Subsampling Methods for Robust Inference in Regression Models

by

Xiao Ling

B.Sc., Beijing Normal University, 2007

A Dissertation Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

University of Victoria

Subsampling Methods for Robust Inference in Regression Models

by

Xiao Ling

B.Sc., Beijing Normal University, 2007

Supervisory Committee

Dr. Min Tsao, Supervisor

(Department of Mathematics and Statistics)

Dr. Julie Zhou, Supervisor

(Department of Mathematics and Statistics)

Dr. William Reed, Departmental Member

(Department of Mathematics and Statistics)

Supervisory Committee

Dr. Min Tsao, Supervisor
(Department of Mathematics and Statistics)

Dr. Julie Zhou, Supervisor
(Department of Mathematics and Statistics)

Dr. William Reed, Departmental Member
(Department of Mathematics and Statistics)

ABSTRACT

This thesis is a pilot study on the subsampling methods for robust estimation in regression models when there are possible outliers in the data. Two basic proposals of the subsampling method are investigated. The main idea is to identify good data points through fitting the model to randomly chosen subsamples. Subsamples containing no outliers are identified by good fit and they are combined to form a subset of good data points. Once the combined sets containing only good data points are identified, classical estimation methods such as the least-squares method and the maximum likelihood method can be applied to do regression analysis using the combined set. Numerical evidence suggest that the subsampling method is robust against outliers with high breakdown point, and it is competitive to other robust methods in terms of both robustness and efficiency. It has wide application to a variety of regression models including the linear regression models, the non-linear regression models and the generalized linear regression models. Research is ongoing with the

aim of making this method an effective and practical method for robust inference on regression models.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would first like to acknowledge my gratitude to my supervisors, Dr. Julie Zhou and Dr. Min Tsao who have guided me into the field of Robust Statistics. During my study at the University of Victoria, they have provided infinite help and patience. I would also like to express my thanks to my committee members, who have been very helpful in assisting in the completion of the thesis. Finally, I would like to thank my family members for their support.

Chapter 1

Introduction

In engineering, the notion of "robustness" is concerned with the ability to withstand stresses, pressures or changes in procedures and circumstances. A system or a design is said to be "robust" if it can withstand different types of variations in its operating environment with minimal damage, alteration or loss of functionality (Wikipedia, 2009). In Statistics, the notion of "robustness" has the same meaning as that in engineering, but instead of some engineering systems or designs, the subjects of interests are now statistical methods or designs.

Most classical statistical methods were developed without much concern about their robustness. The maximum likelihood method, for example, is based on the assumption that the data come from a known parametric model. When the assumption is true, the maximum likelihood estimator (MLE) is the most efficient. Nevertheless, a small departure from the model assumption may render the MLE ineffective. One such a departure occurs when the data set contains a small fraction of outliers, which are not observations from the assumed model. Huber (1964) introduced the robust M-estimator which generalizes the MLE. The M-estimator enjoys the high efficiency of the MLE yet is not heavily influenced by small departures from the model assump-

tion. The importance of robust statistics has been widely recognized and it has been one of the most active research areas in Statistics in recent decades. Indeed, there are now robust versions of most classical (non-robust) statistical methods. Furthermore, robust statistical methods have also becoming increasingly popular with practitioners.

This thesis is concerned with robust estimation of regression models for situations where the data set contains outliers. There are many methods in the literature on robust estimation of location and scale. See, for example, Huber (1981), Hampel, Ronchetti , Rousseeuw and Stahel (1986) and, more recently Maronna, Martin and Yohai (2006). However, there are fewer methods on robust estimation in regression models. Further, existing robust methods for regression models are mostly concerned with simple models such as the linear regression models, *e.g.,* Chapters 4 and 5 in Maronna, Martin and Yohai (2006). They are also typically model-specific in that a robust method for one regression model is designed specifically for that regression model. As such it may not work for other regression models such as non-linear regression models and generalized linear regression models. In this thesis, we propose a simple idea for robust estimation of regression models when a majority of the sample data are "good data" from the underlying regression model. The basic idea is to consider subsamples from the sample and to identify among possibly many subsamples ones that contain only good data (good subsamples). Then estimate the regression model using only the good subsamples through some simple method. The identification of good subsamples is accomplished through fitting the model to the subsamples and then using a criterion, typically a goodness-of-fit measure which is sensitive to the presence of outliers, to determine whether the subsamples contain outliers. We will discuss two different implementations of this idea, which we will refer to as subsampling methods. The main advantages of these methods are that (1) they are applicable in principle to any regression models, (2) under certain conditions

and conditional on the successful removal of outliers through the subsampling process, they provide unbiased estimators for the regression model parameters, and (3) they are conceptually easy to implement. The main disadvantages of these methods are that there is a small probability each time the methods are not robust and that they may be computationally intensive.

The rest of this chapter is organized as follows. In Section 1.1, we introduce the basic idea of robustness through two real life data sets which contain outliers. We examine the non-robust nature of the common estimates of location and scale. We also discuss the identification of the outliers and give examples of robust estimates of location and scale. In Section 1.2, we define the general regression model, and consider the method of least-squares and the robust MM method (Yohai, 1987). We illustrate through a real life example involving outliers that the method of least-squares is not robust and that the MM method is superior for such a case. We also briefly discuss the subsampling methods which are seen as complementary to the MM method. We conclude with an overview of the thesis in Section 1.3.

1.1 Outliers and robustness

Consider the following sample of 24 measurements of copper contents in wholemeal flour (in parts per million) from Analytical Methods Committee (1989). This data set was also analyzed by Maronna, Martin and Yohai (2006, p2).

Example 1.1. *Copper content in wholemeal flour in parts per million (Analytical Methods Committee, 1989).*

2.2	*2.2*	*2.4*	*2.4*	*2.5*	*2.7*	*2.8*	*2.9*
3.03	*3.03*	*3.1*	*3.37*	*3.4*	*3.4*	*3.4*	*3.5*
3.6	*3.7*	*3.7*	*3.7*	*3.7*	*3.77*	*5.28*	*28.95*

The value 28.95 stands out from the rest of the values. It is an outlier in the sense that it is located far away from the bulk of the data. This is clearly seen from the Boxplot in Figure 1.1 where the value 28.95 is all by itself at the upper extreme. This outlier may have been just the consequence of incorrect recording of data or it may have been the result of some unusual contamination to the sample. In any event, it is highly influential to the usual estimates of location and scale, the sample mean and the sample standard deviation. The sample mean $\bar{x}$ and the sample standard deviation (SD) s are 4.28 and 5.30, respectively. Noting that $\bar{x} = 4.28$ is larger than all but two of the data values, it is not representative of a typical value in the data set and is thus a poor measure of central location for the data set. If the outlier 28.95 is removed, then $\bar{x} = 3.21$ which is a much more reasonable measure of central location. The new SD is now $s = 0.69$, which is only one seventh of the original. □

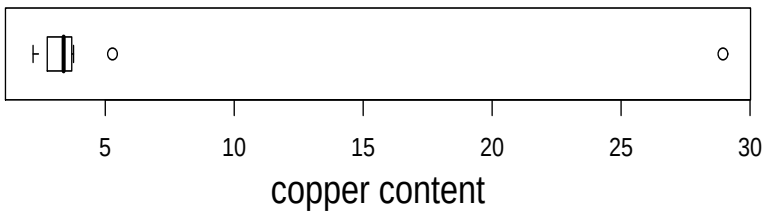

Figure 1.1: Boxplot for Example 1.1.

This is but one example where one single outlier has a big impact on the sample mean and sample SD. In general, we are interested in the behavior of an estimator when a single outlier approaches infinity. In the cases of sample mean and sample standard deviation, suppose that the value 28.95 is replaced by some arbitrary value x and let $|x|$ go to infinity, then $|\bar{x}|$ and the sample SD s will both go to infinity. Because of this, we say that a single outlier has *unbounded influence* on the sample mean and

sample SD. In Chapter 2, we will discuss the concept of the influence function of a statistic to further expand this observation.

It is clear that we need robust inference which either removes or diminishes the influence of the outliers. There are two basic approaches to robust inference (Maronna, Martin and Yohai, 2006): [a] deleting outliers from the data and then using classical methods to make inference and [b] using robust methods with built-in ability to handle outliers to do the inference. Such robust methods typically impose a limit on the amount of influence an outlier can have. Deleting outliers may seem simple, but it poses a couple of problems:

1. The identification of outliers is challenging, especially for high dimensional data.
2. There is always a risk of deleting "good" observations, which may lead to underestimation of the variability.

A traditional measure of the "outlyingness" of an observation x_i in a sample $x_1, x_2, \ldots, x_n$ is the "three-sigma edit" (Pukelsheim, 1994), which is the ratio between its distance to the sample mean and the sample SD:

$$t_i = \frac{x_i - \bar{x}}{s} \tag{1.1}$$

Observations with $|t_i| > 3$ are traditionally deemed as suspicious, based on the fact that they would be "very unlikely" under normality. However, this rule has some drawbacks (Maronna, Martin and Yohai, 2006). In a large normal sample, about three observations out of 1000 will have $|t_i| > 3$. In very small samples with size n, the rule is ineffective. It can be shown that $t_i < \frac{n-1}{\sqrt{n}}$ for all possible data values. Hence if $n \leq 10$, then $|t_i| < 3$ always holds. When there are several outliers, their effects may interact in such a way that some or all of them will not be identified by the "three-sigma edit" (a phenomenon called masking), as we now demonstrate in

the following example.

Example 1.2. *The following data (Stigler, 1977) contains 20 determinations of the time (in microseconds) needed for light to travel a distance of* 7442 *meters. The actual times are the table values*$\times 0.001 + 24.8$.

28	*26*	*33*	*24*	*34*	*-44*	*27*	*16*	*40*	*-2*
29	*22*	*24*	*21*	*25*	*25*	*23*	*29*	*31*	*19*

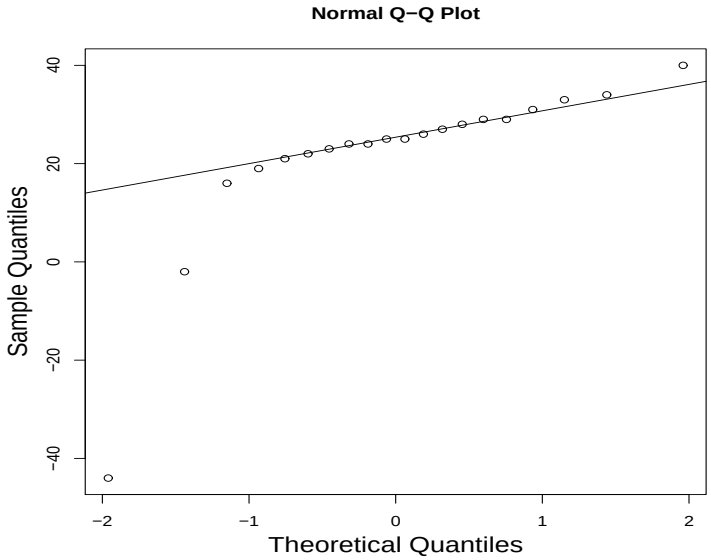

Figure 1.2: Q-Q plot of observed times in Example 1.2.

The normal Q-Q plot in Figure 1.2 reveals that the two lowest observations -44 and -2 as suspicious. The corresponding "three-sigma edit" values $|t_i|$ are -3.73 and -1.35, respectively. Hence -2 would not be identified as an outlier due to its small $|t_i|$ value. However, the reason that -2 has such a small $|t_i|$ value is that the two lowest observations -44 and -2, in particular -44, pulled $\bar{x}$ to the left and substantially inflated s. This led to the small $|t_i|$ for -2 and created a masking effect for -2.

It is clear that we need robust estimators of location and scale for data sets containing outliers, such as those in the above examples. One robust estimate for

location is the sample median. It is a good robust alternative to the sample mean. If the sample is symmetrically distributed around its center, the mean and the median are approximately equal. But if there is an outlier in the sample, like in Example 1.1, the median is a more reliable estimate of location than the mean. Likewise, the median absolute deviation about the median (MAD) (Rousseeuw and Croux, 1993) is a robust alternative to the SD. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$,

$$\text{MAD}(\mathbf{x}) = \text{MAD}(x_1, x_2, \ldots, x_n) = \text{Median}\{|x_i - \text{Median}(\mathbf{x})|\}.$$

This estimator uses the sample median twice. To make the MAD easy to interpret under normal distribution assumption, we define the normalized MAD (MADN) as

$$\text{MADN}(\mathbf{x}) = \frac{\text{MAD}(\mathbf{x})}{0.6745},$$

so that the expected value of MADN equals the standard deviation of the underlying normal distribution. Based on these two robust estimators of location and scale, we can revise the classical "three sigma edit" rule in (1.1):

$$t_i' = \frac{x_i - \text{Median}(\mathbf{x})}{\text{MADN}(\mathbf{x})}.$$

We have seen in this section the impact of outliers on classical location and scale estimates, and discussed robust alternatives to these none robust estimates. In the next section, we turn to regression models and their related robustness issues, which are the main focus of this thesis.

1.2 Regression models and robust estimation

Regression models which characterize the relationships between explanatory variables $\mathbf{x} \in R^q$ and a response variable $y \in R^1$ are among the most important models in Statistics. A parametric regression model typically has the form

$$y = g(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon, \tag{1.2}$$

where $g(\mathbf{x}; \boldsymbol{\beta})$ is a known regression function with parameter vector $\boldsymbol{\beta} \in R^p$, which usually needs to be estimated from the data. The response variable y is sometimes referred to as the dependent variable, and the explanatory variables $\mathbf{x}$ is also referred to as the independent variables. The error term ε is a random variable with mean $E(\varepsilon) = 0$ and variance $Var(\varepsilon) = \sigma^2$. It represents variations in the dependent variable y that are not accounted for by the regression function $g(\mathbf{x}; \boldsymbol{\beta})$. Suppose we have a sample of n (pairs of) observations $(\mathbf{x}_i, y_i)$, we can estimate the unknown parameter $\boldsymbol{\beta}$ using the method of least-squares. This results in an estimated model which minimizes the sums of squared errors. There are other less used methods of estimation such as the method of least absolute deviation which will lead to different estimated models. See Dodge (1987), Berk (2004) and Freedman (2005) for this and other methods of estimation. The least-squares method has a long history going back to the eighteenth century (Wolberg, 2005). When the errors ε_i are uncorrelated and have mean zero and a constant variance, Gauss showed that the least-squares estimates for the linear model where $g(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$ are the best linear unbiased estimators (BLUE) of $\boldsymbol{\beta}$. This result is known as the Gauss-Markov theorem (Hocking, 1985).

The method of least-squares is by far the most commonly used method of estimation due to its optimal BLUE property and its computational simplicity. The latter was essential before the cheap computing power became widely available in recent

decades. Nevertheless, the method of least-squares is not robust. One single outlier can severely affect the accuracy of the least-squares estimates. The following example further illustrates this point.

Example 1.3. *The data set in Table 1.1 is from Bond's (1979) experiment on the speed of learning of rats. Each rat was made to go through a shuttle box in successive attempts until a certain number of successes were reached. At a given attempt, the rat received an electric shock if it was not successful in the first 5 seconds. At the conclusion of the experiment, the number of shocks received by each rat, x, and the average time for all its attempts, y, were recorded.*

Table 1.1: Rat's speed of learning measured in seconds, Bond (1979)

Shocks (x)	Time (y)	Shocks (x)	Time (y)
0	11.4	8	5.7
1	11.9	9	4.4
2	7.1	10	4
3	14.2	11	2.8
4	5.9	12	2.6
5	6.1	13	2.4
6	5.4	14	5.2
7	3.1	15	2

The electric shock serves as a stimulus for learning and hence the number of shocks x is regarded as an explanatory variable, and the time y is the response. We use the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 16, \tag{1.3}$$

to fit this data set. However, the least-squares fitted line (LS, in black) in Figure 1.3(a) does not fit the data set well. It has been pulled up at the left end by the three

observations from rats 1, 2 and 4. Figure 1.3(a) also shows the least-squares fitted line computed without using the three points (LS-, in red). It does a better job at capturing the linear trend of the majority of the data points. A robust estimator, the MM-estimator, was also used to compute a fitted line (MM, in blue). This robust method was applied directly to the data set without having to identify the three outliers in the sample and the resulting MM line does an equally good job at capturing the linear trend as the red line. This is an important advantage as the outliers are not always easy to tell. More details about the MM-estimator will be given in Chapter 2.

Table 1.2: Estimated parameters for the linear model (1.3) for rats data with standard error (s.e.) in the brackets.

