sum of squared vertical distances of data points from the fitted regression line. We will denote this estimator by $\hat{\beta}_{LMS}$. Again, there is no close form solution for $\hat{\beta}_{LMS}$. However, it can be proved that $\hat{\beta}_{LMS}$ is the slope of the start line passing through a particular pair of data points. Thus, $\hat{\beta}_{LMS}$ can be computed by the following steps:

A. Find all straight lines that are determined by any two points in the data set. there are altogether C(n,2) lines.

B. For each of the C(n,2) lines, find the median of the squared vertical distances of the data points from the line. Thus, we will obtain C(n,2) such medians.

C. From the C(n,2) medians, find the smallest one and the corresponding straight line is our estimated regression line.

The computaional method stated above is based on the following theorems developed by Rousseeww (1984):

Theorem 1: There always exists a solution to

$$\min_{\alpha\beta} med_i(y_i - \alpha - \beta x_i)^2$$

Theorem 2: There exists a (α, β) such that least two of the observation satisfy $y_i = \alpha + \beta x_i$ exactly, then the LMS solution equals (α, β) whatever the other observations are.

2.3.5 Least Trimmed Squares Estimator

Rousseeuw (1984) raised least trimmed squares estimator(LTS) which can have a breakdown point up to 50%. LTS developed from OLS, it orders the squared residuals from the smallest to the largest and only takes the smaller 50% of the data. The estimator is denoted as β_{LTS} and it is defined

$$\hat{\beta}_{LTS} = arg\min_{\beta} S(\beta),$$

where $h = [n + k + 1]/2$,

$$S(\beta) = \sum_{i=1}^{h} r^2_{(i)},$$

$r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \ldots \leq r^2_{(n)}(\beta)$ are the ascendant squares of the residuals, $r_i(\beta) = y_i - x_i'\beta$. The above procedures interpret that after excluding the most 50% extreme positive or negative residuals, LTS is proceeded with OLS with the rest of data.

Chapter 3

Construction of Comparison

In order to raise the comparison, the first job is generating the sample data. In this chapter, design of the model with contamination and is introduced in Section (3.1). In the Section (3.2), I would mainly describe the characteristics of the data set produced from (3.1) and also give some rough test to the data and the validity of some assumptions for the model employed.

3.1 Model Designing

In the previous chapter, the definitions of outliers have been introduced. Before starting to present the model, I want to clear what an outlier(or a contaminant) refers to in this manuscript: an outlier is an observation which appears discordantly with the rest data points. On this basis, the first step is to assume the basic probability model which is used to generate the data set without outliers. Then, the model should be adjusted to incorporate some contaminants and part or all of them would appear as outliers, which means they are inconsistant with others. Barnett and Lewis (1984) concluded seven detail different form of contamination models, or outlier-generating models, which have been proposed and studied.

In order to get a sample with different pattern of outliers shown in the figure (2.1), the slippage model, which is one of the outlier generating models, is employed for not only is this the common type as a model for contamination, but also this model can encompass any number of contaminants in the data. According to Barnett and Lewis (1984), we can denote data x_i under the initial probability model as

$$x_i \sim N(\mu, \sigma^2)$$

If contaminants arising from a shift in mean, the slipped distribution could be expressed as

$$x_i \sim N(\mu + a, \sigma^2) \quad a \in \mathcal{N}$$

Contaminants arise from a shift in scale (or dispersion) could be written as

$$x_i \sim N(\mu, b\sigma^2) \quad b > 1$$

Then I am going to apply it to my regression model.

3.1.1 Basic Linear Regression Model with Contaminate Data

Basic linear model

Considering a set of independent observations $(x_i, y_i), i = 1, 2, \ldots, n$, with specified relationship

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{3.1.1}$$

where

(i) Y_i is the response variable in the ith trial;

(ii) X_i is the single explanatory variable, it is either a constant or a random variable measured without error; if X_i is a random variable, it is assumed to be independently identical distributed with $E(X_i) = 0$ and $\sigma^2(X_i) = \sigma_x^2$;

(iii) α and β are the parameters and are preassigned;

(iv) the error term ϵ_i is an random variable comes from $N(0, \sigma^2)$;

(v) $COV(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$, which means that ϵ_i are mutually uncorrelated;

(vi) $COV(X_i, \epsilon_j) = 0$ for all i , j.

Obviously, (3.1.1) is a classical univariate regression model with all main assumptions.

Contamination

The standard regression theory assumes that the error in the model is only associated with the response variable, the explanatory variable is measured error-free, like the equation 3.1.1. An extension to this theory is to assume that both the explanatory variable and the response variable have errors presenting.

This is the well-known errors-in-variables model. Kendall (1952) split it into two categories based on the independent variables:

1. The Functional Model - the independent variable is a mathematical (not a random) variable, taking unobservable values but with a fixed constant mean μ.

2. The Structural Model - the independent variable is a random variable with mean μ and variance σ^2.

Now, considering two variables, u_i and v_i, u_i is a random variable $E(u_i) = 0$ and $var(u_i) = \sigma_x^2$. u_i and v_i are linearly related in the form

$$v_i = \alpha + \beta u_i, \quad i = 1, 2, \ldots, n$$

However, instead of (u_i, v_i), we are getting data points

$$\begin{cases} x_i = u_i + \tau_i \\ y_i = v_i + \delta_i \end{cases} \tag{3.1.2}$$

where τ_i and δ_i are considered to be random error components, or noise. $E(\tau_i)$ and $E(\delta_i)$ are also assumed to be 0 and $var(\tau_i) = \sigma_\tau^2$, $var(\delta) = \sigma_\delta^2$ with $\sigma_\tau^2 > \sigma_x^2$, $\sigma_\delta^2 > \sigma_y^2$ separately. The expression 3.1.2 is the linear structural model. Also the errors δ_i and τ_i are mutually uncorrelated, which means that

$$cor(\tau_i, \delta_j) = 0, \quad i, j = 1, 2, \ldots, n$$

And the chi-square test shows that τ_i and δ_i are mutually independent, we can conclude that how x_i and y_i be outlying are uncorrelated. Other good reviews regarding this model could be found by the research of Gillard (2009).

Besides, as u_i and τ_i follows N(0,σ_x^2) and N(0,σ_τ^2) separately, x_i then follows N(0, $\sigma_x^2 + \sigma_\tau^2$). Same, we can get v_i follows N(0, $\beta\sigma_x^2$) and y_i follows N($\alpha, \beta\sigma_x^2 + \sigma_\delta^2$).

Model

After the brief discussion, consider the above two models and employed a mixture form of contamination of one or both of the error components τ and δ. Specially, τ

is either 0 with probability $(1-\gamma)$ or arises with probability γ from $N(0, \sigma_\tau^2)$. δ has a similarly mixture-type. Notice that the $n * \gamma$ data here are from a slippage model but not all of them are outliers.

Based on the mixture model, four types of data sets will be discussed.

	outlying on x-axis	outlying on y-axis
equally spaced	No	Yes
	No	Yes
normal distributed	Yes	No
	Yes	Yes

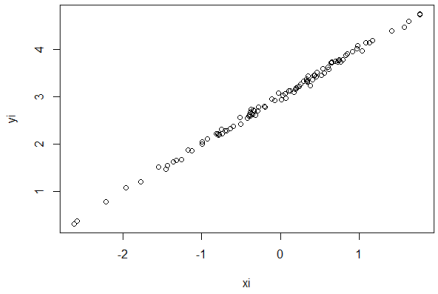
Each data set will be discuss under three types of sample size.

Breakdown Point

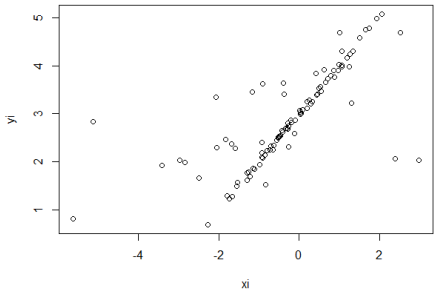
Breakdown point is one of the tools to describe and measure how resistant robust methods could be to the presence of outliers. It is widely used since it is introduced by Donoho (1982). We can see from the previous sections, it is mentioned that the breakdown points of some robust methods could be every high, even up to 50% and this means that the proportion of the contaminants could be up to 50% in the data set. The higher the breakdown point is, the more robust the method is. Given this, in each situation, the contaminate rate γ would be set as 10%, 30% and 50% to see how these robust estimators perform.