Estimation methods	Intercept ($\hat{\beta}_0$) s.e.($\hat{\beta}_0$)	Slope ($\hat{\beta}_1$) s.e.($\hat{\beta}_1$)
LS	10.48 (1.08)	-0.61 (0.12)
LS-	7.22 (0.76)	-0.32 (0.08)
MM	7.83 (1.25)	-0.41 (0.11)

Table 1.2 gives the estimated values for β_0 and β_1 in (1.3) given by the method of least-squares with (LS) and without (LS-) the three outliers. The estimates given by the MM-estimator are also included. From the table, one can see that the MM-estimates are quite different from the least-squares estimates (LS), but they are very similar to the least-squares estimates without the three outliers (LS-). Hence the MM-estimates are little affected by the outliers. This also agrees with our previous observation on Figure 1.3(a) that LS- line is very close to the MM line, and both of them fit the bulk of the data.

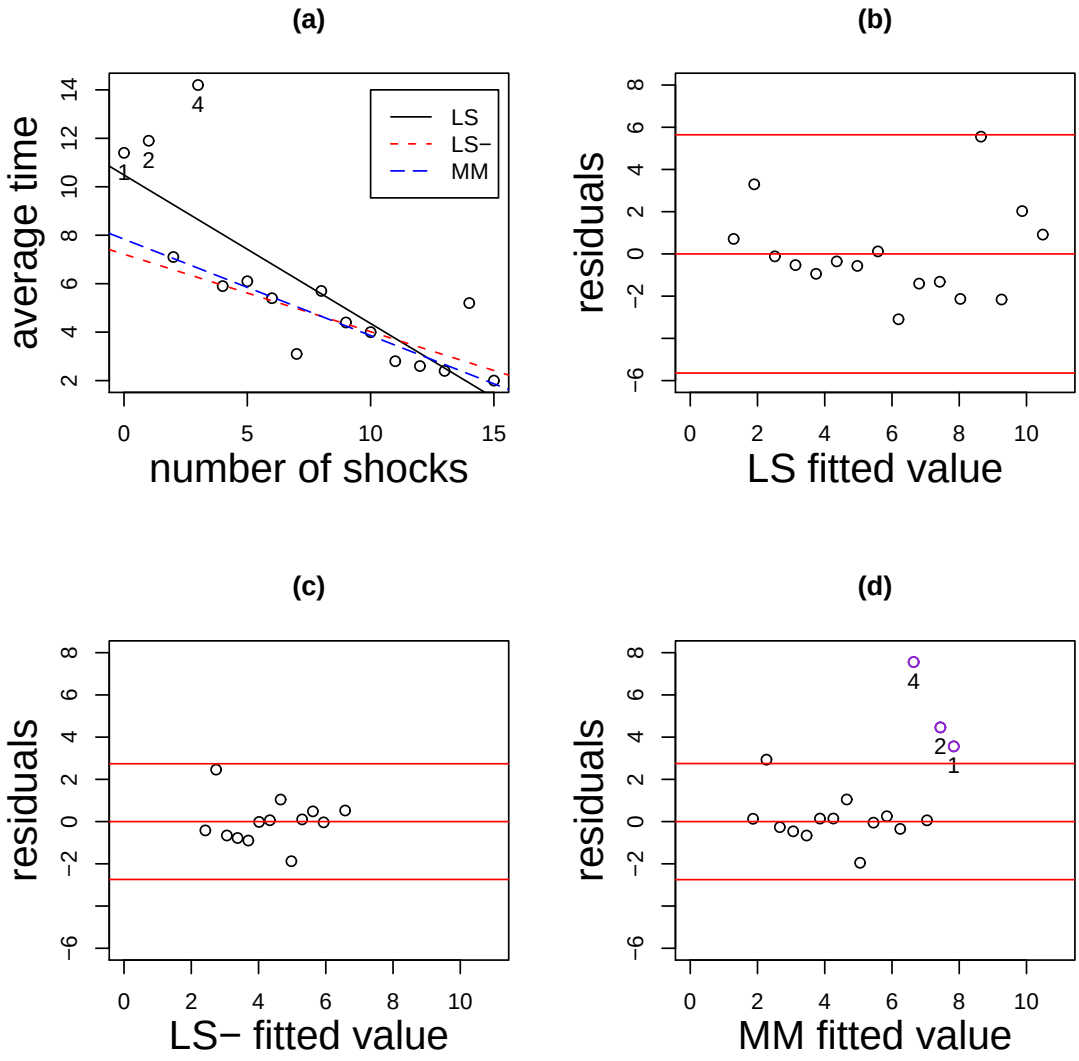

Figure 1.3: (a) LS is the fitted least-squares line to all data points; LS- is the fitted least-squares line after omitting points 1, 2 and 4; MM is the line fitted using the robust MM-estimator. (b) Residual plot for the LS fitted line; the red lines are located at $\pm 2.5 \times \hat{\sigma}$ and zero. (c) Residual plot for LS- fitted line; the red lines are located at $\pm 2.5 \times \hat{\sigma}$ and zero. (d) Residual plot for the MM fitted line; the red lines are located at $\pm 2.5 \times \hat{\sigma}$ and zero; the purple points are identified outliers.

Figure 1.3 (b), (c) and (d) present the residual plots: (b) shows the LS residuals vs the fitted value; (c) gives the LS- residuals vs the fitted values; (d) has the MM residuals vs the fitted values. The red lines are drawn at $\pm 2.5\hat{\sigma}$ where $\hat{\sigma}$ is the estimated value for σ, the standard deviation of ε. The points outside the red lines are identified as outliers. The range between two red lines in the three plots are different because the estimates for σ are different for the three methods. The LS method failed to identify the outliers in the data, but the MM-estimator correctly identified the three outliers colored purple in Figure 1.3 (d).

The above example demonstrates the need for robust methods in regression analysis as the impact of outliers on parameter estimation by non-robust methods is rather damaging. Robust methods such as the MM method are appealing due to its ability to handle outliers automatically. Nevertheless, the MM method can be only applied to the linear regression models. The MM-estimator has no analytic expression and MM-estimates must be computed numerically. Robust M-estimation of other regression models are also complicated and for some case are not well developed. In this thesis, we propose two alternative methods for robust estimation of regression models, the subsampling methods, which complement the M-estimation based robust methods in general and the MM method for linear models in particular. As we have described earlier, the robustness of such methods comes from the elimination of outliers (instead of limiting their impact) in the subsampling step where the model is fitted to the subsamples to determine the presence of outliers. Once a subsample containing only "good data" is identified, simple methods such as the least-squares methods are used for the final parameter estimation. As such, the subsampling estimators may have simple analytic expressions in terms of the "good data" and subsampling methods amount to a way of extending the use of simple non-robust methods to situations with outliers. The general idea of subsampling method is also independent of the

underlying regression model.

It is of interest to note the connection between our subsampling methods, the bootstrap method and the method of trimming outliers. With the subsampling methods, we essentially substitute analytical treatment of the outliers (such as the use of the ρ functions in the MM-estimator) with computing power through the elimination of outliers by repeated fitting of the model to the subsamples. From this point of view, our subsampling methods share the same spirit of trading analytic treatment for intensive computing with bootstrapping. But our subsampling method is not bootstrapping as our objective is to identify a single good subsample instead of making inference based on all bootstrap samples. The subsampling based elimination of outliers is also a generalization of the method of trimming outliers. Instead of relying on some measure of outlyingness of the data to decide which data points to trim, the subsampling methods use a model based trimming by identifying subsamples containing outliers through fitting the model to the subsample. The outliers are in this case outliers with respect to the underlying regression model instead of some measure of central location and they are only implicitly identified.

1.3 Overview of the thesis

In Chapter 2, we review important concepts and tools in robust statistics. We first revisit the problem of robust estimation of location and scale, and introduce the robust M-estimator. We also discuss two important robustness measures for estimators and apply these to several estimators to evaluate their robustness. We then review the robust MM-estimators for regression models.

Chapters 3 and 4 cover our proposed subsampling methods. In Chapter 3, we give the basic setup of the problem and provide a subsampling algorithm which identifies

a pre-specified number of good subsamples and then takes the union of these to form a combined (good) subsample for final model fitting. This is one implementation of the subsampling idea and we will refer to the resulting subsampling method as the *subsampling method proposal I* or SM1. We study the theoretical aspects of this proposal, including its efficiency in terms of the usage of the good data points in the sample. We will also investigate the asymptotic behavior of proposal I estimators and demonstrate the effectiveness of proposal I through numerical examples involving linear regression models and a logistic regression model. Comparisons to the least-squares methods and the MM method are also included in Chapter 3. In Chapter 4, we discuss an alternative implementation of the subsampling idea where we only need to identify one good subsample. We then expand this good subsample by testing points not in this subsample and including those points that are not outliers. We refer to this as *subsampling method proposal II* or SM2. We also examine the differences between the two proposals and compare them using several real examples and a simulation study.

Finally, in Chapter 5 we provide some general discussions on the advantages and disadvantages of our proposed methods. We also briefly discuss ongoing research by my supervisors Dr. Min Tsao and Dr. Julie Zhou on the subsampling methods. This thesis is a pilot project for their ongoing research providing important numerical evidences supporting their further development of the subsampling methods.

Chapter 2

Review of Robust Methods

To prepare for the discussion of our subsampling methods for regression models, we review important robustness concepts and methods in this chapter. In Section 2.1, we discuss robust M-estimators for location and scale. In Section 2.2, we introduce two robustness measures which are then used to quantify the robustness of the M-estimators. In Section 2.3, we first review the method of least-squares and the maximum likelihood method for regression models. We then describe in more details the robust MM method which we have used in Example 1.3.

2.1 Robust estimation of location and scale

A location model is given by

$$x_i = \mu + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where x_i are observed data, μ is the location parameter and ε_i are independent random errors with distribution function $F_0(x)$ and mean 0. Under this location model, $x_1, x_2, \ldots, x_n$ are independently and identically distributed (i.i.d.) random

variables with distribution function

$$F(x) = F_0(x - \mu).$$

An M-estimator $\hat{\mu}$ of location corresponding to a ρ function is defined as

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu), \tag{2.1}$$

which is often computed by solving

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0, \tag{2.2}$$

where $\psi = d\rho/dx$. Typically, function ρ is chosen in order to ensure [1] the estimates are "nearly optimal" when F_0 is exactly normal and [2] the estimates are "nearly optimal" when F_0 is approximately normal. Such ρ functions must satisfy the following properties (Maronna, Martin and Yohai, 2006):

(C1) $\rho(x)$ is a non-decreasing function of $|x|$;

(C2) $\rho(0) = 0$;

(C3) $\rho(x)$ is increasing for $x > 0$ such that $\rho(x) < \rho(\infty)$;

(C4) if ρ is bounded, it is also assumed that $\rho(\infty) = 1$.

Function ψ satisfies that ψ is odd and $\psi(x) \geq 0$ for $x \geq 0$. If $\rho(x)$ is chosen to be $-\log f_0(x)$ where f_0 is the density function corresponding to F_0, then the M-estimator is the maximum likelihood estimator (MLE). Note that if f_0 is symmetric, then ρ is even. Hence ψ is odd and $\psi(x) \geq 0$ for $x \geq 0$.

The sample mean and median are special cases of M-estimates. To see this, if F_0 is a standard normal distribution, by choosing $\rho(x) = -\log f_0(x)$, we have

$\rho(x) = x^2/2 + c$ and $\psi(x) = x$, where c is a constant. By (2.2), we obtain $\hat{\mu} = \bar{x}$. Hence the sample mean is the M-estimate for the mean of a normal random variable x under this particular choice of ρ function. If, on the other hand, F_0 is the double exponential distribution, then $f_0(x) = \frac{1}{2}e^{-|x|}$ and $\rho(x) = \log 2 + |x|$. It follows from (2.2) that

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} |x_i - \mu|,$$

which implies $\hat{\mu}$ is the sample median of x_i. Since $-\log f_0(x)$ is only one of many possible choices for the ρ function, the M-estimator is thus a generalization of the MLE. Most MLEs such as the sample mean are not robust. For robust M-estimation, we are interested in other choices of ρ function.

Two commonly used ρ functions which give rise to robust M-estimators are the Huber function and the bisquare function. The Huber function and its derivative are

$$\begin{aligned} \rho_h(x) &= \begin{cases} x^2, & \text{if } |x| \le k \\ 2k|x| - k^2, & \text{if } |x| > k; \end{cases} \\ \psi_h(x) &= \begin{cases} x, & |x| \le k \\ sgn(x)k, & |x| > k, \end{cases} \end{aligned} \tag{2.3}$$

where $k > 0$ is a constant (Maronna, Martin and Yohai, 2006). The value of k is often chosen to satisfy an asymptotic efficiency requirement when the underlying distribution is normal. The bisquare ρ-function is

$$\rho_b(x) = \begin{cases} 1 - [1 - (x/k)^2]^3, & \text{if } |x| \le k \\ 1, & \text{if } |x| > k; \end{cases} \tag{2.4}$$

with derivative $d\rho_b/dx = 6\psi_b(x)/k^2$ where $k > 0$ is a constant and

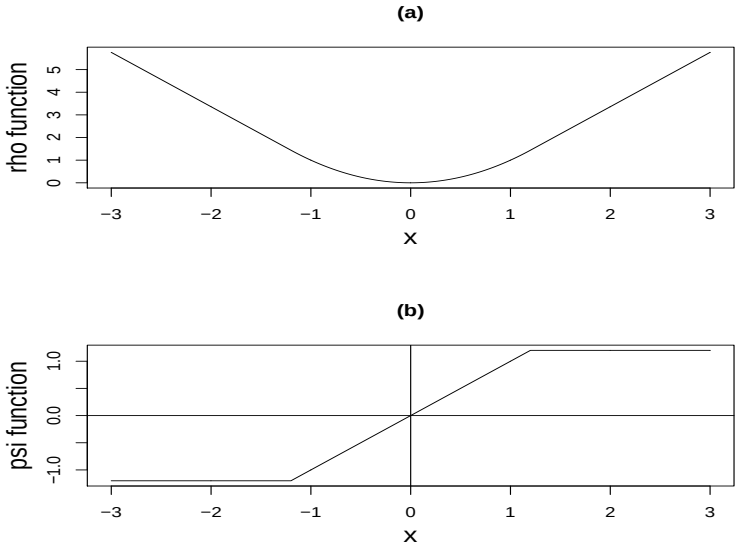

Figure 2.1: Huber function and its derivative with $k = 1.2$. (a) Huber function ρ, (b) the derivative ψ.

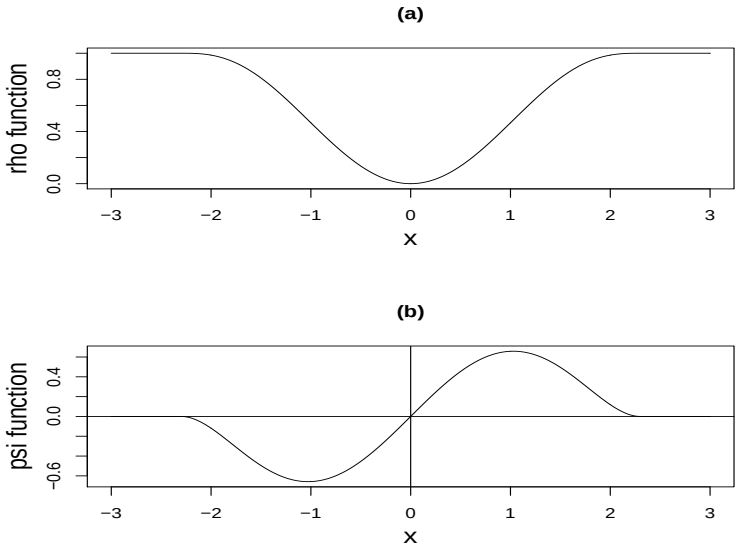

Figure 2.2: Bisquare function and its derivative with $k = 2.3$. (a) bisquare function ρ, (b) the derivative ψ.

$$\psi_b(x) = x\left[1 - \left(\frac{x}{k}\right)^2\right]^2 I(|x| \leq k). \tag{2.5}$$

These functions are plotted as Figures 2.1 and 2.2. The Huber function ρ_h is quadratic in a central region but increases linearly to infinity. The corresponding ψ_h function is monotonic. On the other hand, the bisquare function ρ_b is bounded in $(0, 1)$. Furthermore, its corresponding ψ_b is "redescending", meaning that $\psi_b(x)$ tends to zero when $x \to \infty$. They are different types of ρ functions which lead to different M-estimators. If ρ is everywhere differentiable and ψ is monotonic, then the solutions of (2.1) and (2.2) are equivalent and this common solution is referred to as the "monotone M-estimate". However, if ψ is redescending, then it is not monotonic and there may be multiple solutions to (2.2). In this case, only one of these solutions is the M-estimate that satisfies (2.1). Such an M-estimate is called a "redescending M-estimate".

The location is an important aspect of a random variable that we are interested in. The scale is an equally important aspect that we want to make inference about. It is not only of independent interest itself but also often tied to the estimation of the location through joint estimating equations. For brevity, we only review a simple case of M-estimation for the scale alone here. Suppose $x_i = \sigma u_i$, where the u_i are i.i.d. with density f_0 and $\sigma > 0$. Then the density of x_i is

$$\frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right),$$

where the parameter σ is the scale parameter of x_i. For a chosen ρ function, an M-estimator of scale σ is the solution to the following equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i}{\hat{\sigma}}\right) = \delta, \tag{2.6}$$

where $\delta \in (0, \rho(\infty))$ is a constant and function ρ must also satisfy the conditions $\mathbf{C1} - \mathbf{C4}$. If function ρ is bounded, according to $\mathbf{C4}$, we have $\delta \in (0, 1)$.

2.2 Robustness measures

In the previous section, we have introduced the M-estimators of location and scale but have yet to look into the robustness of these M-estimators. In the present section, we discuss how to measure the robustness of an estimator using two important tools: the influence function (IF) and the breakdown point (BP). A more detailed discussion may be found in Maronna, Martin and Yohai (2006).

Influence functions provide an attractive means by which one can formally assess the robustness of a statistical procedure and identify more robust alternatives. The empirical influence function of an estimator refers to the value of the estimator as a function of a single data point, and it tells us the behavior of the estimator when we change one data point in the sample (Jolliffe, 1986). To define the empirical influence function, suppose $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is an estimator based on a random sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ of size n. The empirical influence function is given by

$$IF_{\hat{\theta}}(x) = \hat{\theta}(x_1, x_2, \ldots, x_{i-1}, x, x_{i+1}, x_n), \tag{2.7}$$

where we have replaced the ith value in the sample by an arbitrary value x. By setting x to extremely large or small values, we can observe the changes in the estimate $IF_{\hat{\theta}}(x)$ in response to such extreme values, and hence the robustness or the lack of it of the estimator $\hat{\theta}(\mathbf{x})$. If $IF_{\hat{\theta}}(x)$ is unbounded, then the estimator $\hat{\theta}(\mathbf{x})$ is not robust. In this case, a single outlier can have infinite influence on the estimator. If $IF_{\hat{\theta}}(x)$ is bounded, then estimator is robust in the sense that any point, even an extreme outlier, will have only limited influence on the estimator.

The influence function captures the robustness of an estimator against one single outlier. It does not tell us the robustness of an estimator where there are more than one outliers present. Although one may generalize the influence function to handle multiple-outlier situations by replacing $k < n$ sample points with arbitrary values, this would result in a multivariate influence function which is difficult to work with. For this reason, we study instead the breakdown point of an estimator, which is a one-number measure of robustness against multiple outliers.

The breakdown point of an estimator is the proportion of outliers or arbitrarily large observations that an estimator can handle before becoming essentially unbounded. Alternatively, it is the largest amount of contamination that the data set may contain such that the estimator still gives some information about true parameter value. The finite sample breakdown point of an estimator is the fraction of data that can be given arbitrary values without making the estimator arbitrarily bad (Donoho and Huber, 1983). More formally, denote by χ_m the set of all data sets $\mathbf{y}$ of size n having $n - m$ elements in common with $\mathbf{x}$:

$$\chi_m = \{\mathbf{y} : \#(\mathbf{y}) = n, \#(\mathbf{x} \cap \mathbf{y}) = n - m\}.$$

Then the finite sample breakdown point (FBP) ε_n^* is

$$\varepsilon_n^*(\hat{\theta}) = \frac{m^*}{n}, \tag{2.8}$$

where $m^* = \max\{m \geq 0 : \hat{\theta}(\mathbf{y})$ is bounded and is also bounded away from the boundary of the space of $\theta, \forall \mathbf{y} \in \chi_m\}$. Clearly, if $\varepsilon_n^*(\hat{\theta}) = 0$, then the estimator is not robust. A robust estimator $\hat{\theta}$ must have $\varepsilon_n^*(\hat{\theta}) > 0$ and the higher the breakdown point the more robust the estimator. Here high breakdown means a value at or slightly smaller than 0.5 as the notion of breakdown under contaminations exceeding

50% of the data set is not well defined. By the definition of the breakdown point, an estimator $\hat{\theta}$ can handle $m^* = n\varepsilon_n^*(\hat{\theta})$ outliers without breaking down.

The following example illustrates the influence function and breakdown point for various estimators.

Example 2.1. *Copper content data set ($n = 24$) from Example 1.1 revisited: Influence functions and finite sample breakdown points for the sample mean, sample median, sample SD, MAD and the M-estimators for location and scale.*

For the 24 copper content data, we plotted the influence functions of the six estimators by changing observation 24. The plots are shown in Figure 2.3. From Figure 2.3(a), we see that the influence function of the sample mean is unbounded. Thus the sample mean is not robust. From Figure 2.3(b), we see that the influence function of the sample standard deviation is also unbounded, and hence it is also not robust. The influence functions of the sample median, MAD and the M-estimators for location and scale are bounded; hence the corresponding estimators are all robust. Since the influence function can reveal the impact of only one outlier, we need to look at the finite sample breakdown points for these estimators to assess their robustness in handling situations with multiple outliers. The finite sample breakdown points of the sample mean and standard deviation are both zero because their influence functions are unbounded. The breakdown points for the robust estimators of location and scale are shown in Figure 2.4. The plot for the median in Figure 2.4(a), for example, was generated as follows. For $m = 1, 2, \ldots, 12$, we replaced the largest m values in the copper content data set with $50, 50 + 1, \ldots, (50 + m - 1)$, which are very large values (outliers) in comparison to a typical value in the data set. We then computed the median for this new data set of n observations. The solid line in 2.4(a) is that of the median verses the value m for $m = 0, 1, 2, \ldots, 12$. The breakdown of the sample median occurred at $m = 12$ where the plot takes a steep rise. Hence the

breakdown point of the sample median is $\varepsilon_n^* = 11/24$. Similarly, the plot for the M-estimator of location indicates that its breakdown point is also 11/24. For this plot, the M-estimator used is based on the Huber ρ function. Plots for MAD and the M-estimator for scale in 2.4(b) were generated analogously. They indicate the breakdown points for these two scale estimators are also 11/24. □

2.3 Robust estimation of regression models

In the last section, we reviewed an important robust estimator, the M-estimator for location and scale. We also discussed how to measure the robustness of estimators. We now return to the main topic of this thesis, robust estimation of regression models. We first briefly review the method of least-squares and illustrate its lack of robustness with an example. We then discuss the robust MM method for linear models, which is related to the M-estimate for location and scale.

Recall the general regression model (1.2) introduced in Section 1.2. For a given sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, it can be written as

$$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.9}$$

where function g is a given regression function with unknown parameter (vector) $\boldsymbol{\beta}$ and ε_i are independent random errors with mean zero and constant variance σ^2. When $g(\mathbf{x}_i, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, model (2.9) is called a linear regression model. For example, (1.3) is the simple linear regression model.

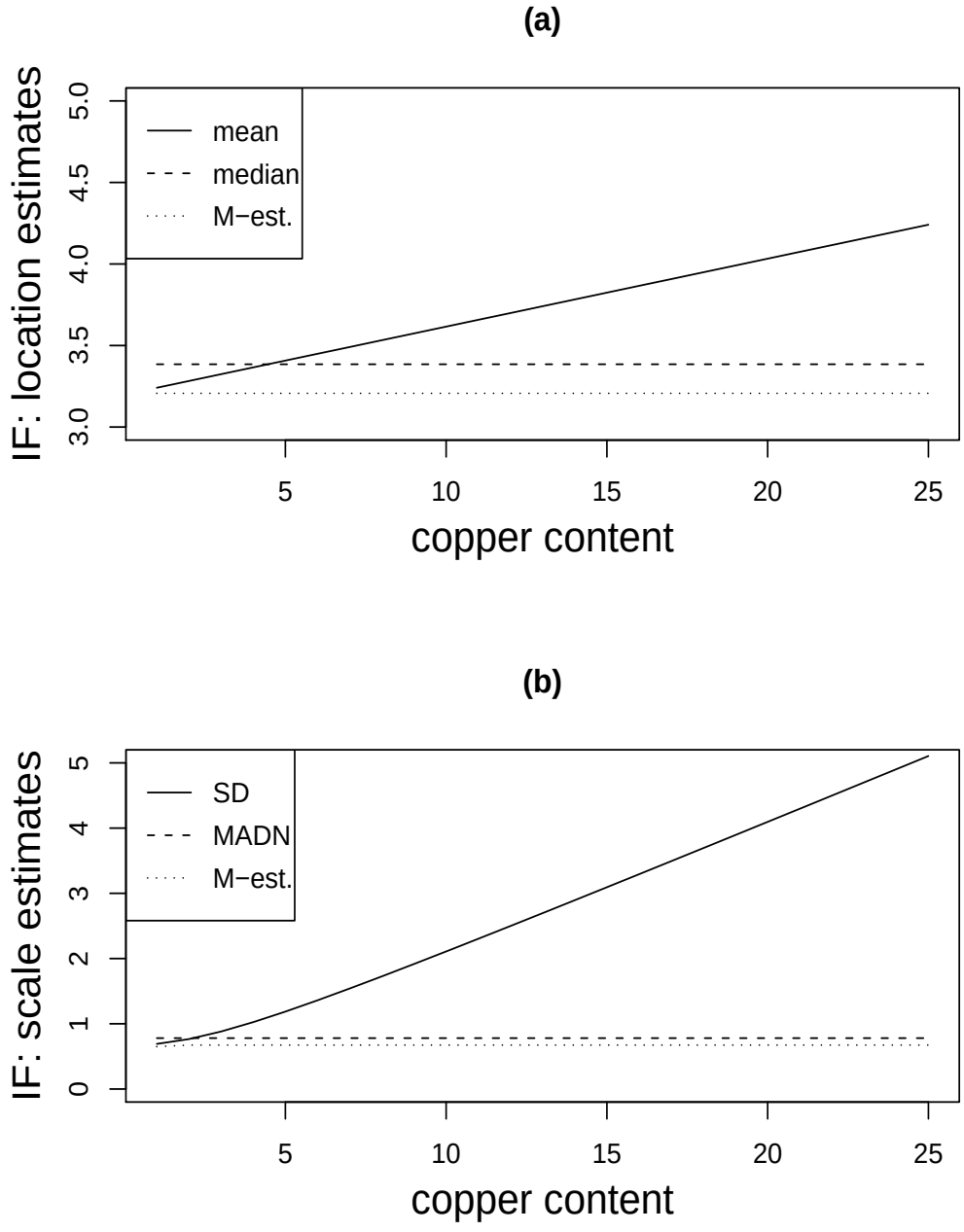

Figure 2.3: Empirical influence function for location and scale estimators: (a) location estimators; (b) scale estimators.

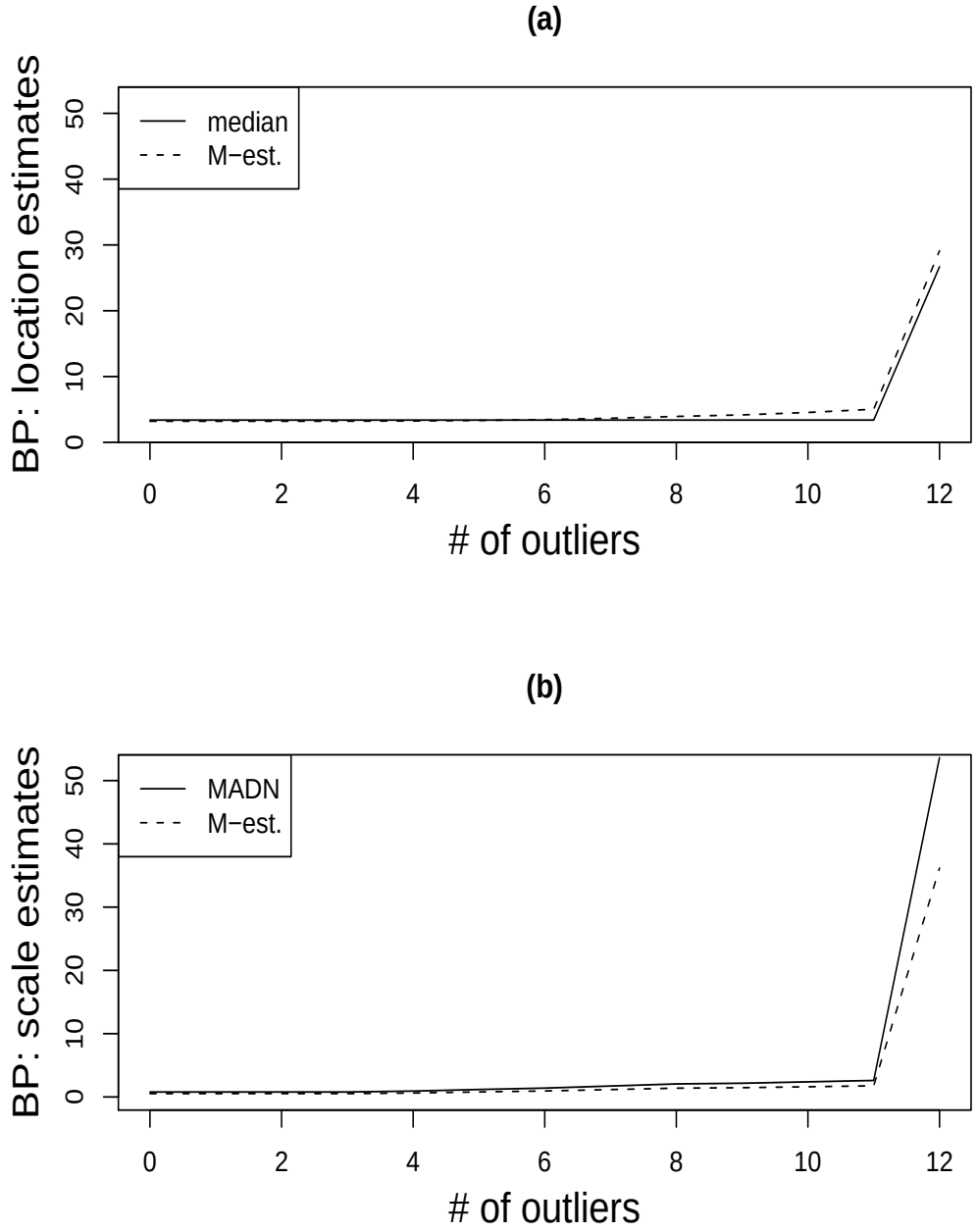

Figure 2.4: Copper content data: plots for identifying the finite sample breakdown points for location and scale estimators: (a) plots for location estimators; (b) plots for scale estimators. Each curve is formed by plotting the value of an estimator against the number of artificial outliers in the copper content sample.

2.3.1 The method of least-squares

The least-squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ is the minimizer of the sums of squares (of residuals) in the right-hand side of the following equation:

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2.$$

For the linear model where $g(\mathbf{x}_i, \boldsymbol{\beta})) = \mathbf{x}_i^T \boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{LS}$ satisfies

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{LS}) \mathbf{x}_i = 0,$$

which leads to the following expression for $\hat{\boldsymbol{\beta}}_{LS}$,

$$\hat{\boldsymbol{\beta}}_{LS} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i.$$

Under standard conditions on the random error ε_i, the least-squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ is the best linear unbiased estimator for $\boldsymbol{\beta}$. Furthermore, if the errors are normally distributed, the least-squares estimator is also the maximum likelihood estimator.

However, the least-squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ is not robust against outliers in the sample. The asymptotic influence function for $\hat{\boldsymbol{\beta}}_{LS}$ is

$$IF_{\hat{\boldsymbol{\beta}}_{LS}}((\mathbf{x}_0, y_0), F) = (E_F(\mathbf{x}\mathbf{x}^T))^{-1} (y_0 - \mathbf{x}_0^T \hat{\boldsymbol{\beta}}_n) \mathbf{x}_0,$$

where F is the joint distribution of $(\mathbf{x}, y)$, $\hat{\boldsymbol{\beta}}_n = \lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{LS}$ and $(\mathbf{x}_0, y_0)$ is an arbitrary point of the data set. $IF_{\hat{\boldsymbol{\beta}}_{LS}}((\mathbf{x}_0, y_0), F)$ is not bounded in either $\mathbf{x}_0$ or y_0. Hence one single outlier in either y or $\mathbf{x}$ can have large influence on the least-squares estimator. Consequently, the finite sample breakdown point of the least-squares estimator is zero since its influence function is unbounded.