3.2 Data Properties

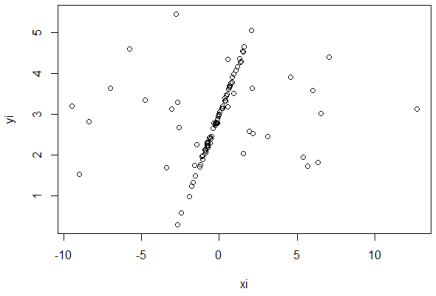
Now, let's see what the patterns of generating data look like. Below figures show when the sample size=100, contaminate rate=0.3, and when $\sigma_\tau = 2, 5$ and/or $\sigma_\delta = 2, 5$:



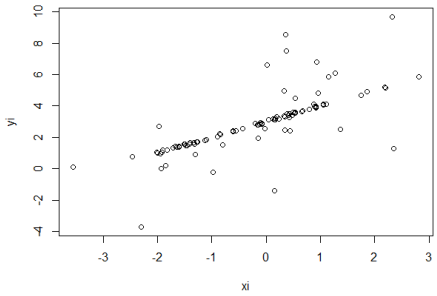
(a) $\sigma_\tau = \sigma_\delta = 0$



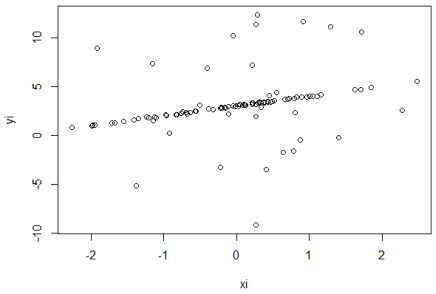
(b) $\sigma_\tau = 2, \sigma_\delta = 0$



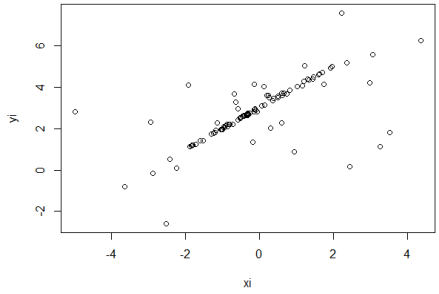
(c) $\sigma_\tau = 5, \sigma_\delta = 0$



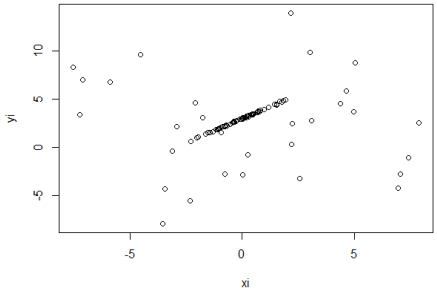
(d) $\sigma_\tau = 0, \sigma_\delta = 2$



(e) $\sigma_\tau = 0, \sigma_\delta = 5$



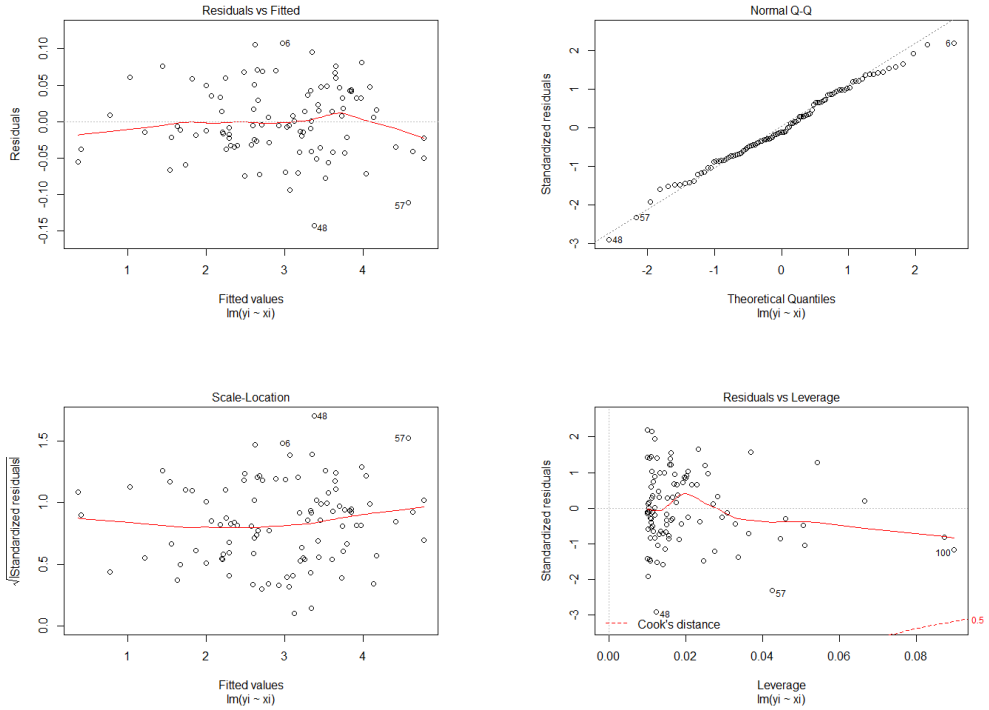
(f) $\sigma_\tau = \sigma_\delta = 2$



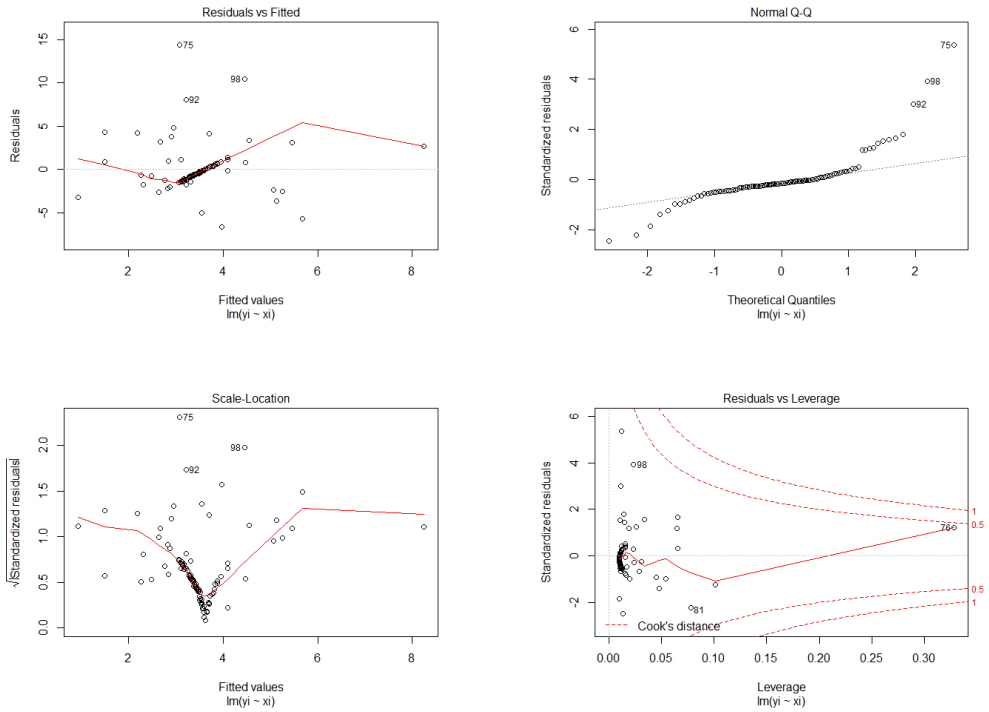
(g) $\sigma_\tau = \sigma_\delta = 5$

From the above scatter plot, we can see that the outlying data are spreading as the expectation. Through preassigning different values to σ_τ and σ_δ, we can see the more different patterns of outliers as we want.

Without contaminants ($\tau_i = 0$, $\delta_i = 0$), we can see that the data set is well satisfied the assumption of linear regression(still with n=100, γ=0.3):



Furthermore, when there is slippage in the model(like $\tau_i = 5$, $\delta_i = 5$ as an example), we can see the preliminary rough test result as below and the assumption for linear regression is no longer satisfied:



3.3 Comparison Procedure

People are either trying to detect outliers by the tests of discordance or in attempts to accommodate outliers through robust statistical procedures. In the first way, when the outliers are found out, the choice of how to deal with an outlier should depend on the cause, like retention or exclusion. In regression problems, an common approach could also be to only exclude points which exhibit a large degree of influence on the estimated coefficients, using a measure such as Cook's distance.

3.3.1 How to Deal with Outliers

In this section, I illustrate how the mentioned measurement are used to obtained the estimators. The procedure will follow the below three steps:

A. Compute the diagnostic measure, e.g. Studentized Residual, Cook's Distance for each sample point.

B. Delete the sample points whose diagnostic measures are larger than assigned percentage point of the distribution of the diagnostic influence function.