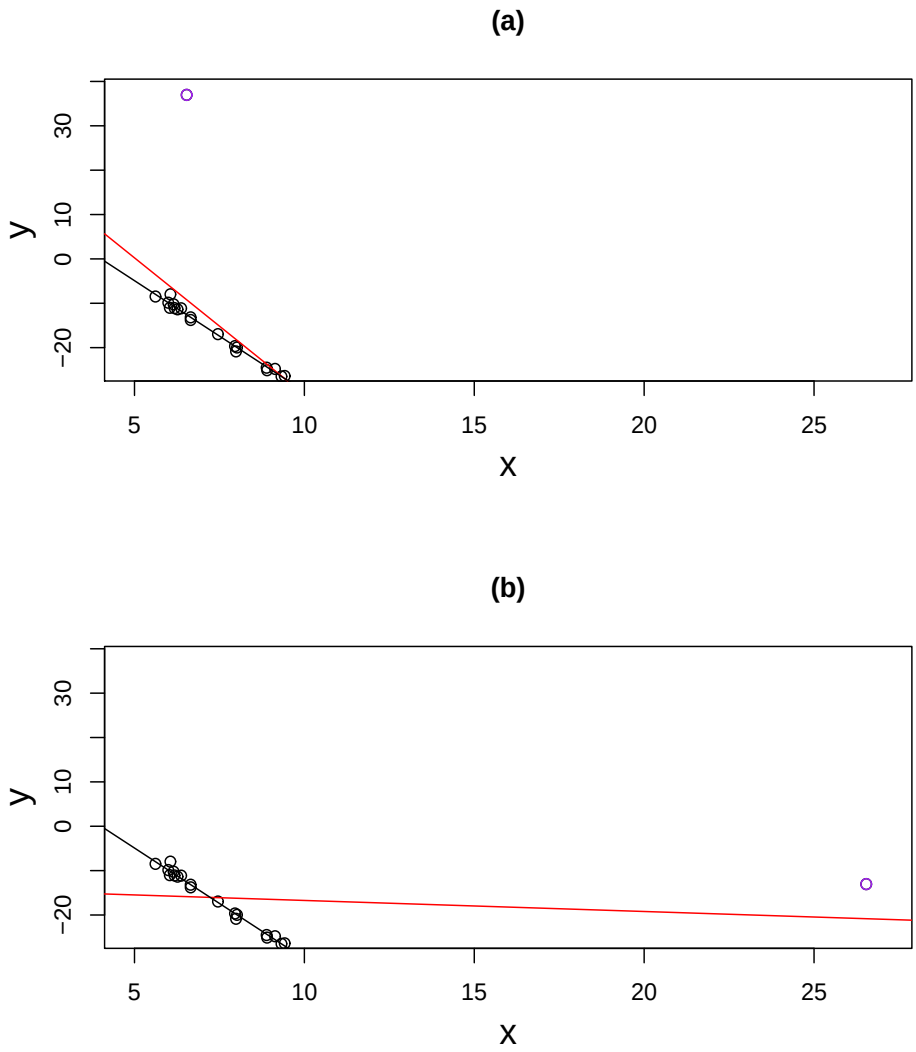

Figure 2.5: Least-squares estimated regression lines: (a) the case of a y-outlier (the purple point) and (b) the case of an x-outlier (the purple point). The red lines are least-squares estimated regression lines based on the sample with the outlier. The black lines are least-squares estimated regression lines based on the original sample (without the outlier). The plots show an outlier of either type can dramatically affect the least-squares fitted line.

Example 2.2. *The influence of one single outlier on the least-squares estimated regression line based on a sample from model*

$$y_i = 20 - 5x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 20, \tag{2.10}$$

where ε_i *are i.i.d.* $N(0, 1)$ *and* x_i *are i.i.d.* $Unif(5, 10)$.

We generated $n = 20$ pairs of (x_i, y_i) observation from (2.10). We first computed the least-squares estimate of the line using this sample of size 20, and this is the black line in Figures 2.5(a) and (b). Then, we drag one selected point far away from the other 19 points in the vertical direction as shown in Figure 2.5(a) to create a y-outlier. We estimated the regression line using the method of least-squares with this altered sample of size 20. This line is in red in Figure 2.5(a), which (in contrast to the black line) does not go through the bulk of the data points. Similarly, we drag the selected point far away from the other 19 points in the horizontal direction as in Figure 2.5(b) to generate an x-outlier. We then computed the least-squares line for this altered sample. This line is shown in red in Figure 2.5(b) and again it does not go through the bulk of the data. These plots show that an outlier in either the x or the y direction can dramatically affect the least-squares fitted line. □

2.3.2 Robust M-estimators

In the Section 2.1, we reviewed the robust M-estimators for location and scale. We now discuss the extension of M-estimation to regression models. The M-estimators for location and scale are special cases of a general M-estimator (for a general unknown parameter θ) defined by the following equation

$$\sum_{i=1}^{n} \Psi(x_i, \theta) = 0.$$

For estimating location, Ψ has the form $\Psi(x,\theta) = \psi(x-\theta)$ with $\theta \in R$. For estimating scale, $\Psi(x,\theta) = \rho(|x|/\theta) - \delta$ with $\theta > 0$. If ψ (or ρ) is non-decreasing, then Ψ is non-increasing in θ. Here we are interested in the special M-estimator for parameters of a linear model given below.

Let $r_i(\beta) = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$. Then the M-estimator $\hat{\boldsymbol{\beta}}$ is the solution of equation

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\mathbf{x}_i = 0, \tag{2.11}$$

where $\hat{\sigma}$ is a robust estimate of scale. Often $\hat{\sigma}$ is chosen to be $\text{MAD}(r_i(\hat{\boldsymbol{\beta}}))$. Here function ψ is still the derivative of function ρ which satisfies properties **C1**-**C4** in Section 2.1. If ψ is bounded and $\hat{\sigma}$ is chosen to be $\text{MAD}(r_i(\hat{\boldsymbol{\beta}}))$, then the influence function for $\hat{\boldsymbol{\beta}}$ is bounded in y. So the M-estimator $\hat{\boldsymbol{\beta}}$ is robust against y-outliers. However, it is not robust against x-outliers as shown in the example below.

Example 2.2 continued: To see that the M-estimator with bounded ψ function is robust against y-outliers but not robust against x-outliers, we use the same data generated from model (2.10) that was used in Figure 2.5, and we generated the y-outlier and x-outlier as described before. Figure 2.6(a) shows that the M-estimator with Huber's ψ function (2.3) works very well when there is only a y-outlier as the estimated line (in green) fits the data well. This is in contrast to the least-squares regression line (in red) which does not fit as well. However, Figure 2.6(b) shows that this M-estimator is not robust against even one single x-outlier, since the fitted line (in green) does not go through the bulk of the data. This fitted line is dragged to the direction of the x-outlier, which is the same to the fitted line estimated by the method of least-squares. □

Because the M-estimator defined in (2.11) is not be robust against x-outliers, other robust estimators (Rousseeuw and Leroy, 1987) have been studied in the literature

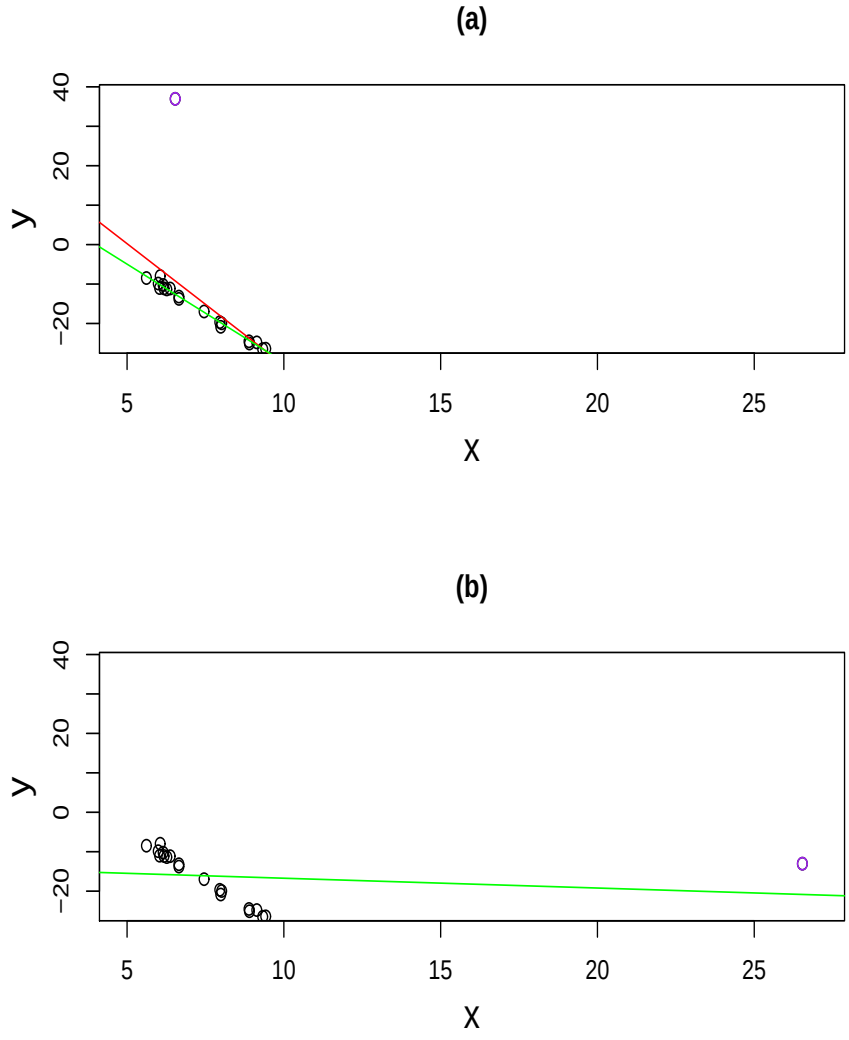

Figure 2.6: In both plots, the red line is fitted by the method of least-squares. The green line is the fitted by the M-estimator, and the purple point in (a) is the y-outlier and the purple point in (b) is the x-outlier.

such as the least median squares, the least trimmed squares and the MM-estimator. The MM-estimator introduced by Yohai (1987) is highly efficient with high breakdown point. It is robust against both y-outliers and x-outliers. This estimation has the following three steps.

1. Compute an initial consistent estimate $\hat{\boldsymbol{\beta}}_0$ with high breakdown point but possibly low normal efficiency.

2. Compute a robust scale $\hat{\sigma}$ of the residuals $r_i(\hat{\boldsymbol{\beta}}_0)$.

3. Find a solution $\hat{\boldsymbol{\beta}}$ of Equation (2.11) using $\hat{\sigma}$ in step 2 and an iterative procedure starting at $\hat{\boldsymbol{\beta}}_0$.

In step 1, we compute an initial consistent estimate $\hat{\boldsymbol{\beta}}_0$ that is robust against outliers with high breakdown point, such as the least trimmed squares estimate (Rousseeuw, 1984) or the least median of squares estimate (Hampel, 1975). Then we compute a robust scale M-estimate $\hat{\sigma}$ based on the vector of residuals $\mathbf{r}(\boldsymbol{\beta}_0) = (r_1(\boldsymbol{\beta}_0), \ldots, r_n(\boldsymbol{\beta}_0))$ by equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i}{\hat{\sigma}}\right) = 0.5, \tag{2.12}$$

where the asymptotic breakdown point of $\hat{\sigma}$ is 0.5. In the last step, we put the robust scale $\hat{\sigma}$ into (2.11) and find the solution $\hat{\boldsymbol{\beta}}$ of (2.11) using an iterative procedure. Note that in each iteration, we compute $\hat{\sigma}$ and $\hat{\boldsymbol{\beta}}$ with two different functions ρ_0 and ρ satisfying $\rho(x) \leq \rho_0(x)$. Both ρ_0 and ρ are bounded satisfying conditions **C1**-**C4** in the Section 2.1. Beaton and Tukey (1974) introduced a family of functions ρ_d, which can be used to compute the MM-estimate, given by

$$\rho_d(x) = \begin{cases} 3(x/d)^2 - 3(x/d)^4 + (x/d)^6, & \text{if } |x| \leq d \\ 1, & \text{if } |x| > d; \end{cases} \tag{2.13}$$

where the tuning constant d is positive. If $d = 1.548$ in ρ_d and $\delta = 0.5$ in (2.12) and $d = 4.68$ in (2.11), the MM-estimate will have 50% asymptotic breakdown point and 95% efficiency when the errors have a normal distribution (Salibian-Barrera, 2003). The useful implementation is implemented in R software (*e.g.* code *lmrob* for linear regression model). It uses an S-estimator (Rousseeuw and Yohai, 1984) for the errors which is also computed with the bisquare function using the Fast-S algorithm of Salibian-Barrera and Yohai (2006) (the function *lmrob.S* for linear model, for example).

Example 2.2 continued 2: The following example illustrates the practice of the MM-estimator and its finite breakdown point. we still use the same data generated from model (2.10) that was used in Figures 2.5 and 2.6, and we generated the y-outlier and x-outlier as described before. Figure 2.7 (a) and (b) show that the MM-estimator works well when there is one y-outlier and one x-outlier, since the fitted lines in these two plots go through the majority of the data. How about the finite breakdown point of these MM-estimates? Figures 2.7(c) and (d) give us the answer. There are eight outliers in Figure 2.7(c). The fitted line from MM still goes through the bulk of the data. However, when there are 9 outliers in Figure 2.7(d), the MM line is heavily influenced by the outliers. So the MM breaks down with 9 outliers. Hence the finite breakdown point for the MM-estimator in this example is $8/20 = 40\%$. This finite breakdown point is very high and the finite breakdown point is always smaller than the asymptotic breakdown point (Donoho and Huber, 1983), even though the asymptotic breakdown point for the MM-estimator is 0.5. In this example, the MM-estimates are robust against both the x-outliers and the y-outliers with high breakdown point. □

However, it is always challenging to compute the MM-estimator and the MM-

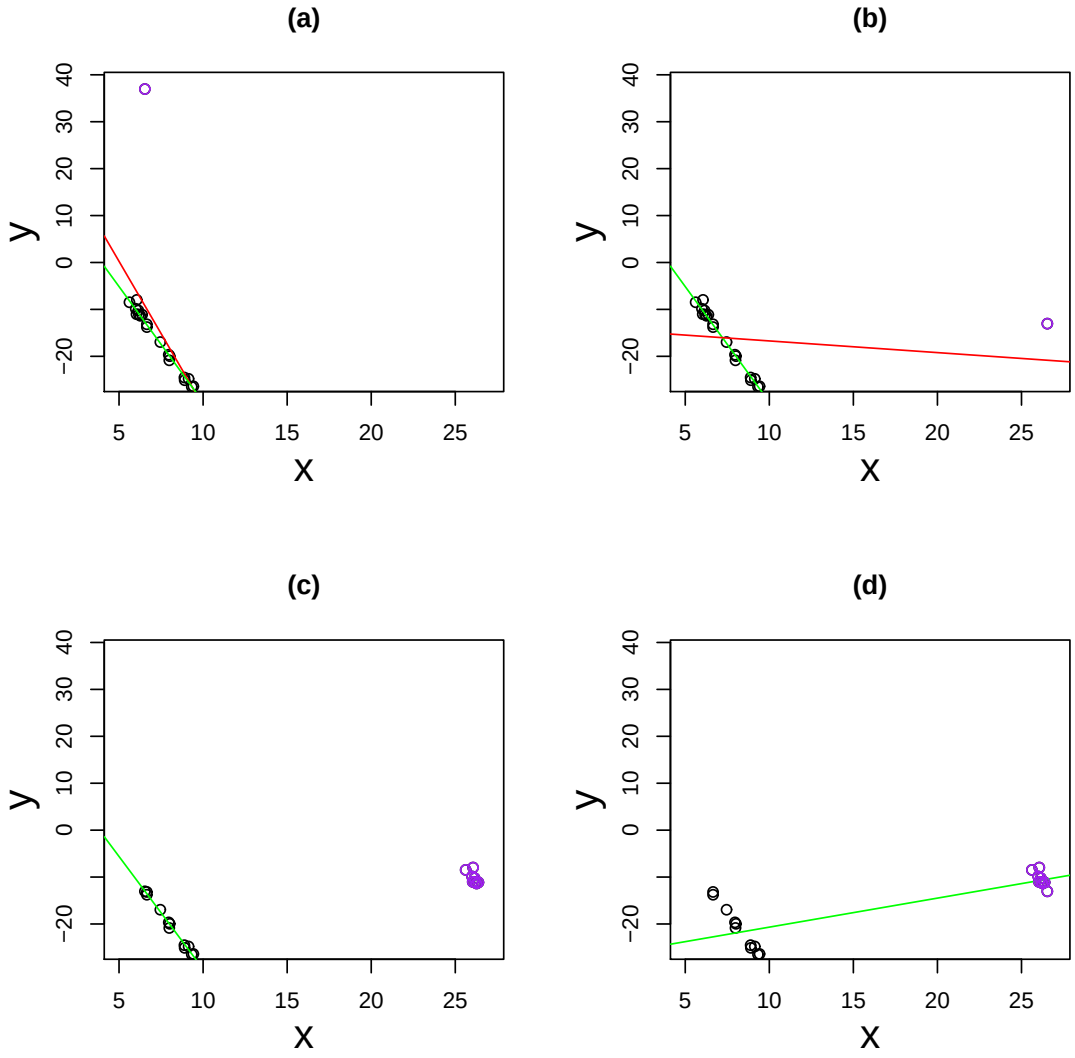

Figure 2.7: The fitted line (red for the LSE, green for the MM) for various outliers colored purple: (a) one y-outlier, (b) one x-outlier, (c) 8 outliers, (d) 9 outliers.

estimator is designed and good for the linear regression models. The robust method to analyze generalized linear regression models is different. The implementation used in R software uses Mallows or Huber type robust estimators, as described in Cantoni and Ronchetti (2001) and currently no other method is implemented.

Chapter 3

Subsampling Method - Proposal I

In this chapter, we propose an alternative robust method, subsampling method, to deal with outliers when fitting regression models. The basic idea is to consider subsamples from the data set and to identify among possibly many subsamples ones that contain only good data (good subsamples). Then estimate the regression model using only the good subsamples through some simple methods. There are different implementations of the subsampling idea. We will describe one implementation in this chapter which we will refer to as *subsampling method proposal I* or simply *proposal I*. In principle, proposal I can be applied for the robust estimation of *any* regression model, provided a reasonable goodness-of-fit measure for the model is available.

The rest of this chapter is organized as follows. In Section 3.1, we introduce subsampling method proposal I and provide a detailed algorithm for its implementation. Some theoretical results concerning this proposal are also given in this section. In Section 3.2, we demonstrate the use of this method and its versatility through several numerical examples on robust estimation of linear regression and logistic regression models. In Section 3.3, we investigate the finite sample performance of this method. We will focus on the bias and the efficiency of the proposal I. We will also examine the

coverage accuracy of the resulting confidence intervals. Further, we compare proposal I with robust MM-method and the method of least-squares. Then in Section 3.4, we will make some remarks on the proposal I.

3.1 Subsampling method and related theory

We now introduce the subsampling method proposal I. To set up notation, suppose we have a sample of $N = n + m$ observations $\mathcal{S}_N = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, \boldsymbol{z}_{n+1}, \ldots, \boldsymbol{z}_{n+m}\}$ from regression model (3.1),

$$E(Y_i) = g(\mathbf{x}_i, \boldsymbol{\beta}), \tag{3.1}$$

where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, y_i is the response and $\mathbf{x}_i$ is the corresponding covariates. Here $E(Y_i)$ denotes the expectation of random variable Y_i underlying the observation y_i and $g(\mathbf{x}_i, \boldsymbol{\beta})$ is the regression function with parameter (vector) $\boldsymbol{\beta}$. To completely specify a parametric regression model, we also need a distributional assumption on Y_i. But since our general discussion of subsampling method in this section does not involve the distribution of Y_i, we leave this distribution unspecified. Throughout this and the next chapter, we assume that $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ are n "good data points" randomly generated by the underlying model (3.1), and $\boldsymbol{z}_{n+1}, \ldots, \boldsymbol{z}_{n+m}$ are m "bad points" or outliers, which are contaminated observations. Note that regression model (3.1) includes as special cases the linear models, non-linear models and generalized linear models. So the subsampling method we describe below applies to all such models.

3.1.1 Subsampling algorithm-proposal I

The objective here is to construct a "good subsample" S_g of $\mathcal{S}_N$ that contains only good data points. To find such a good subsample, consider a random subsample of size n_s, S_{n_s}, taken without replacement from $\mathcal{S}_N$. We assume $m < n_s \leq n$ which

ensures that S_{n_s} cannot be consisted entirely of bad data points and that there exists at least one S_{n_s} containing only good data points. Further, for some simple non-robust method Π of fitting the regression model to the subsamples (such as the method of least-squares), we assume there is an associated quantitative goodness-of-fit criterion Γ (such as the mean squared error, AIC or BIC) which may be indicative of the presence of outliers in S_{n_s}. Let γ be the numerical score given by the criterion Γ upon fitting the model (3.1) to S_{n_s} using method Π, and suppose a small γ value means a good fit. We compute the good subsample S_g from S_{n_s} through following algorithm.

Algorithm SA1(n_s, r^*, k): *Subsampling algorithm-proposal I* For specified n_s, r^* and k:

Step 1: Randomly draw a subsample $S^1_{n_s}$ without replacement from the original sample $\mathcal{S}_N = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, \boldsymbol{z}_{n+1}, \ldots, \boldsymbol{z}_{n+m}\}$.

Step 2: Fit the regression model (3.1) to the subsample obtained in Step 1 and compute the corresponding goodness-of-fit measure γ_1.

Step 3: Repeat Steps 1 and 2 for $j = 1, 2, \ldots, k$ times. Each time record $(S^j_{n_s}, \gamma_j)$, the subsample taken and the associated goodness-of-fit measure at the jth repeat.

Step 4: Sort the k subsamples by the size of their associated γ values; denote by $\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(k)}$ the ordered values of γ_j, and by $S^{(1)}_{n_s}, S^{(2)}_{n_s}, \ldots, S^{(k)}_{n_s}$ the correspondingly ordered subsamples.

Step 5: Take the union of the r^* subsamples with the smallest γ values to form the good subsample S_g. That is

$$S_g = \bigcup_{j=1}^{r^*} S^{(j)}_{n_s}. \tag{3.2}$$

We make the following remarks on the algorithm and terminology:

1. An alternative to Step 5, **Step 5***, is to use a cutoff point for $\gamma_{(j)}$, say γ_C, and take the union of those subsamples with $\gamma_{(j)} \leq \gamma_C$ to form S_g. That is

$$S_g = \bigcup_{\gamma_{(j)} \leq \gamma_C} S_{n_s}^{(j)}. \tag{3.3}$$

   This gives a better control over the subsamples to be included in the union as they all have γ values smaller than the cutoff point. But the number of subsamples included in the union r^* becomes a random variable.

2. When n, m and n_s are sufficiently small such that the *total number* of different subsamples of size n_s is not large, instead of drawing random subsamples in Steps 1-3, we may simply go through all possible subsamples one-by-one in these three steps. In this case k is the total number of such subsamples and j is the unique label for a subsample.

3. On terminology, since S_g is a union, it is also referred to as the *combined subsample*. Also, the term *subsampling method-proposal I* refers to using the sampling algorithm SA1(n_s, r^*, k) to compute the combined sample S_g **and** then estimating and making inference of the parameters of model (3.1) using a simple, existing (typically non-robust) method on S_g. Although this method does not have to be the same method Π involved in the algorithm, for convenience and consistency we will use the same method Π for both computing S_g through SA1(n_s, r^*, k) and the subsequent estimation and inference based on S_g.

For successful implementation of this algorithm, the goodness-of-fit criterion Γ and the values of parameters in (n_s, r^*, k) must be carefully chosen. We will discuss theoretical considerations concerning the selection of the latter in the next section.

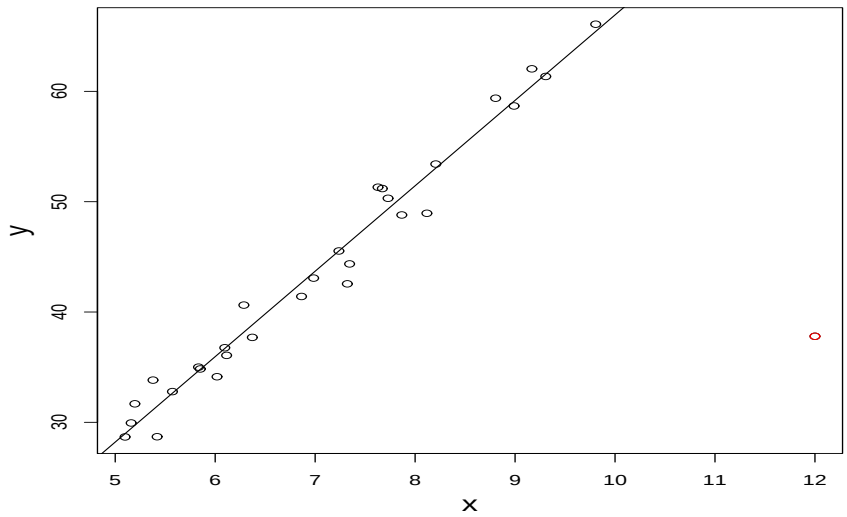

Figure 3.1: Scatter plot of a contaminated sample of 30 from model (3.4)

The basic assumption underlying this algorithm is that, with careful selection of the Γ and the parameter values, those subsamples containing outliers will have large γ values and hence be excluded from the combined sample S_g. There are two problems with this algorithm: [a] the basic assumption may be invalid, and [b] even when the assumption is valid and S_g contains no outliers, S_g may not be a random subsample of the good data points and hence not a random sample from the underlying model. Problem [a] may lead to outliers going undetected into the combined subsample S_g. This may be handled by employing additional criteria to identify subsamples with outliers and keeping such subsamples from the union in (3.3) and (3.2). Later in Section 3.4, we will discuss modified algorithms based on such additional criterion. Problem [b] will affect the interpretation of S_g as well as that parameter estimation and inference based on S_g. We will address these issues in Sections 3.1.2 and 3.4.

We conclude this section with a simulation example typical of situations where we expect the above algorithm to be successful in finding a good sample S_g.

Example 3.1. *Consider a sample of size* $N = 30$ *from linear model (3.4) with one*

outlier.

$$y = -10.7 + 7.87x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 = 4), \tag{3.4}$$

with values of x *generated by uniform distribution on* $[5, 10]$.

We generated $N = 30$ (x, y) observations from model (3.4) and then doubled the x-value of one randomly selected point from $x = 6$ to $x = 12$ to create an x-outlier. See Figure 3.1 for a scatter plot of the resulting data set of size $N = n + m = 29 + 1$. Now consider subsamples of size $n_s = 29$. There are a total of 30 such subsamples, among which exactly one contains no outliers. We fitted the simple linear model using the method of least-squares to all such subsamples and computed the mean squared error (MSE=$SS_{Residual}/(n_s - 2)$) of each fit. Figure 3.2 shows the plot of the MSE versus the subsample label/number. The MSE of the sample without the outlier is seen to be substantially smaller than that of the other subsamples. In this case, the SA1(29, 1, 29) recovered all 29 good data points through S_g. The use of $r^* = 1$ was partly based on the plot in Figure 3.2. Alternatively, we may use Step 5* where r^* is the number of subsamples satisfying $\gamma \leq \gamma_C = 2\gamma_{(1)}$. This would also recover all 29 good data points as well. This point is discussed further in the next section. As the last step of the subsampling method-proposal I, we fitted the linear model to S_g using the method of least-squares. The fitted line in Figure 3.1 is not affected by the outlier. In this case, all standard least-squares method based inference for linear models would apply as S_g may be viewed as a random sample from model (3.4). □

In this example, the number of outliers is known and the total number of subsamples of size 29 is merely 30, which made it possible to run SA1(n_s, r^*, k) algorithm on all possible subsamples. In real applications, however, $N = n + m$ may be large and the number of all possible subsamples $\binom{n+m}{n_s}$ may be so large that enumerating all subsamples is computationally impossible. In this case, the algorithm will have to be implemented by doing random subsampling in Steps 1-3. We now calculate the

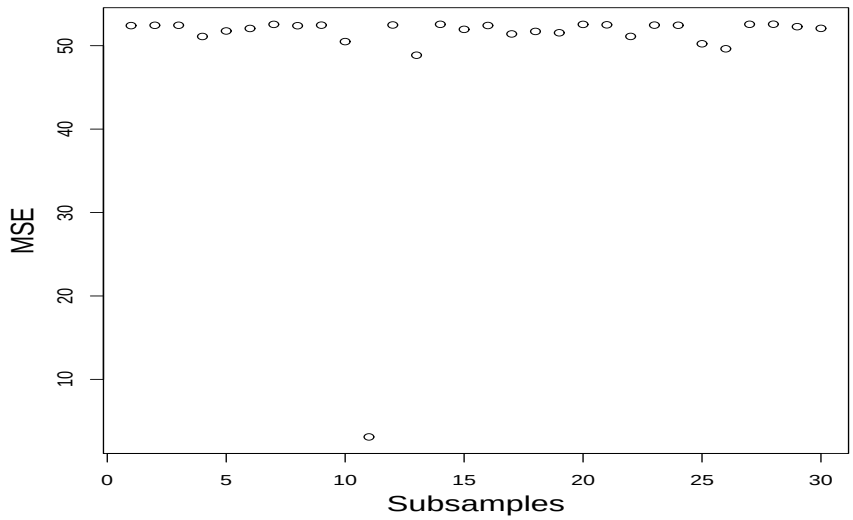

Figure 3.2: MSE plot for all combinations of original data

theoretical considerations for parameters selection for the algorithm.

3.1.2 Theoretical considerations

It is clear that the subsampling method for removing outliers will not be one-hundred percent successful. Regardless of the selection of (n_s, r^*, k) and Γ, there is always a small probability that outliers will be present in the combined sample S_g. What we can hope to do is to control the magnitude of such a probability. Another issue with the subsampling method is the size of S_g relative to that of $\mathcal{S}_N$. We can control the probability that S_g containing outliers by limiting its size but the resulting S_g may be only a small fraction of the good data points. This is inefficient as most of the good data points are not in the S_g, and hence not utilized for the estimation of and inference on the regression model. Nevertheless, with the proper selection of (n_s, r^*, k), we can achieve a desired level of efficiency and control the probability at an acceptable level. We address these issues in this section.

Denote by A a subsample with size n_s from S_N taken without replacement. The

probability that A contains only good data (A is a good subsample) is

$$p_A = \frac{\binom{n}{n_s}\binom{m}{0}}{\binom{n+m}{n_s}} > 0. \tag{3.5}$$

Here, p_A is positive due to the condition that $n_s \leq n$. Now consider a sequence of independent subsamples $A_1, A_2, \ldots, A_k$ from $\mathcal{S}_N$, each with size n_s and taken without replacement. Let T be the total number of subsamples which contain no outliers and denote by p_i the probability that at least i of these subsamples contain no outliers. Then T has a binomial distribution,

$$T \sim BIN(k, p_A), \tag{3.6}$$

and as a simple consequence of (3.6),

$$p_i = P(\text{at least } i \text{ good subsamples}) = 1 - \sum_{j=0}^{i-1} \binom{k}{j} p_A^j (1 - p_A)^{(k-j)}. \tag{3.7}$$

Probability p_i will be used in determining parameter settings in SA1(n_s, r^*, k). For convenience, we will refer to p_i as the probability of at least i good subsamples.

The combined sample S_g may be viewed as an "estimation" of the good data $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$. The following is a consistency result related to this estimation.

Theorem 3.1. *Let $A_1, A_2, \ldots$ be an infinite sequence of independent random subsamples of size $n_s < n$, each taken without replacement from $\mathcal{S}_N$. Let $A_1^*, A_2^*, \ldots$ be the subsequence of those containing no outliers and define a partial union B_j as*

$$B_j = \bigcup_{i=1}^{j} A_i^*. \tag{3.8}$$

Then

$$P(B_\infty = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}) = 1. \tag{3.9}$$

Proof: We first note that with probability 1, the subsequence $A_1^*, A_2^*, \ldots$ is an infinite sequence. This is so because the event that this subsequence contains only finitely many subsamples is equivalent to the event that there exists an l for $A_1, A_2, \ldots$ such that after l, $A_l, A_{l+1}, \ldots$ contains no more good subsamples. Since $p_A > 0$, the probability of this latter event and hence that of the former are both zero.

Next, under the condition that $A_j^* \subseteq \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, A_j^* may be viewed as a random sample of size n_s taken directly without replacement from $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$. Hence with probability 1, $A_1^*, A_2^*, \ldots$ is an infinite sequence of random samples from the finite population $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$. It follows that, for any fixed $1 \leq j \leq n$, the probability that z_j is in at least one of the A_i^* is 1. Hence $P(\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\} \subseteq B_\infty) = 1$. This and the fact that $B_\infty \subseteq \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ imply (3.9). □

The above theorem avoids the issue of identification of subsamples with outliers and this greatly simplifies the discussion. It does show, however, that we can eventually recover the good data by repeated subsampling and then combining good subsamples, provided we are able to separate good subsamples (without outliers) from bad ones (without outliers). The combined sample S_g given by SA1(n_s, r^*, k) may be viewed as a special case of the partial sum B_j. We say that S_g is consistent in the sense that its counterpart in the idealized situation, B_j, is consistent. We now examine the finite sample efficiency of S_g, again by focusing on the corresponding B_j.