C. Use the OLS estimator from the remaining sample points of the data.

The estimator is denoted by $\hat{\beta}_D$ (including $\hat{\beta}_{CD}$ when Cook's Distance is applied and $\hat{\beta}_{SR}$ otherwise). Studentized residual is used when $\sigma_\tau^2 = \sigma_x^2$. It is possible that all the computed diagnostic measures are less than the assigned percentage points and therefore $\hat{\beta}_D$ is the same as the OLS estimator calculated from the original sample. A mathematical expression for $\hat{\beta}_D$ is

$$\hat{\beta}_D = \frac{\sum_{j=1}^{n-m}(X_{(j)} - \bar{X}_m)(Y_{(j)} - \bar{Y}_m)}{\sum_{j=1}^{n-m}(X_{(j)} - \bar{X}_m)^2}$$

where

(a) m is the number of deleted outliers and we assume that $m = n\gamma$, $0 \leq m \leq (n/2)$

(b) For each j, $j = 1, ..., n-m$, there exits some i, $i = 1, ..., n$ so that

$$x_{(j)} = x_i; \quad y_{(j)} = y_i$$

and

$$D_i < f_0 \quad if \quad \sigma_\tau^2 > \sigma_x^2$$

$$R_i < t_0 \quad if \quad \sigma_\tau^2 = \sigma_x^2$$

Here, D_i and R_i are the Cook's distance of the sample point (x_i, y_i) and f_0 and t_0 are the assigned percentage points of the corresponding F distribution and t distribution, respectively.

(c) $\bar{X}$ and $\bar{Y}$ are the sample means of the remaining data after deletion of the m outliers.

For fixed m, the values of $\bar{X}$ vary for different remaining data set of size $n-m$, there are C(n,m) possible values of $\bar{X}$ and $\bar{Y}$ for fixed m. The mathematical expression for D_i and R_i are

$$D_i = (\frac{n-2}{2})(\frac{e_i^2}{SSE})[\frac{h_{ii}}{(1-h_{ii})^2}]^{1/2}$$

and

$$R_i = e_i[\frac{n-3}{SSE(1-h_{ii}-e_i^2)}]^{1/2}$$

where

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\alpha}_L + \hat{\beta}_L x_i),$$

$$SSE = \sum_{i=1}^{n} e_i^2,$$

and

$$h_{ii} = \begin{bmatrix} 1 & x_i \end{bmatrix} \left[\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right]^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

It is difficult to choose a suitable percentage point f_0 or t_0 so that the corresponding estimator $\hat{\beta}_{DC}$ will be the best in the family of all estimators of the same type. However, in our comparison, we will use three different levels in each of the following two case:

(i) If $\sigma_\tau^2 = \sigma_x^2$ (so that we use R_i as the diagnostic measure), we use .95,$100(1-\frac{.05}{n})$ and $100(1-\frac{0.1}{n})$ and we denote the corresponding estimators by $\hat{\beta}_{SR}(.95)$,$\hat{\beta}_{SR}(1-\frac{.05}{n})$,$\hat{\beta}_{SR}(1-\frac{.01}{n})$,respectively.

(ii) If $\sigma_\tau^2 > \sigma_x^2$ (so that we use D_i as the diagnostic measure). We use 20, 50,and $100(1-\frac{.5}{n})$,and we denote the corresponding estimators by $\hat{\beta}_{CD}(.20)$, $\hat{\beta}_{CD}(.05)$,$\hat{\beta}_{CD}(1-\frac{.5}{n})$,respectively.

Remark: The high percentages $100(1-\frac{.05}{n})$ and $100(1-\frac{.01}{n})$ in cash (i) and $100(1-\frac{.5}{n})$ in case (ii) in consideration of the viewpoint of simultaneous testing (see Weisberg (1980) or Myers (1990)).

3.3.2 Steps for Computing MSE

Now the mean square errors (MSE) of all estimators discussed in the previous section is going to be computed.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_i - \beta_i)^2$$

The comparison is carried out by Monte Carlo simulations which consist of the following steps:

1. Using software R, n data points $(x_i, y_i), n = 1, 2, \ldots, n$ are generating following the method mentioned in section 2.3.

2. From the generated data in Step 1, each estimator is calculated.

3. Repeat the above two steps ten thousand times with fixed value of α,β,σ_x^2,σ_τ^2,σ_δ^2.

4. Compute the average of the ten thousand $\hat{\beta}_i$, namely,

$$\bar{\hat{\beta}} = \sum_{i=1}^{10000} \frac{\hat{\beta}_i}{10000}$$

.

5. The Monte Carlo estimator of bias, variance and mean square error of $\hat{\beta}$ are computed as

$$\widehat{Bias(\hat{\beta})} = \bar{\hat{\beta}} - \beta$$

,

$$\widehat{Var(\hat{\beta})} = \sum_{i=1}^{10000} \frac{(\hat{\beta} - \bar{\hat{\beta}})^2}{(10000 - 1)}$$

$$\widehat{MSE(\hat{\beta})} = \widehat{Bias(\hat{\beta})} + \widehat{Var(\hat{\beta})}$$

Using the above procedure, we obtain the Monte Carlo MSE of $\hat{\beta}_L$, $\hat{\beta}_{SR}$, $\hat{\beta}_{CD}$, $\hat{\beta}_{LAD}$, $\hat{\beta}_{LMS}$, $\hat{\beta}_{LTS}$ and two M-estimators for different parameter values combinations.

Chapter 4

Simulation Studies

The following values are chosen in our simulation:

$$\beta = 0(0.2)2(2)10(10)50;$$

($\beta \geq 0$ is assumed for the sake of convenience)

$$n = 10, 100, 500;$$

$$\gamma = 0.1, 0.3, 0.5;$$

$$\sigma_x^2 = 1(1)10(5)50;$$

$$\sigma_y^2 = 1(1)10(5)100;$$

$$\sigma_\tau^2 = 1(1)20(5)100(5)1000;$$

$$\sigma_\delta^2 = 1(1)20(5)200(5)1000.$$

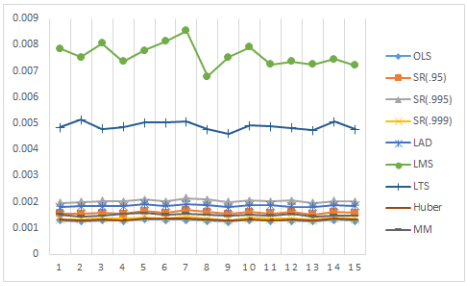
In the previous chapter, I explained the MSE values for different robust estimators for the all preassigned β. This Chapter is the interpretation of the outcome of the

simulation. In order to facilitate the comparison, I would present some tables and graphs with typical properties of four types of data patterns. Due to limitation of space, only certain combination of parameters simulation are presented and if not mentioned specifically, other charts under the same condition have the same characteristics, like the simulation outcome with sample size of 500 is exactly same to the sample size 100 but the relative magnitude of the value of MSE; and the extreme large σ_τ and σ_δ result in extraordinary large of MSE which could not be shown appropriately. The R code used to get all the tables is attached in the appendix.

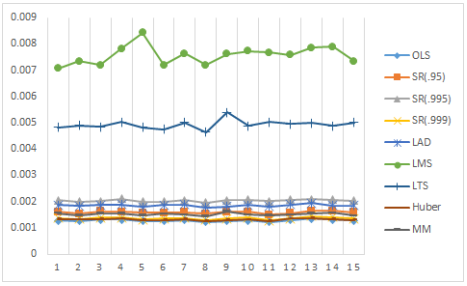
The summary of interpretation of the simulation outcome is made from four angles: the different outlying levels, various values of β, the different contaminant rates and different sample sizes.

4.1 Data without Outliers

Before we start the simulation with the contaminated data, all methods are applied to a data set without outliers to test their performance. The model is $y_i = 3+\beta x_i+e_i$, x_i is either equally spaced in [.5,5] or normally distributed following N(0,1). Below charts shows the results and from the value of y-axis, we can tell that all of them have a good performance. If we have to make a comparison, we could say $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$ are slightly weaker.



(a) x_i is equally spaced



(b) x_i is normally distributed

Figure 4.1

4.2 Data Outlying in y-space (I)

The first simulation is about the case that all x_i are constants and equally spaced in [.5,5]. The model is $y_i = 3 + \beta x_i + e_i$ with the contaminating rate γ=30%, n=10 or 100 or 500, $\beta = 0(.2)1(2)10(10)50$. The influential data are outlying only in the $y - axis$ direction and $\sigma_\delta = 1(1)20(5)200(5)1000$. Studentized residual is used to detect outliers in this experiment.

(1) Changing β

When the contaminants are only outlying in y-space and x_i is equally spaced constant, all robust methods don't show any sensitive to the value of β. The corresponding outcome are like the tables (4.1) and (4.2).