Let W_j be the number of elements $(\boldsymbol{z}_j, j \leq n)$ in the B_j given by (3.8). Clearly, a large W_j means B_j is efficient in recovering the good data $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$. By Theorem 3.1, $P(W_j = n) \to 1$ as $j \to \infty$. Hence it is asymptotically 100% efficient. However, at a fixed j, we want a measure of efficiency for B_j. Since W_j is a random

variable, a natural efficiency measure $E_F(B_j)$ is given by

$$E_F(B_j) = \frac{E(W_j)}{n}, \quad j = 1, 2, \ldots,$$

where $E(W_j)$ is the expected number of (good) data points picked up by B_j. The following theorem gives a simple expression of $E_F(B_j)$ in terms of n and n_s.

Theorem 3.2. *The efficiency of B_j in recovering good data points is*

$$E_F(B_j) = \frac{E(W_j)}{n} = 1 - \left(\frac{n - n_s}{n}\right)^j, \quad j = 1, 2, \ldots. \tag{3.10}$$

Proof: We prove (3.10) by induction.

For $j = 1$, since $B_1 = A_1^*$ which contains exactly n_s points, W_1 is a constant and

$$E_F(B_1) = \frac{E(W_1)}{n} = \frac{n_s}{n}.$$

Hence (3.10) is true for $j = 1$.

To find $E_F(B_2)$, we need to find the expected value for $E(W_2)$. To this end, we first find the distribution of W_2. Divide the good data $\{z\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ into two groups, A_1^* containing n_s points and its complement $\bar{A}_1^*$ which contains $n - n_s$ points. Since A_2^* is a random sample of size n_s taken without replacement from $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, we assume without loss of generality that it contains U_1 data points from $\bar{A}_1^*$, and hence it contains $n_s - U_1$ data points from A_1^*. It follows that U_1 has a hypergeometric distribution

$$U_1 \sim \text{Hyperg}(n, n - n_s, n_s), \tag{3.11}$$

with expected value

$$E(U_1) = n_s \frac{n - n_s}{n}.$$

Now, since $B_2 = A_1^* \cup A_2^*$, its number of data points W_2 is

$$W_2 = n_s + U_1.$$

Hence

$$E(W_2) = n_s + E(U_1) = n_s + n_s \frac{(n - n_s)}{n} = n \left[1 - \left(\frac{n - n_s}{n} \right)^2 \right].$$

It follows that

$$E_F(B_2) = 1 - \left(\frac{n - n_s}{n} \right)^2.$$

Similarly, $W_3 = \#(A_1^* \cup A_2^* \cup A_3^*) = \#(B_2 \cup A_3^*) = W_2 + U_2$, where U_2 is the number of new points not already in B_2 and $\#(A)$ denotes the total number of points in set A. Further the conditional distribution of U_2 given w_2 is also hypergeometric

$$U_2 | w_2 \sim \text{Hyperg}(n, n - w_2, n_s).$$

In general, $W_j = \#(B_{j-1} \cup A_j^*) = W_{j-1} + U_{j-1}$ and

$$U_{j-1} | w_{j-1} \sim \text{Hyperg}(n, n - w_{j-1}, n_s).$$

Assume $E(W_{j-1}) = nE_F(B_{j-1}) = n[1 - ((n - n_s)/n)^{j-1}]$. Then

$$\begin{aligned} E(W_j) = E(W_{j-1}) + E(E(U_{j-1}|W_{j-1})) &= E(W_{j-1}) + E\left[n_s\left(\frac{n - W_{j-1}}{n}\right)\right] \\ &= n_s + \frac{n - n_s}{n}E(W_{j-1}) \\ &= n_s + (n - n_s)\left[1 - \left(\frac{n - n_s}{n}\right)^{j-1}\right] \\ &= n - \frac{(n - n_s)^j}{n^{j-1}} \\ &= n\left[1 - \left(\frac{n - n_s}{n}\right)^j\right]. \end{aligned}$$

Thus

$$E_F(B_j) = \frac{E(w_j)}{n} = 1 - \left(\frac{n - n_s}{n}\right)^j,$$

which proves Theorem 3.2. □

Theorem 3.2 indicates that $E_F(B_j) \to 1$ as $j \to \infty$. The convergence is very fast when $(n - n_s)/n$ is not close to 1. Recall that the combined sample S_g given by Step 5 of the SA1(n_s, r^*, k) algorithm is the union of r^* good subsamples. If these r^* good subsamples are independent, then S_g is just a B_{r^*} and thus

$$E_F(S_g) = E_F(B_{r^*}). \tag{3.12}$$

In practice, equation (3.12) holds only approximately since the r^* good subsamples in Step 5 may not be independent (see further discussion below). Nevertheless, due to the connection given by (3.12), we denote the *efficiency* of algorithm SA1(n_s, r^*, k) as $E_F(S_g)$ and define it to be $E_F(B_{r^*})$.

Finally, we comment on problem [b] of the subsampling algorithm described in Section 3.1.1. In the idealized situation where we have independent subsamples $A_1^*, A_2^*, \ldots$ from the good data $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, each B_j may be viewed as a ran-

dom sample from the good data and thus it is a random sample from the underlying regression model. So when we fit the regression model to data in B_j using any existing method, say the method of least-squares, inference procedures associated with the least-squares method would be valid. However, the good subsamples $S_{n_s}^{(1)}, S_{n_s}^{(2)}, \ldots, S_{n_s}^{(r^*)}$ that we use to form S_g in Step 5 of SA1(n_s, r^*, k) may not be independent random samples from the good data. For example, due to the γ-ordering, $S_{n_s}^{(1)}$ may be viewed as the subsample of the good data that is most "favorable" to the model. Because of this, the combined sample S_g may be a "biased" subsample from the good data in that it may contain only those good data points that are most favorable to the model. This biased nature of S_g will have an impact on the validity of the least-squares estimation and inference. Research is ongoing to investigate the bias of the S_g and its impact on the subsequent inferences based S_g. To tentatively deal with this problem in this thesis, we will set the efficiency rate at the high value of 99%. Although such a high efficiency is computationally expensive to achieve, empirical evidence suggests that the resulting combined sample S_g is very effective in recovering the good data points. As such, S_g may be viewed as the set of good data points itself and hence is a random sample from the underlying model. Because of this, inferences associated with the least-squares method based on S_g are valid.

3.1.3 How to choose parameter values for the subsampling algorithm SA1(n_s, r^*, k)

We now discuss the application of the above theoretical developments for setting the parameter values in SA1(n_s, r^*, k). Before we can apply the theoretical results, we need to know the number of outliers m in $\mathcal{S}_N$. In real applications, this information is not available. Thus we make a working assumption that we have an estimate of the (maximum) percentage of outliers which leads to estimates of m and n. As the default

setting, we assume this percentage is 10%. Correspondingly, the default values for m and n are set to $0.1N$ and $0.9N$, respectively.

The subsample size n_s can be determined through a combination of qualitative and quantitative considerations. First, a necessary condition for n_s is that it must satisfies $n_s > m$ so that a subsample consisting entirely of outliers will not be undetected. To see this is necessary, consider the example where 95% of $\mathcal{S}_N$ are from one linear model but the remaining 5% (outliers) are from another very different linear model. If a subsample happens to contain only points from this 5%, then a goodness-of-fit criterion for linear model may indicate this subsample is a good subsample. Once labeled a good subsample, outliers from this subsample may be included in the combined sample S_g which will result in the failure of the subsampling algorithm. In practice, whenever feasible we recommend setting n_s to the much larger value of $0.5N + 1$ to ensure the subsampling method has the high finite sample breakdown point.

The selection of r^* is tied to the efficiency that we want. Suppose we want an efficiency of at least 99%, *i.e.,* $E_F(B_{r^*}) \geq 0.99$. Once n and n_s are determined, we can find the r^* that satisfies this efficiency requirement through (3.10). Specifically, we choose r^* to be

$$r^* = \arg\min_j \{E_F(B_j) \geq 0.99\}. \tag{3.13}$$

Then we take B_{r^*} to be S_g.

The choice of k is tied to the desired r^* value. This is so since in practice we do not have an infinite sequence of good subsamples $A_1^*, A_2^*, \ldots$ to work with. In order to have at least r^* good subsamples, we need to run the algorithm with a suitably large k value so that among the k subsample generated, there is a high probability that there will be at least r^* good subsamples. We call this probability the *probability of having required number of good subsamples* and denote it by p^*. The probability of

Table 3.1: The number of subsamples k and the number of good subsamples r^* required to achieve a 99% efficiency

Sample Size N	Number of good data n	Number of Outliers m	Subsample size n_s	r^*	Number of subsamples k
$N = 30$	$n = 30$	$m = 0$	$n_s = 16$	$r^* = 7$	$k = 7$
	$n = 27$	$m = 3$	$n_s = 16$	$r^* = 6$	$k = 143$
	$n = 25$	$m = 5$	$n_s = 16$	$r^* = 5$	$k = 823$
$N = 50$	$n = 50$	$m = 0$	$n_s = 26$	$r^* = 7$	$k = 7$
	$n = 45$	$m = 5$	$n_s = 26$	$r^* = 6$	$k = 650$
	$n = 42$	$m = 8$	$n_s = 26$	$r^* = 5$	$k = 8468$

at least i good subsamples, p_i, in (3.7) is a monotone increasing function of k and it approaches 1 as k approaches infinity. Thus if we wish to have a k such that there is a probability of $p^* = 0.99$ of having at least r^* good subsample in $A_1, A_2, \ldots, A_k$, we set k to

$$k = \arg\min\{p_{r^*} \geq p^* = 0.99\}. \tag{3.14}$$

To illustrate the determination of r^* and k and to prepare for the applications of the subsampling algorithm in subsequent examples, we include in Table 3.1 the r^* values and k values required in order to achieve a 99% efficiency for various combinations of m and n values. To arrive at this table, the p^* value used to compute k was set to 0.99, and the subsample size was set to $n_s = 0.5N + 1$ for all cases.

The computationally intensive nature of SA1(n_s, r^*, k) can be seen in the case given by $(N, m, n) = (50, 8, 42)$, where to achieve an efficiency of 99% we need to fit the model in question to 8468 subsamples of size 26 each. This would not be a problem if the model to be fitted is a simple linear model. But it could be very time consuming if the model is more complicated and the computation required for fitting the model to each subsample is time consuming.

To visualize the relationship among various parameters of the algorithm, for the

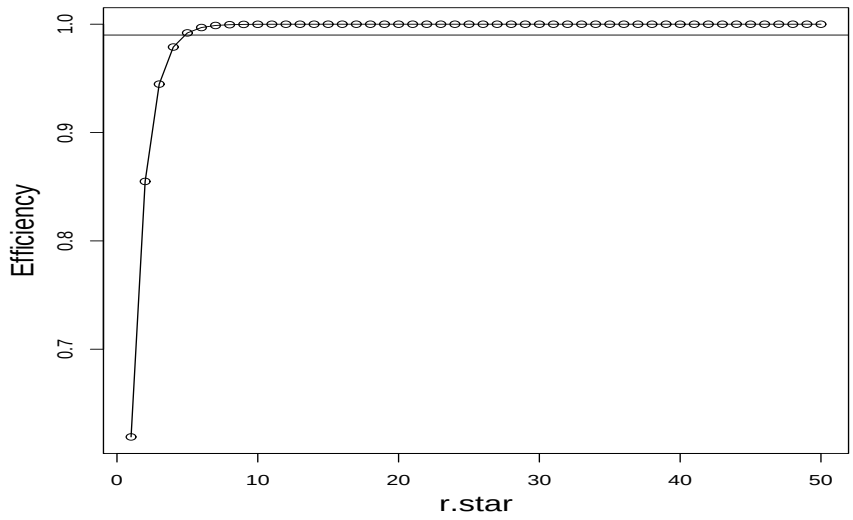

Figure 3.3: The efficiency of the subsampling algorithm SA1(n_s, r^*, k) as a function of the number of good subsamples r^* that form the combined sample S_g. The black line is at the 99% efficiency.

case of $(m, n) = (8, 42)$, Figure 3.3 shows the efficiency of the algorithm $E_F(S_g)$ verses r^* the number of good subsamples involved in S_g. The solid black line in the plot represents the 99% efficiency line. With subsample size $n_s = 26$, we see that the efficiency curve rises rapidly as r^* values increases. At $r^* = 5$, the efficiency reaches 99%. This example shows that, with a properly chosen n_s value, the number of good subsamples required to achieve a highly efficient S_g does not have to be very large.

For this example, we also plotted the probability of having at least i good subsamples p_i verses i when the total number of subsamples is fixed at $k = 7000$ in 3.4(a). Here we see that the probability of having at least $r^* = 5$ (required for 99% efficiency) is only about 0.91. To ensure a high probability of having at least 5 good subsamples, we need to take more subsamples. In 3.4(b), we plotted the same curve but at $k = 8468$ as was shown in Table 3.1. There we see that the probability of having at least 5 good subsamples is now at 0.99.

To summarize, to set the parameters r^* and k for SA1(n_s, r^*, k), we need the

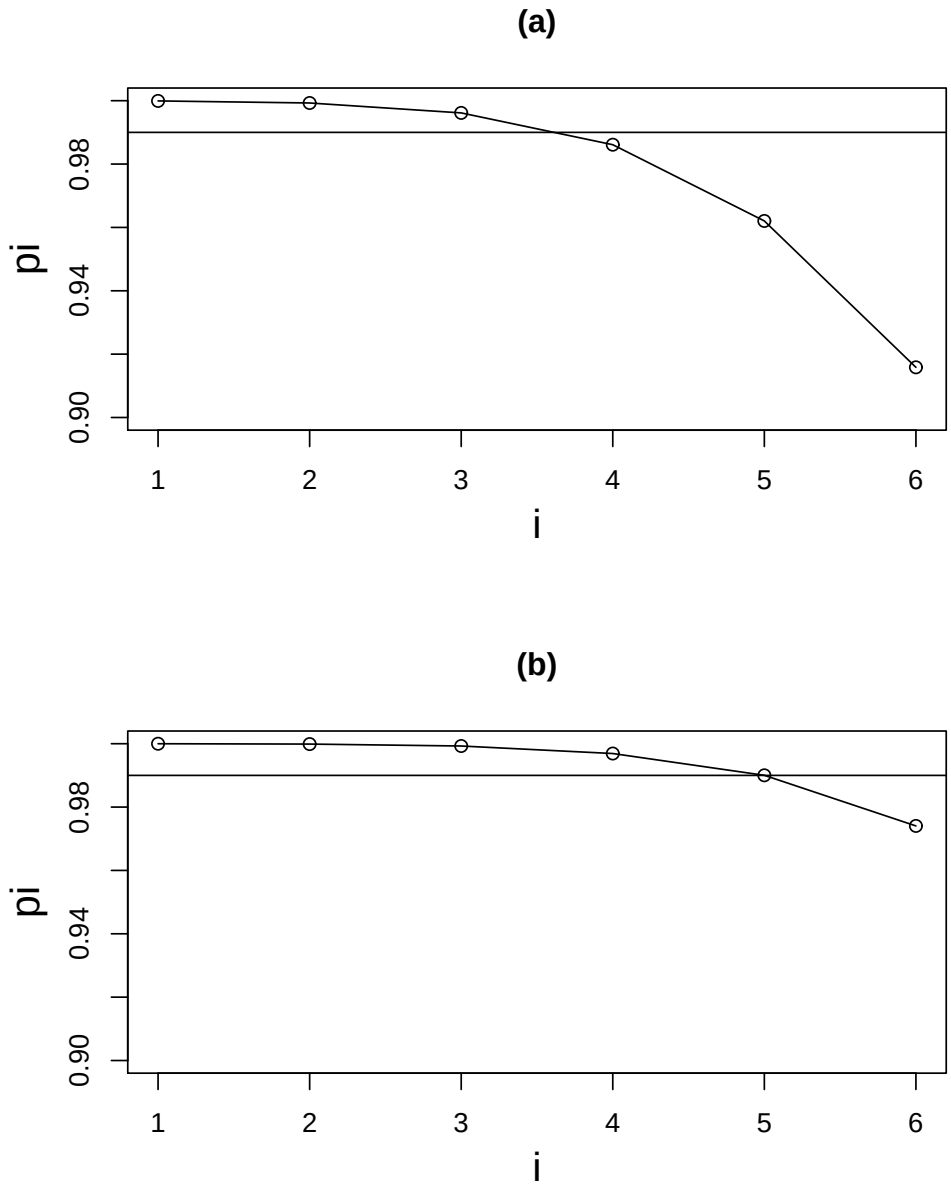

Figure 3.4: (a) The probability of having at least i good subsamples in 7000 subsamples. (b) the probability of having at least i good subsamples in 8468 subsamples. The solid black line is the $p_i = 0.99$ line.

follow ingredients: (1) an estimated percentage of outliers in the sample $\mathcal{S}_N$ (default=10%), (2) an efficiency requirement (default $E_F(S_g) = 99\%$) which reflects the percentage of good data that we wish to recover with the combined sample S_g and [3] a desired probability p^* (default=0.99) of having the required r^* good subsamples. We have developed an R program which computes the required r^* and k for any combination of $(N, m, n, n_s, E_F(S_g), p^*)$. The default value of this input vector is $(N, 0.1N, 0.9N, 0.5N + 1, 99\%, 0.99)$. Note that the determination of such algorithm parameters does not depend on the actual model being fitted. It does assume implicitly that the good and bad subsamples can be accurately separated by the criterion Γ that we employ. Since the criterion is model dependent, we have thus far avoided this issue in our general discussion of the subsampling algorithm. We will examine the choice of an criterion when we apply the subsampling method on specific examples and comment to supplementary/secondary criterion in Section 3.4.