(2) Sample Size

In this experiment, as all data spread in a certain interval, the larger sample size means the closer x_is are. From the table (4.1) to (4.4), we can see that when the sample size expands from 10 to 100, β_L, $\beta_{SR(.999)}$ become less robust, $\beta_{SR(.95)}$ performs a little bit better and the other estimators are resistant to the change of sample size.

Table 4.1: MSEs of $\hat{\beta}$ (σ_y=1,σ_δ=4,x_i is equally spaced in [.5,5], γ=0.3, n=10)

β	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.1883	0.0025	0.0007	0.1883	0.0015	0.0027	0.0020	0.0015	0.0008
0.2	0.1936	0.0025	0.0007	0.1936	0.0016	0.0026	0.0020	0.0016	0.0008
0.4	0.1934	0.0025	0.0007	0.1934	0.0015	0.0026	0.0020	0.0015	0.0007
0.6	0.1976	0.0024	0.0007	0.1976	0.0015	0.0026	0.0019	0.0015	0.0008
0.8	0.1915	0.0024	0.0007	0.1915	0.0015	0.0027	0.0020	0.0016	0.0008
1	0.1925	0.0024	0.0007	0.1925	0.0016	0.0026	0.0020	0.0015	0.0008
2	0.1896	0.0025	0.0007	0.1896	0.0015	0.0026	0.0020	0.0016	0.0008
4	0.1874	0.0025	0.0007	0.1874	0.0016	0.0027	0.0020	0.0016	0.0008
6	0.1914	0.0025	0.0007	0.1914	0.0015	0.0028	0.0020	0.0016	0.0008
8	0.1910	0.0025	0.0007	0.1910	0.0015	0.0026	0.0019	0.0015	0.0007
10	0.1882	0.0024	0.0007	0.1882	0.0015	0.0027	0.0020	0.0015	0.0008
20	0.1880	0.0025	0.0007	0.1880	0.0015	0.0026	0.0019	0.0016	0.0008
30	0.1895	0.0025	0.0007	0.1895	0.0016	0.0027	0.0020	0.0015	0.0008
40	0.1903	0.0025	0.0007	0.1903	0.0015	0.0027	0.0020	0.0015	0.0008
50	0.1888	0.0025	0.0007	0.1888	0.0015	0.0027	0.0020	0.0016	0.0008

Table 4.2: MSEs of $\hat{\beta}$ (σ_y=1,σ_δ=8,x_i is equally spaced in [.5,5], γ=0.3, n=100)

β	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	1.4770	1.4123	1.3715	1.4770	0.3623	0.0040	0.0005	0.0058	0.0002
0.2	1.4776	1.4058	1.3644	1.4776	0.3463	0.0040	0.0005	0.0054	0.0002
0.4	1.4677	1.4034	1.3621	1.4677	0.3577	0.0040	0.0005	0.0058	0.0002
0.6	1.4413	1.3721	1.3185	1.4413	0.3340	0.0040	0.0005	0.0058	0.0002
0.8	1.4815	1.4164	1.3270	1.4815	0.3539	0.0040	0.0005	0.0058	0.0002
1	1.4354	1.3822	1.3304	1.4354	0.3289	0.0040	0.0005	0.0057	0.0002
2	1.4553	1.3891	1.3301	1.4553	0.3400	0.0039	0.0005	0.0056	0.0002
4	1.4958	1.4360	1.3358	1.4958	0.3515	0.0041	0.0005	0.0052	0.0002
6	1.4134	1.3553	1.3160	1.4134	0.3198	0.0041	0.0005	0.0058	0.0002
8	1.4337	1.3790	1.3277	1.4337	0.3435	0.0040	0.0005	0.0059	0.0002
10	1.4317	1.3709	1.3182	1.4317	0.3129	0.0039	0.0005	0.0057	0.0002
20	1.4761	1.4109	1.3221	1.4761	0.3548	0.0040	0.0005	0.0056	0.0002
30	1.4826	1.4146	1.3058	1.4826	0.3433	0.0041	0.0005	0.0061	0.0002
40	1.4319	1.3737	1.3176	1.4319	0.3387	0.0040	0.0005	0.0052	0.0002
50	1.4353	1.3774	1.3706	1.4353	0.3529	0.0040	0.0005	0.0051	0.0002

(3) Increasing Outlying Level

Since there is no significant difference between the MSE with respect to the parameter β, I take the average of the MSE of each robust estimator to see how MSE responds to the change of the value of σ_δ.

It can be seen that both $\hat{\beta}_M$ and $\hat{\beta}_A$ work well when x_i is equally spaced no matter the sample size is small or large. The studentized residual method looks having no advantage compared to OLS method in all combination of parameters. so we can take the average of the MSE of each robust estimator to see how MSE change responding to the change of σ_δ.

Table 4.3: MSEs of robust estimators, x_i equally spaced, n=10

σ_δ	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
1	0.0123	0.0046	0.0008	0.0123	0.0014	0.0029	0.0021	0.0014	0.0009
2	0.0478	0.0039	0.0007	0.0478	0.0015	0.0028	0.0020	0.0015	0.0008
4	0.1907	0.0025	0.0007	0.1907	0.0015	0.0027	0.0020	0.0015	0.0008
6	0.4298	0.0019	0.0007	0.4298	0.0015	0.0026	0.0020	0.0016	0.0007
8	0.4298	0.0019	0.0007	0.4298	0.0015	0.0026	0.0019	0.0016	0.0007
10	1.1866	0.0014	0.0007	1.1866	0.0016	0.0026	0.0019	0.0016	0.0007
20	4.7740	0.0011	0.0007	4.7740	0.0016	0.0026	0.0019	0.0016	0.0007

Table 4.4: MSEs of robust estimators, x_i equally spaced, n=100

σ_δ	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
1	0.0231	0.0225	0.0216	0.0231	0.0087	0.0064	0.0007	0.0008	0.0006
2	0.0912	0.0877	0.0837	0.0912	0.0250	0.0053	0.0006	0.0015	0.0005
4	0.3668	0.3513	0.3366	0.3668	0.0893	0.0045	0.0005	0.0027	0.0003
6	0.8184	0.7843	0.7524	0.8184	0.1945	0.0042	0.0005	0.0039	0.0003
8	1.4557	1.3933	1.3345	1.4557	0.3427	0.0040	0.0005	0.0056	0.0002
10	2.2684	2.1696	2.0796	2.2684	0.5318	0.0039	0.0005	0.0074	0.0002
20	9.1024	8.6919	8.3200	9.1024	2.1326	0.0038	0.0005	0.0213	0.0002

4.3 Data Outlying in y-space (II)

Similar to the previous experiment, the model is still $y_i = 3 + \beta x_i + e_i$ with sample size n=10,100,500, $\beta = 0(.2)1(2)10(10)50$. The different settings are that x_i is assumed to follow normal distribution $N(0, 1)$, the contaminating rates are 10%,30% and 50% separately. As there is still no data outlying in x-space, we continue using the studentized residuals as the adjustment to OLS.

(1) Changing β

Same as the last example, tables (4.5) and (4.6) demonstrate that all simulated robust methods don't react to the change of the value of β.

(2) Sample Size

The increase of the sample size only improve the accuracy of the regression of the coefficients. Different sample size doesn't have substantial impact on the performance of the robust methods.

(3) Different contaminating rates

As MSE isn't sensitive to the change of the value of β, the average of MSE under all β values would be employed again to show how MSE responds to the different outlying levels and contaminating rates. The figure (4.2) shows the comparison outcome where the x-axis represents the robust estimation methods. Different lines represent different value of σ_δ and the y-axis represents the value of the average of MSEs.

From these line charts, we can see that $\hat{\beta}_{Huber}$ and $\hat{\beta}_{SR(0.995)}$ are vulnerable to the contaminating rate, but enlarging the sample size could well offset this influence to $\hat{\beta}_{Huber}$. Although when the sample size is small and the contaminating rate increases

Table 4.5: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=0,σ_δ=2, x_i is normal distributed, n=10, γ=0.1)

β	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_A$	$\hat{\beta}_M$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.0599	0.0422	0.0156	0.0546	0.0039	0.0102	0.0068	0.0045	0.0021
0.2	0.0591	0.0447	0.0179	0.0571	0.0041	0.0137	0.0063	0.0049	0.0021
0.4	0.0594	0.0431	0.0177	0.0562	0.0039	0.0114	0.0061	0.0048	0.0020
0.6	0.0573	0.0432	0.0170	0.0542	0.0034	0.0100	0.0081	0.0050	0.0021
0.8	0.0612	0.0406	0.0144	0.0553	0.0035	0.0095	0.0066	0.0042	0.0021
1	0.0606	0.0432	0.0157	0.0551	0.0035	0.0110	0.0062	0.0044	0.0021
2	0.0590	0.0405	0.0149	0.0546	0.0036	0.0100	0.0063	0.0045	0.0020
4	0.0583	0.0384	0.0136	0.0533	0.0035	0.0112	0.0063	0.0044	0.0020
6	0.0581	0.0423	0.0152	0.0550	0.0035	0.0102	0.0065	0.0045	0.0021
8	0.0583	0.0411	0.0164	0.0542	0.0036	0.0104	0.0064	0.0049	0.0020
10	0.0588	0.0445	0.0138	0.0549	0.0039	0.0103	0.0068	0.0045	0.0021
20	0.0565	0.0432	0.0163	0.0551	0.0033	0.0106	0.0070	0.0039	0.0020
30	0.0577	0.0432	0.0168	0.0529	0.0038	0.0103	0.0063	0.0049	0.0021
40	0.0564	0.0431	0.0160	0.0525	0.0035	0.0093	0.0067	0.0042	0.0022
50	0.0599	0.0413	0.0153	0.0534	0.0038	0.0117	0.0064	0.0042	0.0022

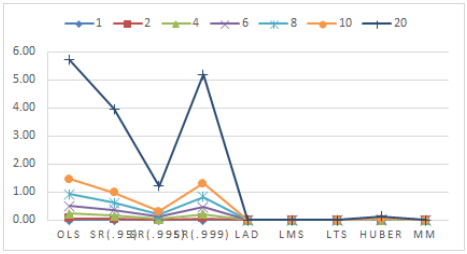
Table 4.6: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=0,σ_δ=10,x_i is normal distributed, n=100, γ=0.5)

β	$\hat{\beta}_L$	$\hat{\beta}_{SR(.95)}$	$\hat{\beta}_{SR(.995)}$	$\hat{\beta}_{SR(.999)}$	$\hat{\beta}_A$	$\hat{\beta}_M$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.3432	0.3432	0.3439	0.3432	0.0008	0.0009	0.0003	0.0323	0.0047
0.2	0.3260	0.3260	0.3268	0.3260	0.0008	0.0009	0.0003	0.0290	0.0047
0.4	0.3286	0.3287	0.3293	0.3286	0.0008	0.0009	0.0003	0.0286	0.0048
0.6	0.3298	0.3298	0.3297	0.3298	0.0008	0.0009	0.0003	0.0291	0.0047
0.8	0.3352	0.3352	0.3356	0.3352	0.0008	0.0009	0.0003	0.0308	0.0045
1	0.3334	0.3334	0.3327	0.3334	0.0008	0.0009	0.0003	0.0292	0.0044
2	0.3300	0.3300	0.3294	0.3300	0.0008	0.0009	0.0003	0.0290	0.0047
4	0.3234	0.3234	0.3234	0.3234	0.0008	0.0009	0.0003	0.0279	0.0046
6	0.3303	0.3303	0.3301	0.3303	0.0008	0.0009	0.0003	0.0290	0.0045
8	0.3290	0.3290	0.3290	0.3290	0.0008	0.0009	0.0003	0.0298	0.0048
10	0.3324	0.3324	0.3320	0.3324	0.0007	0.0009	0.0003	0.0289	0.0046
20	0.3286	0.3285	0.3290	0.3286	0.0008	0.0009	0.0003	0.0275	0.0047
30	0.3266	0.3266	0.3265	0.3266	0.0008	0.0009	0.0003	0.0302	0.0045
40	0.3288	0.3288	0.3281	0.3288	0.0007	0.0009	0.0003	0.0290	0.0048
50	0.3274	0.3274	0.3274	0.3274	0.0008	0.0009	0.0003	0.0289	0.0046

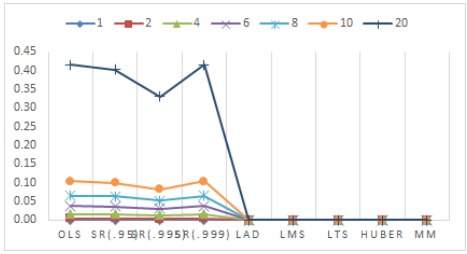
to 50%, MSEs of $\hat{\beta}_{LTS}$ and $\hat{\beta}_{MM}$ become larger, they are still very robust in other cases. What is worthy to be mentioned is no matter how the situation varies, $\hat{\beta}_{LMS}$ always has a good performance.

(4) Increasing Outlying Level

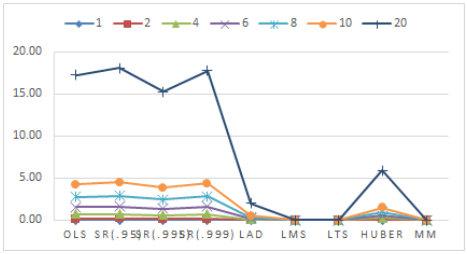
From the figure (4.2), we can also see that $\hat{\beta}_{LMS}$, $\hat{\beta}_{LTS}$ and $\hat{\beta}_{MM}$ are not affected by the increasing value of σ_δ, when the sample size is large, $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$ also become more resistant to the outliers.



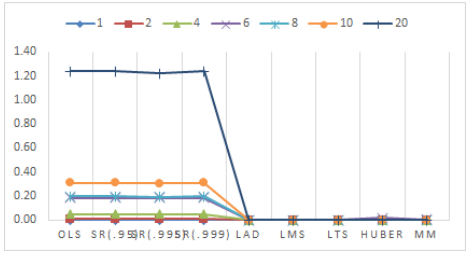
(a) n=10,$\gamma = 10\%$



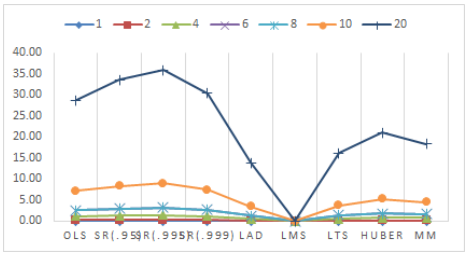
(b) n=100,$\gamma = 10\%$



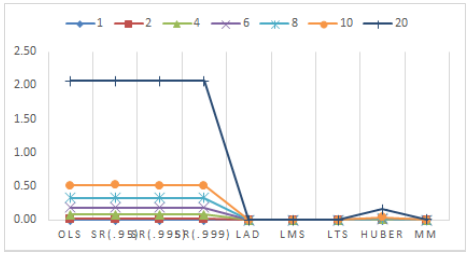
(c) n=10,$\gamma = 30\%$



(d) n=100,$\gamma = 30\%$



(e) n=10,$\gamma = 50\%$



(f) n=100,$\gamma = 50\%$

Figure 4.2: σ_x=σ_y=1,σ_τ=0

Summary

In conclusion, we could say that when the outliers only present in y-space, $\hat{\beta}_{LMS}$, $\hat{\beta}_{LTS}$ and $\hat{\beta}_{MM}$ are better than $\hat{\beta}_{LAD}$, $\hat{\beta}_{Huber}$ than the other methods derived from OLS.

4.4 Data Outlying in Both x-space and y-space

In this series of experiments, the case that the influential data are outlying in both $x-axis$ and $y-axis$ directions. The underlying model is $y_i = 3 + \beta x_i + e_i$, where x_i are normal random variables with $E(X)$=0 and σ_x^2=1,and $\beta = 0(.2)1(2)10(10)50$. The contaminating rates are $\gamma = 10\%$ or 30% or 50%, $n = 10, 100, 500$, σ_x^2 and σ_y^2 vary from 1 to 1000. As $\sigma_\tau \neq \sigma_x$, rather than Studentized residual, Cook's Distance is employed as the detection method for outliers.

Same as the previous example, there are many possible combinations of parameter values in this series of experiments, only those representative ones are presented.

(1) Changing β

First, we are going to see whether the value of β affect the performs of the robust estimators or not when the specific outlying points, leverage, exist. We can see from table (4.7) and (4.8), it is evident that only $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$ are still resistant to the existence of leverage. For some reason, in the second table, $\hat{\beta}_{MM}$ becomes no longer robust, further discussion about this would be presented in the following text.