3.2 Applications of subsampling method

In this section, we apply the subsampling method-proposal I to examples involving linear and logistic regression model. Through these different examples, we demonstrate the implementation of the subsampling method for different regression models and, in particular, discuss the selection of criterion Γ select for the subsampling algorithm SA1(n_s, r^*, k). We begin with an example involving a simple linear regression model. This is followed by an example on the multiple linear regression model. We then consider a logistic regression example. This last example also demonstrates the point that subsampling method can indeed be applied to different regression models.

Example 3.2. *Consider $N = 50$ random observations (x_i, y_i), where 45 of these are from a simple linear regression model of interest (3.15) and the remaining 5 are from*

a "contaminating" model (3.16).

$$y_i = -15.7 + 10.87x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma = 2), \tag{3.15}$$

$$y_i = -15.7 + \frac{10.87}{2}x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma = 2), \tag{3.16}$$

with values of x_i *generated by uniform distribution on* $[5, 10]$ *in (3.15) and on* $[8, 11]$ *in (3.16). We wish to make inference on parameters of (3.15) with this sample of* $N = 50$.

Here, the 45 observations from (3.15) represent the good data and the remaining 5 the outliers. To apply SA1(n_s, r^*, k) to find the combined sample S_g, we first set the percentage of outliers to the default value of 10%, which coincides with the true percentage of outliers for this example. So $(N, m, n) = (50, 5, 45)$. Next, we set the subsample size n_s, the efficiency level $E_F(S_g)$ and the probability $p*$ to their default values. The complete input vector for the R program for computing algorithm parameters r^* and k is now $(N, m, n, n_s, E_F(S_g), p^*) = (50, 5, 45, 26, 99\%, 0.99)$. Since this particular case is covered by Table 3.1, from the table we have $(r^*, k) = (6, 650)$.

A model-dependent issue for the implementation of SA1(n_s, r^*, k) is the selections of the estimation method Π and the goodness-of-fit criterion Γ. For linear models, we use the least-squares method as our choice of Π. We find the mean squared error (MSE) a good choice for Γ in that the MSE induced ordering of the subsamples is effective in separating good subsamples from the bad ones; subsamples with small MSEs are usually good subsamples and those with large MSEs are bad ones. To further illustrate this point, Figure 3.5 shows the MSEs of all 650 subsamples (plotted against the subsample index/label) that the SA1(n_s, r^*, k) algorithm has generated. There are 12 dots at the bottom, representing 12 good subsamples containing no outliers. From these 12 good subsamples, the $r^* = 6$ subsamples with the smallest

MSEs are plotted in Figure 3.6. See plots (a)-(f) in this Figure. These 6 good subsamples then form the combined sample S_g which contains 43 distinct points. Hence the *observed efficiency* (recovery rate of good data) is 43/45% for this particular run of the algorithm, which is roughly 96%. For this example, the observed efficiency fluctuates between 95% to 100%.

Plot (a) in Figure 3.7 shows the entire data set of $N = 50$ observations as well as two fitted lines; the red line is the least-squares fitted regression line based on the entire sample of 50 points and green line is the subsampling regression line, which is obtained by using the least-squares method to the combined sample S_g. We see that the subsampling method has successfully excluded the outliers from the combined sample S_g and the subsampling regression line accurately captured the trend in the good data. This is further confirmed in Figure 3.7(b) which shows the plot of the points in the combined sample S_g only. Clearly, no outliers are in S_g.

To compare the estimated parameter values given by the subsampling method with those given by the least-squares method, Table 3.2 shows these estimated values and their estimated standard errors. We see that the 5 outliers have a big impact on the estimated slope and intercept given by the least-squares method but have no impact on the subsampling estimates. We also observed that the subsampling method underestimates σ. This is tied to the insufficient observed efficiency of 43/45; it's likely that the two good data points further away from the true regression line were excluded from the combined sample. Hence the error standard deviation is underestimated.

Finally, in Figure 3.8, we plotted the residuals based on subsampling estimated model and the least-squares estimated model (using all $N = 50$ data points). There we see the subsampling residuals of the 5 outliers are easy to identify in plot (b), but least-squares residuals of the outliers in plot (a) are less distinguished. □

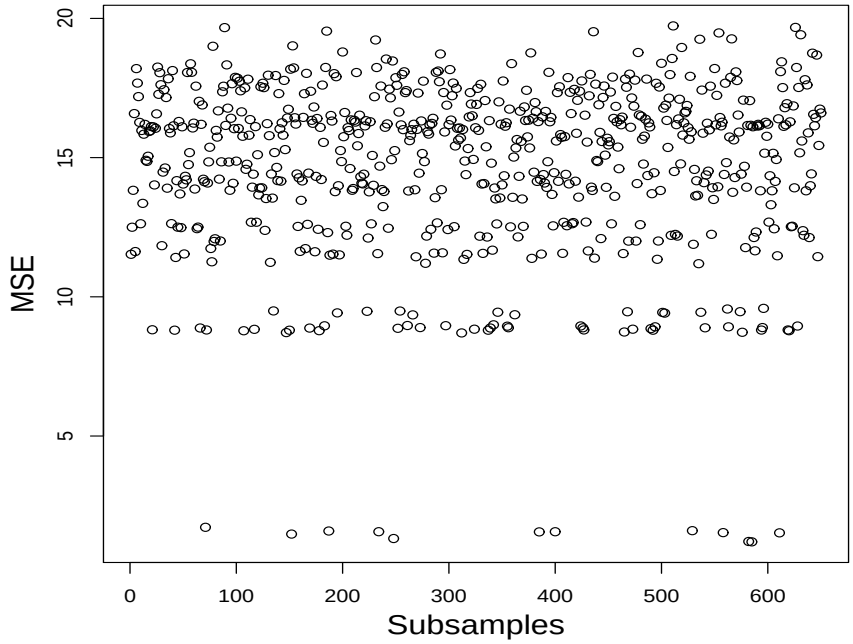

Figure 3.5: Mean squared errors (MSE) of $k = 650$ subsamples from the sample of $N = m + n = 5 + 45$ observations

Table 3.2: Least-squares (LS) estimates and subsampling estimates (SM1) of model (3.15) parameters based on a contaminated sample of $N = 50$. Standard errors (s.e.) are in brackets

	True value (s.e.)	LS estimates (s.e.)	SM1 estimates (s.e.)
$\hat{\beta}_0$	-15.7	5.984 (10.634)	-15.599 (1.119)
$\hat{\beta}_1$	10.87	7.399 (1.356)	10.862 (0.146)
$\hat{\sigma}$	2	15.028 (14.874)	1.548 (1.416)

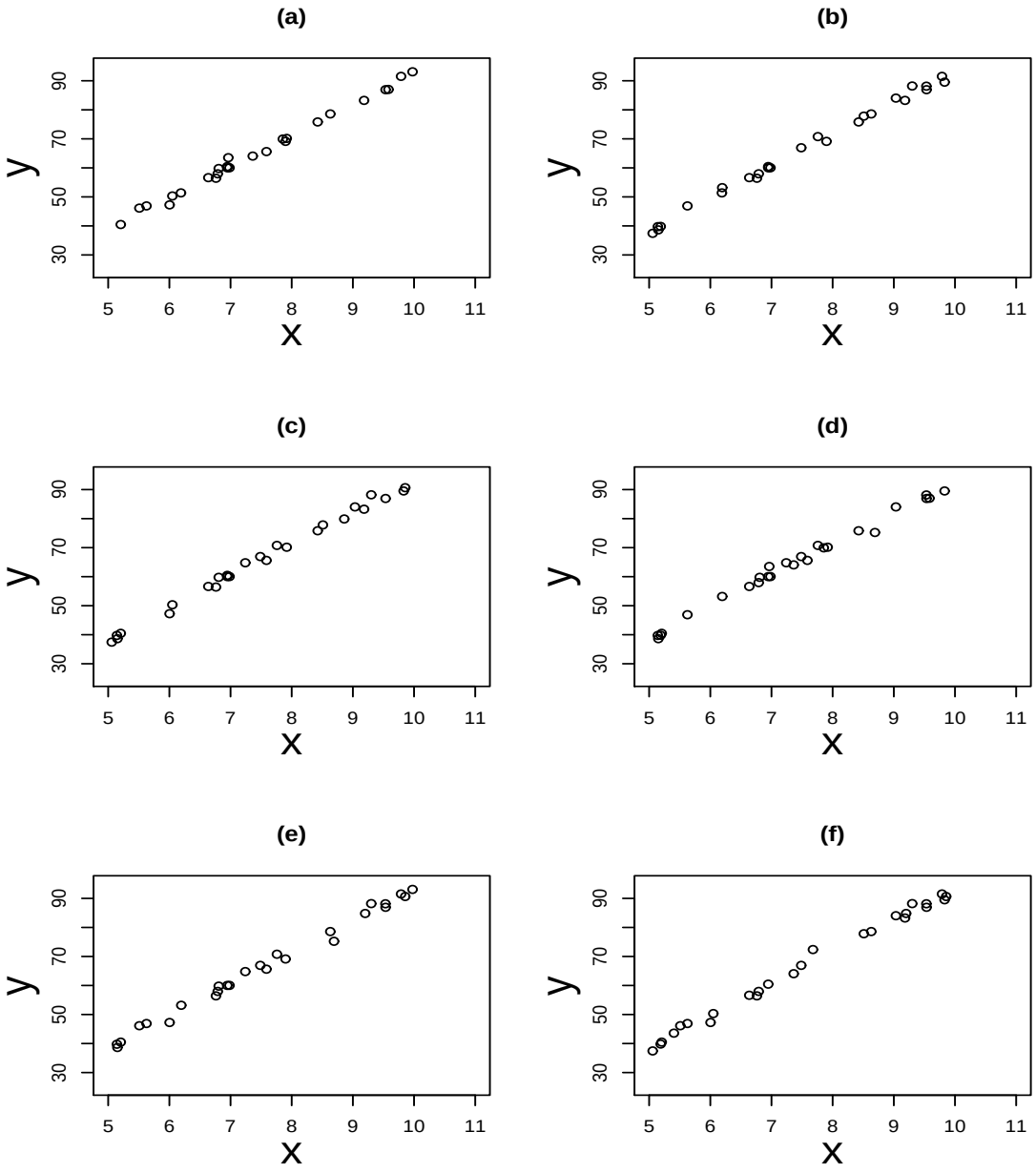

Figure 3.6: The scatter plots of $r^* = 6$ selected good subsamples.

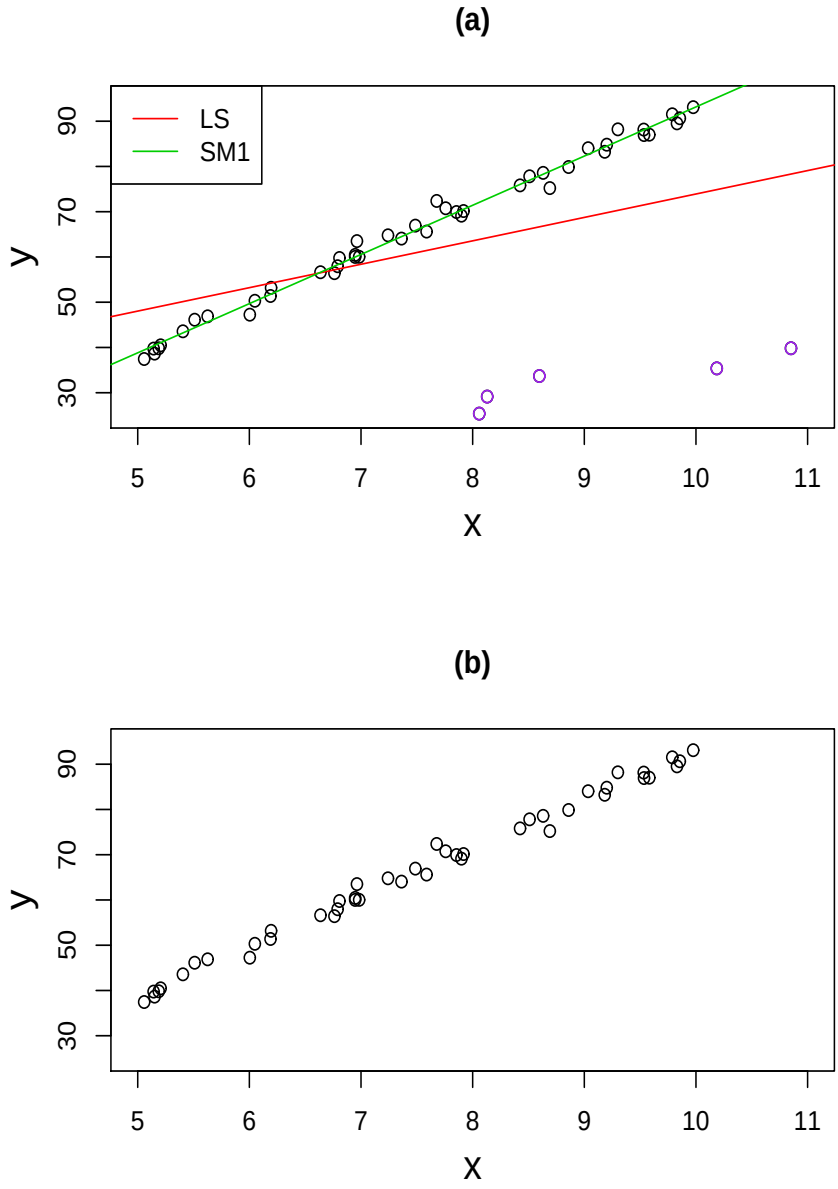

Figure 3.7: (a) Scatter plot of the $N = 50$ observations, the $m = 5$ outliers are show in purple. The red line is the L-S line. The green line is subsampling line. (b) The union of 6 good subsamples selected in the procedure of proposal I for the model (3.15).

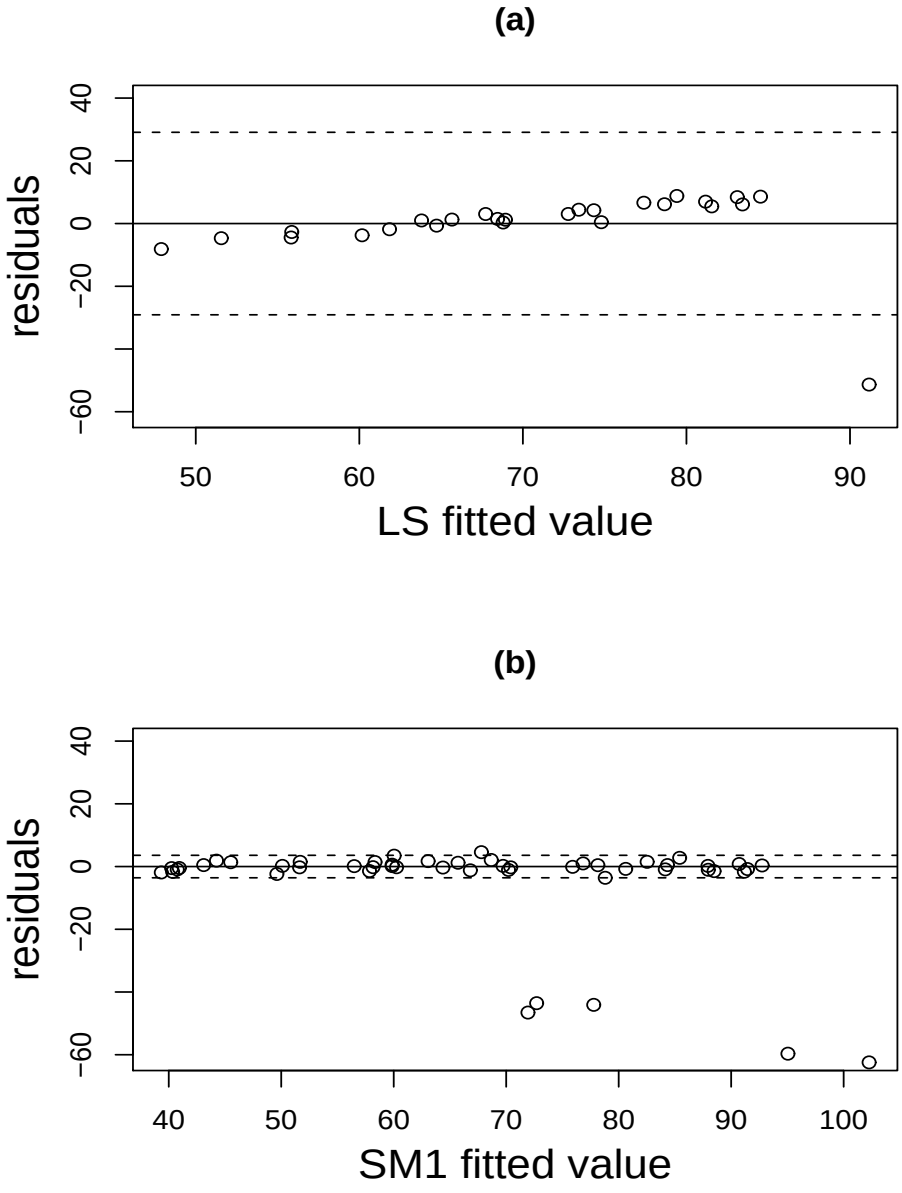

Figure 3.8: Residual plots with $\pm 2.5\hat{\sigma}$ lines (in red): (a) LS residuals. (b) SM1 residuals.