Table 4.7: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=σ_δ=8,x_i is normal distributed, n=10, γ=0.1)

β	$\hat{\beta}_L$	$\hat{\beta}_{CD(.95)}$	$\hat{\beta}_{CD(.995)}$	$\hat{\beta}_{CD(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	1.11	0.12	0.02	0.30	0.37	0.01	0.01	0.60	0.00
0.2	1.13	0.13	0.03	0.31	0.37	0.01	0.01	0.62	0.00
0.4	1.18	0.14	0.03	0.29	0.44	0.01	0.01	0.67	0.00
0.6	1.32	0.13	0.03	0.32	0.52	0.01	0.01	0.76	0.00
0.8	1.51	0.14	0.03	0.33	0.72	0.01	0.01	0.91	0.00
1	1.66	0.15	0.03	0.33	0.80	0.01	0.01	1.06	0.00
2	3.29	0.17	0.03	0.39	2.13	0.01	0.01	2.44	0.00
4	9.56	0.23	0.05	0.65	6.92	0.01	0.01	8.14	0.00
6	20.38	0.36	0.11	1.07	15.49	0.01	0.01	18.04	0.00
8	35.21	0.50	0.08	1.59	27.51	0.01	0.01	31.13	0.00
10	54.71	0.72	0.09	2.45	42.62	0.01	0.01	49.32	0.00
20	216.94	2.01	0.32	8.90	169.84	0.01	0.01	188.80	0.00
30	483.96	4.70	0.56	19.26	380.40	0.01	0.01	427.80	0.00
40	845.46	10.28	1.01	36.22	655.80	0.01	0.01	753.49	0.00
50	1329.04	14.35	1.20	55.04	1041.63	0.01	0.01	1168.31	0.00

Table 4.8: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=σ_δ=4,x_i is normal distributed, n=100, γ=0.5)

β	$\hat{\beta}_L$	$\hat{\beta}_{CD(.95)}$	$\hat{\beta}_{CD(.995)}$	$\hat{\beta}_{CD(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.02	0.02	0.02	0.02	0.01	0.00	0.00	0.01	0.02
0.2	0.05	0.05	0.05	0.05	0.02	0.00	0.00	0.04	0.02
0.4	0.14	0.14	0.14	0.14	0.05	0.00	0.00	0.11	0.02
0.6	0.30	0.30	0.30	0.30	0.15	0.00	0.00	0.24	0.03
0.8	0.53	0.53	0.52	0.53	0.30	0.00	0.00	0.44	0.04
1	0.81	0.81	0.80	0.81	0.52	0.00	0.00	0.69	0.06
2	3.17	3.17	3.16	3.17	2.63	0.00	0.00	2.94	0.17
4	12.65	12.65	12.61	12.65	11.80	0.00	0.00	12.22	0.50
6	28.37	28.37	28.30	28.37	27.19	0.00	0.00	27.82	1.18
8	50.57	50.56	50.47	50.57	49.10	0.00	0.00	49.79	2.07
10	78.94	78.93	78.75	78.94	77.04	0.00	0.00	77.94	3.07
20	315.29	315.28	314.77	315.29	310.17	0.00	0.00	312.87	11.91
30	709.86	709.78	708.57	709.86	699.86	0.00	0.01	703.88	26.40
40	1260.78	1260.70	1258.66	1260.78	1242.08	0.00	0.01	1252.98	50.42
50	1968.91	1968.75	1965.12	1968.91	1940.37	0.00	0.02	1958.76	79.06

(2) Sample Size and Contaminating rate

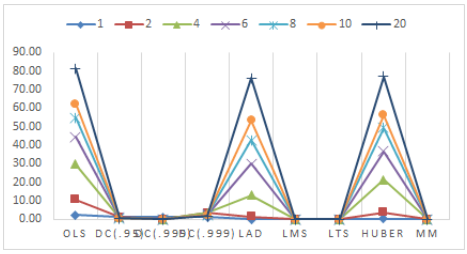
As some of MSEs of different robust estimators are changing with the value of β, MSE of $\beta = 10$ is selected to make the below line charts to make it comparable from other angles. From figure (4.3), we can see that the larger sample size doesn't decrease the relative magnitude of MSE as expected, but it did change the behaviour of some estimators. When the contaminating rate is as large as 50%, the large sample size did improve the performance of $\hat{\beta}_{LTS}$ and $\hat{\beta}_{MM}$. However, when the contaminating is small, $\hat{\beta}_{CD(.95)}$ and $\hat{\beta}_{CD(.995)}$ show more robust with smaller sample size.

(3) Increasing Outlying Level

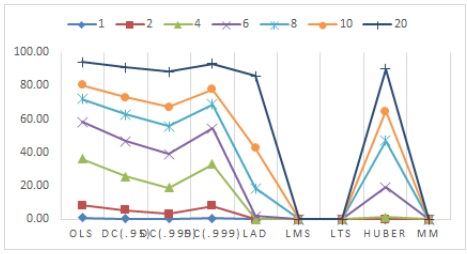
From the figure (4.3), except those exceptions mentioned in the last paragraph, we can see that all the estimators could be classified into two group. One group includes $\hat{\beta}$ related to OLS, $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$, they are sensitive to the slippage of the distribution. The other group includes the rests whose performance are quite independent from the value of σ_δ. Specifically, if the contaminating rate is extreme large and the sample size is not big enough, $\hat{\beta}_{MM}$ would be highly influenced by the leverage.

Summary

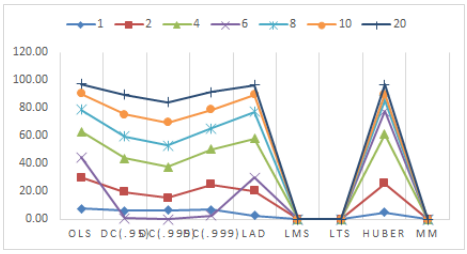
A rough summary could be made in the section that $\hat{\beta}_{LMS}$ is still the best robust estimator, the second would be $\hat{\beta}_{LTS}$ and next would be $\hat{\beta}_{MM}$.



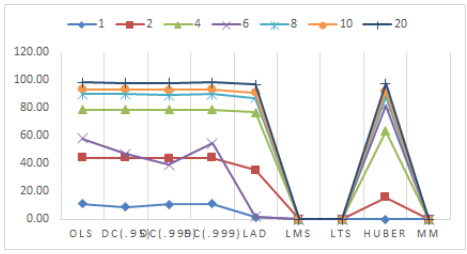
(a) n=10,$\gamma = 10\%$



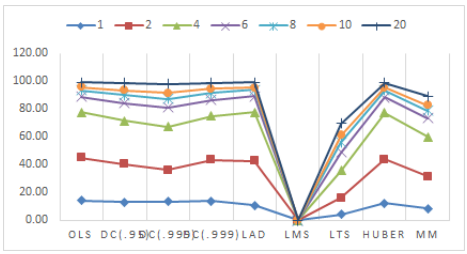
(b) n=100,$\gamma = 10\%$



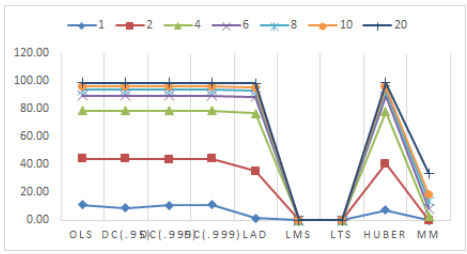
(c) n=10,$\gamma = 30\%$



(d) n=100,$\gamma = 30\%$



(e) n=10,$\gamma = 50\%$



(f) n=100,$\gamma = 50\%$

Figure 4.3: σ_x=σ_y=1,σ_τ=σ_δ=1(2)10,20

4.5 Data Outlying in x-space

The series of experiments in this section are data sets with outliers only in x-space. All settings are the same as section (4.4) except σ_δ=0.

(1) Changing β

Similar to the following two tables, we can actually see from all the tables resulting from data sets with outliers only in x space that only $\hat{\beta}_{MM}$, $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$ work and others are destroyed. Among these ineffective robust estimator, we can see that

Cook's Distance did work but not obviously. Actually, when $\sigma_\tau/\sigma_x > 100$, $\hat{\beta}_{CD}$ has way smaller MSE than others, however, in this case, $\sigma_\tau^2/\sigma_x^2 > 10000$, which doesn't make sense.

Table 4.9: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=10,σ_δ=0,x_i is normal distributed, n=10, γ=0.3)

β	$\hat{\beta}_L$	$\hat{\beta}_{CD(.95)}$	$\hat{\beta}_{CD(.995)}$	$\hat{\beta}_{CD(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.04	0.03	0.03	0.03	0.04	0.02	0.02	0.04	0.02
0.4	0.14	0.12	0.11	0.13	0.15	0.03	0.02	0.14	0.03
0.6	0.32	0.27	0.26	0.28	0.33	0.03	0.02	0.33	0.02
0.8	0.58	0.48	0.45	0.50	0.58	0.03	0.01	0.58	0.01
1	0.90	0.76	0.71	0.79	0.91	0.02	0.01	0.90	0.01
2	3.61	3.02	2.82	3.17	3.64	0.02	0.01	3.62	0.00
4	14.38	11.98	11.23	12.57	14.47	0.01	0.01	14.41	0.00
6	32.39	27.16	25.43	28.40	32.59	0.01	0.01	32.47	0.00
8	57.61	48.35	45.19	50.35	58.09	0.01	0.00	57.81	0.00
10	90.09	75.45	70.80	79.06	91.03	0.01	0.01	90.35	0.00
20	359.90	302.29	284.45	315.17	363.36	0.01	0.00	361.19	0.00
30	809.50	679.84	636.60	711.54	817.32	0.01	0.00	812.09	0.00
40	1436.18	1205.29	1126.76	1262.28	1443.48	0.01	0.01	1440.65	0.00
50	2248.16	1878.24	1759.49	1968.80	2269.38	0.01	0.00	2256.37	0.00

(2) Sample Size

Line charts in the figure (4.4) are also under the condition β=10. It's interesting that figure (4.4) looks so close to figure (4.3), even from the tables, we can only see slight difference between numbers, which means that the influence of outlying in x-space in linear regression is much stronger then the outlying in y-space. Sample size only affects the performance of estimators related OLS in a opposite way.

Table 4.10: MSEs of $\hat{\beta}$ (σ_x=1,σ_τ=8,σ_δ=0,x_i is normal distributed, n=100, γ=0.1)

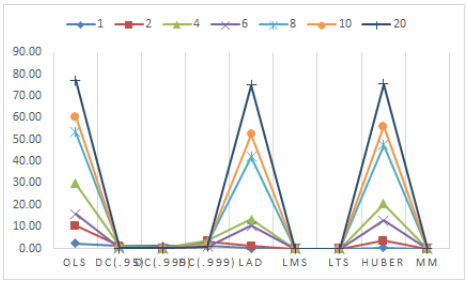
β	$\hat{\beta}_L$	$\hat{\beta}_{CD(.95)}$	$\hat{\beta}_{CD(.995)}$	$\hat{\beta}_{CD(.999)}$	$\hat{\beta}_{LAD}$	$\hat{\beta}_{LMS}$	$\hat{\beta}_{LTS}$	$\hat{\beta}_{Huber}$	$\hat{\beta}_{MM}$
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.02	0.02	0.02	0.02	0.01	0.00	0.00	0.02	0.00
0.4	0.09	0.08	0.06	0.09	0.02	0.00	0.00	0.05	0.00
0.6	0.21	0.17	0.14	0.20	0.02	0.00	0.00	0.09	0.00
0.8	0.37	0.30	0.25	0.35	0.03	0.00	0.00	0.15	0.00
1	0.58	0.47	0.39	0.55	0.04	0.00	0.00	0.23	0.00
2	2.35	1.91	1.60	2.20	0.13	0.00	0.00	0.89	0.00
4	9.30	7.50	6.25	8.70	0.38	0.00	0.00	3.30	0.00
6	20.99	17.01	14.20	19.68	0.85	0.00	0.00	7.50	0.00
8	37.24	29.98	25.01	34.82	1.44	0.00	0.00	13.18	0.00
10	58.30	46.97	39.20	54.65	2.46	0.00	0.00	20.83	0.00
20	234.01	189.05	158.29	219.07	9.81	0.00	0.00	84.03	0.00
30	522.60	420.06	350.50	488.71	20.38	0.00	0.00	182.45	0.00
40	930.15	753.09	626.72	871.76	37.14	0.00	0.00	329.17	0.00
50	1458.82	1175.41	983.00	1364.24	55.51	0.00	0.00	519.81	0.00

(3) Different contaminating rate

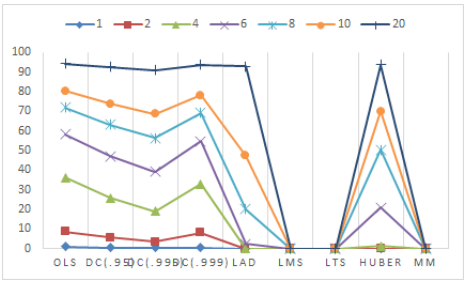
From the perspective of different contaminating lever, we can also see that $\hat{\beta}_{DC}$ is not appropriate with high contaminating rates, so as $\hat{\beta}_{LAD}$ and $\hat{\beta}_{MM}$.

(4) Increasing Outlying Level

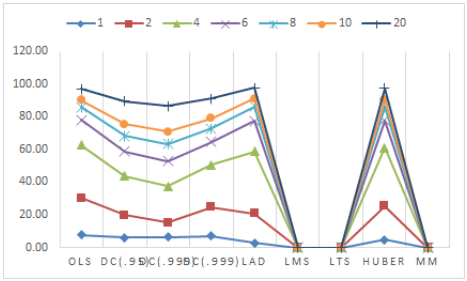
According to the simulation results, we can sort the robust estimators from best to worst like: $\hat{\beta}_{LMS}, \hat{\beta}_{LTS} > \hat{\beta}_{MM} > \hat{\beta}_{LAD}, \hat{\beta}_{Huber} > \hat{\beta}_{CD}$ (">" means better).



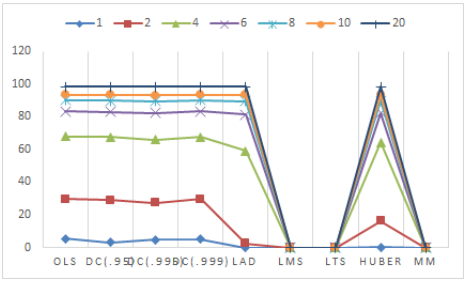
(a) n=10,$\gamma = 10\%$



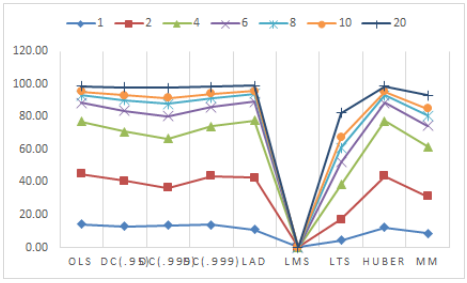
(b) n=100,$\gamma = 10\%$



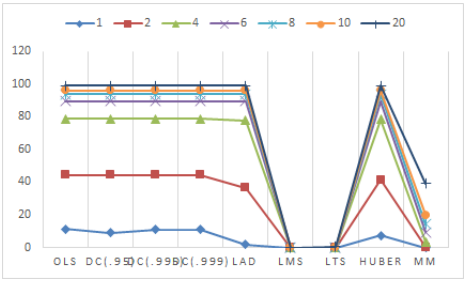
(c) n=10,$\gamma = 30\%$



(d) n=100,$\gamma = 30\%$



(e) n=10,$\gamma = 50\%$



(f) n=100,$\gamma = 50\%$

Figure 4.4: σ_x=1,σ_δ=0,σ_τ=1(2)10,20

Chapter 5

Conclusion

After considering all combinations of the above parameter values, the below conclusion of the simulation is given:

1. When there is no outlying data, i.e. $\gamma = 0$, the OLS estimator is always the best estimator.

2. When $\gamma > 0$, there are two situations:

   (a) If $\sigma_\tau^2 = 0$, the spurious data are outlying only in the $y - axis$ direction. When x_i is equally spaced, if x_i are very close to each other, Studentized residual may not work, modern robust methods should be obtained. In the case of x_i is normal distributed, the OLS estimator $\hat{\beta}_L$ becomes more and more inefficient as σ_δ^2 increases, whereas both $\hat{\beta}_{LMS}$, $\hat{\beta}_{LTS}$ and $\hat{\beta}_{MM}$ become more and more efficient because of their robustness with respect to outliers (MSEs are decreasing). The estimator $\hat{\beta}_{LMS}$ seems to be the best one when the σ_δ^2 is extremely large under the small sample size whereas $\hat{\beta}_{LTS}$ is the best under large sample size.

(b) If $\sigma_\tau^2 > 0$, the spurious data are outlying in both $x-axis$ and $y-axis$ directions or only in $x-axis$. Then, the LAV estimator $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$ are no longer robust if β is large, whereas $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$ are still as robust as in previous case. When there are outliers exist in both x-space and y-space, the estimator $\hat{\beta}_{CD}$, which is obtained by using the Cook's distance as influence function, has a very good performance when both sample size is small and contaminating rate is low. It's much more robust than $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$ when β is large.

3. The robustness of $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$ depend on the ratio $\sigma_\tau^2 / \sigma_x^2$. The larger the ratio is, the less robust the estimators are. On the other hand, if the ratio is large, say less than 10, both $\hat{\beta}_{LAD}$ and $\hat{\beta}_{Huber}$ can be destroyed and the mean square errors of both estimators increase rapidly, so they should not be used.

4. The robustness of $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$ have very stable performance in all mentioned patterns of data.

5. For the three different levels of percentage point we used for the estimator β_{SR} and β_{CD} doesn't produce very good estimators, it looks like its benefit is only the simplicity and easy calculation.

At the end, we can conclude that

1. If the actual observations are consistent with model assumptions, all the selected robust estimators have a good performance.

2. If the actual observations are slightly break the assumed assumptions, $\hat{\beta}_{LAD}$,$\hat{\beta}_{LMS}$,$\hat{\beta}_{LTS}$, $\hat{\beta}_{Huber}$ and $\hat{\beta}_{MM}$ are still robust.

3. If the assumptions are badly broken, the perfect $\hat{\beta}_{LMS}$, $\hat{\beta}_{LTS}$ are still very robust, especially $\hat{\beta}_{LMS}$.

So if we have prior information how the data distributed, we can choose a robust estimator with easy calculation good behaviour. If no prior information is available, a conservative choice is $\hat{\beta}_{LMS}$, which has amazing performance under all conditions.

Appendix A

Appendix: Computer Simulation Codes

```
##The software used for simulation is R Version 3.2.2.
##Parameter Assignment
N=10000
n<-10
gamma<-.5
n2<-n*gamma
n1=n-n2
alpha<-3
beta<-c(seq(0,1,by=0.2),seq(2,10,by=2),seq(20,50,by=10))
sigmax<-1
sigmatau<-0
sigmay<-1
sigmadelta<-10
sigmaep<-0.1
```

```
##Conducting Simulation
terrorI1 <-0.05
terrorI2 <-0.05/n
terrorI3 <-0.01/n
ferrorI1 <-0.2
ferrorI2 <-0.5
ferrorI3 <-0.5/n

f1<-qf(1-ferrorI1 , df1=2,df2=n-2)
f2<-qf(1-ferrorI2 , df1=2,df2=n-2)
f3<-qf(1-ferrorI3 , df1=2,df2=n-2)
t1<-qt(1-terrorI1 , df=n-2-1)
t2<-qt(1-terrorI2 , df=n-2-1)
t3<-qt(1-terrorI3 , df=n-2-1)


bho<-matrix(0,N,1)
mseo<-rep(NA, length(beta))
bho1<-matrix(0,N,1)
mseo1<-rep(NA, length(beta))
bho2<-matrix(0,N,1)
mseo2<-rep(NA, length(beta))
bho3<-matrix(0,N,1)
mseo3<-rep(NA, length(beta))
bha<-matrix(0,N,1)
```

```
msea<-rep(NA,length(beta))
bhm<-matrix(0,N,1)
msem<-rep(NA,length(beta))
bhl<-matrix(0,N,1)
msel<-rep(NA,length(beta))
bhh<-matrix(0,N,1)
mseh<-rep(NA,length(beta))
bhb<-matrix(0,N,1)
mseb<-rep(NA,length(beta))

for (j in 1:length(beta)){
        for (i in 1:N){
                para<-c(alpha,beta[j])

                ## case 1. equally spaced x_i
                x<-seq(5/n,5*(1-gamma),by=5/n)
                ep<-rnorm(n1,0,sigmaep)
                xt<-model.matrix(~x)
                y<-xt%*%para+ep
                u<-seq(5*(1-gamma)+5/n,5,by=5/n)
                ut<-model.matrix(~u)
                v<-ut%*%para
                tau<-rnorm(n2,0,sigmatau)
                delta<-rnorm(n2,0,sigmadelta)
                xo<-u
```

```
yo<-v+delta

## case 2. normal distributed x_i
x<-rnorm(n1,0,sigmax)
ep<-rnorm(n1,0,sigmaep)
xt<-model.matrix(~x)
y<-xt%*%para+ep
u<-rnorm(n2,0,sigmax)
ut<-model.matrix(~u)
v<-ut%*%para
tau<-rnorm(n2,0,sigmatau)
delta<-rnorm(n2,0,sigmadelta)
xo<-u+tau
yo<-v+delta

## model
xi<-c(x,xo)
yi<-c(y,yo)
mydata<-data.frame(xi,yi)
model<-lm(yi~xi)

##(1)OLS
require(MASS)
ols<-lm(yi~xi,data=mydata)
bho[i]<-summary(ols)$coef[2,1]
```

```
##(2)Least Squares Estimator After Deletion of
        Influential Outliers:
require(MASS)
Di<-cooks.distance(modeli)
Di<-round(Di,digits=2)
Ri<-studres(modeli)
Ri<-round(Ri,digits=2)
data1<-cbind(mydata,Di,Ri)
data11<-data1[Di<f1,]
data12<-data1[Di<f2,]
data13<-data1[Di<f3,]
data14<-data1[Ri<t1,]
data15<-data1[Ri<t2,]
data16<-data1[Ri<t3,]
        if (sigmax==sigmatau){
                ols1<-lm(yi~xi,data=data14)
                ols2<-lm(yi~xi,data=data15)
                ols3<-lm(yi~xi,data=data16)}
                else{
        ols1<-lm(yi~xi,data=data11)
        ols2<-lm(yi~xi,data=data12)
        ols3<-lm(yi~xi,data=data13)}
bho1[i]<-summary(ols1)$coef[2,1]
bho2[i]<-summary(ols2)$coef[2,1]
```

```
bho3[i]<-summary(ols3)$coef[2,1]

##(3)The Least Absolute Value Estimator
library(quantreg)
lavi<-rq(yi~xi)
bha[i]<-summary(lavi)$coef[2]

##(4)The Least Median of Squares Estimator
library(MASS)
lmsi<-lqs(yi~xi,method="lms")
bhm[i]<-lmsi$coef[2]

#(5)Least Trimmed Squares Estimator
library(MASS)
ltsi<-lqs(yi~xi,method="lts")
bhl[i]<-ltsi$coef[2]

##(6)M-estimator - Huber
mhubi<-rlm(yi~xi)
bhh[i]<-mhubi$coef[2]


##(7)M-estimator - MM
mbsi<-rlm(yi~xi,data=mydata,method="MM")
bhb[i]<-mbsi$coef[2]
```

```
    }

    mseo[j]<-var(bho)+(mean(bho)-beta[j])^2
    mseo1[j]<-var(bho1)+(mean(bho1)-beta[j])^2
    mseo2[j]<-var(bho2)+(mean(bho2)-beta[j])^2
    mseo3[j]<-var(bho3)+(mean(bho3)-beta[j])^2
    msea[j]<-var(bha)+(mean(bha)-beta[j])^2
    msem[j]<-var(bhm)+(mean(bhm)-beta[j])^2
    msel[j]<-var(bhl)+(mean(bhl)-beta[j])^2
    mseh[j]<-var(bhh)+(mean(bhh)-beta[j])^2
    mseb[j]<-var(bhb)+(mean(bhb)-beta[j])^2

}

outcome<-list(beta=beta, betaL=mseo, betaDC1=mseo1, betaDC2=mseo2,
              betaDC3=mseo3, betaA=msea, betaM=msem)
```

Bibliography

Aggarwal, C. C. (2013). Outlier analysis.

Barnett, V. and Lewis, T. (1984). *Outliers in statistical data*. Wiley, Chichester [West Sussex]; New York, 2nd edition.

Charnes, A., Cooper, W. W., and Ferguson, R. O. (1955). Optimal estimation of executive compensation by linear programming: 1. introduction. *Management Science (pre-1986)*, 1(2):138.

Chen, K., Ying, Z., Zhang, H., and Zhao, L. (2008). Analysis of least absolute deviation. *Biometrika*, 95(1):107–122.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall, New York.

Dielman, T. E. (2005). Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, 75(4):263–286.

Donoho, D. (1982). *Breakdown Properties of Multivariate Lacation Estimators*. PhD thesis, Harvard University.

Gillard, J. (2009). An overview of linear structural model in error-in-variables regression.

Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall, New York; London.

Hossain, A. and Naik, D. (1991). A comparative study on detection of influential observations in linear regression. *Statistical Papers*, 32(1):55–69.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35.

Iglewicz, B. and Hoaglin, D. (1993). The asqc basic references in quality control: Statistical techniques. *Volume 16: How to Detect and Handle Outliers*, 16.

Kendall, M. G. (1952). Regression, structure and functional relationship. part ii. *Biometrika*, 39(1/2):96–108.

Kutner, M. H., Nachtsheim, C., and Neter, J. (2004). *Applied Linear Regression Models*. McGraw-Hill/Irwin, New York;Boston;, 4th edition.

Myers, R. H. (1990). *Classical and modern regression with applications*. PWS-KENT, Boston, 2nd edition.

Park, S. K. and Xu, L. (2013). Observation influence diagnostic of a data assimilation system. *Data Assimilation*, 11.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: Computability of squared- error versus absolute-error estimators. *Statistical Science*, 12(4):279–296.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Stigler, S. M. (2010). The changing history of robustness.(author abstract). *The American Statistician*, 64(4):277.

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *Journal of the American Statistical Association*, 88(421):237–244.

Weisberg, S. (1980). *Applied linear regression*. Wiley, New York.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Mathematical Statistics*, 15(2):642–656.