Example 3.3. *Consider $N = 50$ random observations (x_i, y_i), where 45 of these are from a multiple linear regression model of interest (3.17) and the remaining 5 are from a "contaminating" model (3.18).*

$$\begin{aligned} y_i &= -15.7 + 16.87x_{1i} + 19x_{2i} + 11x_{3i} + \varepsilon_i, \quad i = 1, 2, \ldots, 45, \\ &\quad \varepsilon_i \sim N(0, 4), \end{aligned} \tag{3.17}$$

$$\begin{aligned} y_i &= -15.7 + 16.87x_{1i} + \frac{19}{2}x_{2i} + 22x_{3i} + \varepsilon_i, \quad i = 46, 47, \ldots, 50, \\ &\quad \varepsilon_i \sim N(0, 4), \end{aligned} \tag{3.18}$$

with x_{1i} generated by uniform distribution on $[5, 10]$, x_{2i} on $[1, 5]$, and x_{3i} on $[3, 8]$ in (3.17) and $[10, 12]$ in (3.18) Our goal is to estimate the parameters of model of interest (3.17).

Since this sample size N (50) and the percentage of outliers (10%) are the same as the sample in Example 1, the input vector $(N, m, n, n_s, E_F(S_g), p^*)$ is the same as before, that is $(50, 5, 45, 26, 99\%, 0.99)$. Hence the value of (n_s, r^*, k) is also $(26, 6, 650)$. We ran the SA1(n_s, r^*, k) with this parameter setting and obtained the combined sample S_g containing 44 good data points. The algorithm efficiency for this case is 44/45 or roughly 97%. Table 3.3 contains the subsampling and simple least-squares estimates for β_i and σ for model (3.17). As in Example 1, we see that the subsampling estimates outperform the least-squares estimates. For this example, the data points are 4-dimensional and thus cannot be easily displayed. We only plotted the residuals using the fitted models given by the two methods in Figure 3.9. We see that the $\pm 2.5\hat{\sigma}$ for the least-squares line is much wider than that of the subsampling line. Like the previous example, in this case the 5 outliers have been successfully excluded from

the combined sample S_g and thus had no impact on the subsampling estimates. But the simple least-squares method has failed because of the outliers.

Example 3.3 is a case with multivariate data for which outlier detection becomes difficult. The outliers in this case are outliers with respect to the unknown true model. They are not easily characterized as extremes point in some geometric sense. Conventional methods of outlier detection which rely on the geometry or the distribution of the data points are not effective. The subsampling method which uses model based outlier detection has been found very effective in this case. □

Table 3.3: Estimates of 50 data generated by the model (3.17) with s.e. in brackets

	True value (s.e.)	LS estimates (s.e.)	SM1 (s.e.)
$\hat{\beta}_0$	-15.7	-658.8700 (180.9200)	-16.8890 (1.5454)
$\hat{\beta}_1$	16.87	20.9400 (20.5000)	16.7618 (0.1576)
$\hat{\beta}_2$	19	30.1500 (25.3600)	19.0281 (0.1877)
$\hat{\beta}_3$	11	122.7000 (14.7600)	11.3504 (0.1615)
$\hat{\sigma}$	2	214.2000 (207..536)	1.4900 (1.3124)

To demonstrate the versatility of the subsampling method, we now apply it for robust estimation of a logistic regression model.

Example 3.4. *Consider logistic regression model (3.19) for binomial variables* $Y_i \sim BIN(n_i, \pi_i)$ *for* $i = 1, 2, \ldots, 50$*:*

$$\log \frac{\pi_i}{1-\pi_i} = -19.95 + 0.348 x_i. \tag{3.19}$$

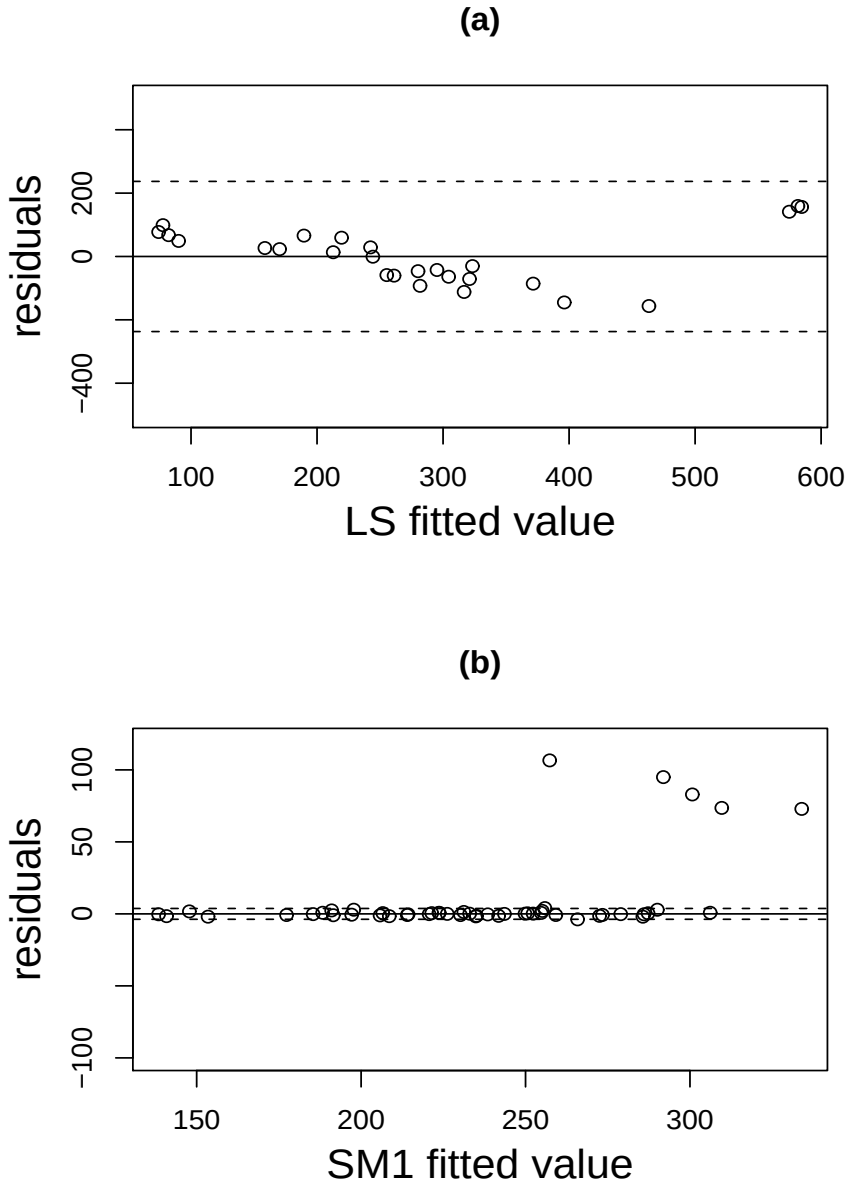

Figure 3.9: Residual plots with $\pm 2.5\hat{\sigma}$ dashed lines: (a) LS residuals. (b) SM1 residuals.

To generate random observations from Y_i, suppose $n_i \sim \text{Unif}(30, 50)$ and $x_i \sim \text{Unif}(50, 70)$. As before, we wanted to create a sample of size $N = 50$ with $m = 5$ outliers. We first generated each x_i value and computed its corresponding π_i using model (3.19) for $i = 1, 2, \ldots, 50$. We then replaced the 5 largest π_i values by 0.1. These largest 5 values are above 0.8 and hence their replacement by 0.1 creates 5 outliers in π_i. With this new set of π_i (containing the 5 outliers), we then generated for each i an n_i from Unif(30,50) and finally a y_i from $\text{BIN}(n_i, \pi_i)$. Hence the 5 y_i values corresponding to the five largest x_i values are outliers, much smaller than what they would be had the corresponding π_i values were not changed.

To apply subsampling method to fit the regression model to the sample of 50 generated above, first note that since the sample size N and the number of outliers m are the same as the two previous examples, the parameter values of the subsampling algorithm $\text{SA1}(n_s, r^*, k)$ are the same as before, *i.e.,* $\text{S}(n_s, r^*, k) = (26, 6, 650)$. The maximum likelihood method is the natural choice for Π and the goodness-of-fit criterion Γ is chosen to be the deviance $D(\mathbf{y})$ given by (here $\mathbf{y}$ denotes the data vector)

$$D(y) = -2\{\log[l(\mathbf{y})|\hat{\boldsymbol{\beta}})] - \log[l(\mathbf{y})|\hat{\boldsymbol{\beta}}_{max})]\},$$

where $\hat{\boldsymbol{\beta}}$ denotes the estimated parameter values in the (reduced) model of interest and $\hat{\boldsymbol{\beta}}_{max}$ denotes the fitted parameter values for the "full model".

Figure 3.10 shows the 650 deviances resulted from fitting the logistic regression model using the maximum likelihood method to 650 subsamples chosen randomly without replacement from the sample of 50 data points with 5 outliers. There are 12 points at the bottom of the plot, representing 12 subsamples containing no outliers. We take the 6 subsamples with the smallest deviances to form the combined sample S_g. The combined sample contains 45 distinct data points. So the observed efficiency is $45/45 = 100\%$. As has been observed in the above two examples, with the probability

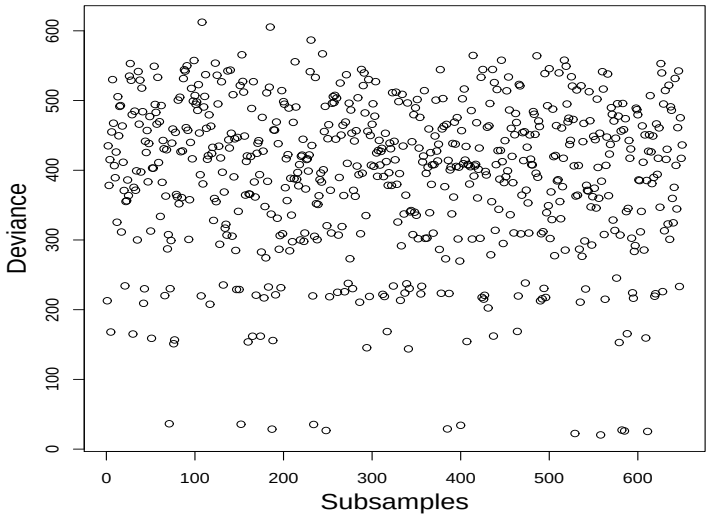

Figure 3.10: Deviances of $k = 650$ subsamples.

of having at least r^* good subsamples p^* set to a high value of 0.99, we usually obtain more than r^* good subsamples but only r^* of which are taken union of. In the present case $r^* = 6$ but we obtained 12. Step 5^* may be modified to take advantage of these extra good subsamples to further increase the observed efficiency. We will revisit this point in the last section of this chapter.

In Figure 3.11, we plotted the 6 selected good subsamples. The entire sample of 50 observations are plotted in Figure 3.12(a). We see that the 6 good subsamples contain no outliers. This is also seen in plot of the combined sample S_g in Figure 3.12(b) which contains no outliers.

Finally, we apply the maximum likelihood method to estimate the logistic regression model parameters. Table 3.4 contains the estimated values and the associated standard deviations. The subsampling estimates are based on only data points in S_g and are in the SM1 column. Those given by the original maximum likelihood method applied to the entire sample of 50 are shown in the MLE column. Comparing to the MLE, the SA1 estimates are much closer to the true values. Furthermore, the

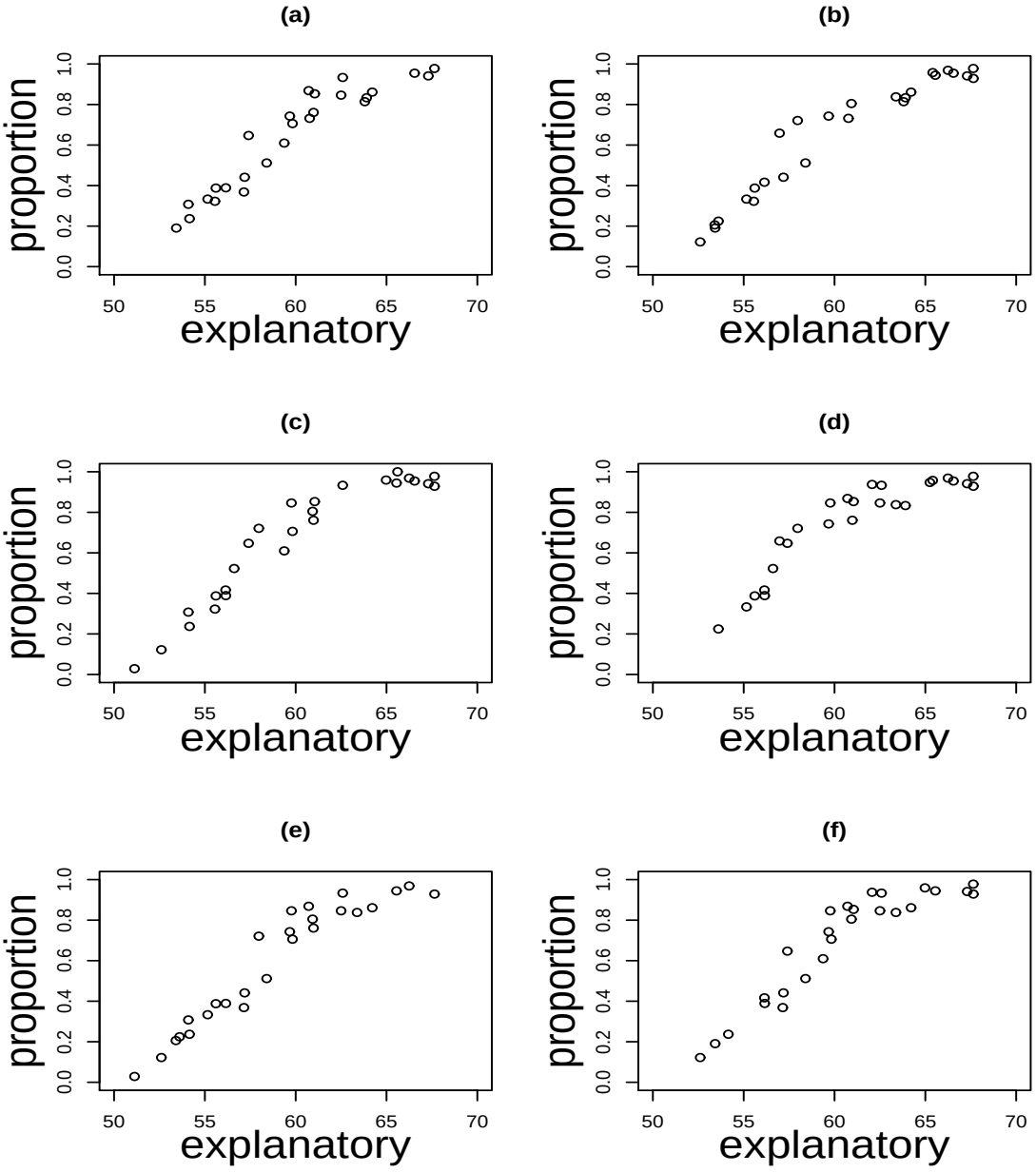

Figure 3.11: (a) Scatter plot of sample of 50; purple points are outliers. (b) scatter plots of five selected good subsamples.

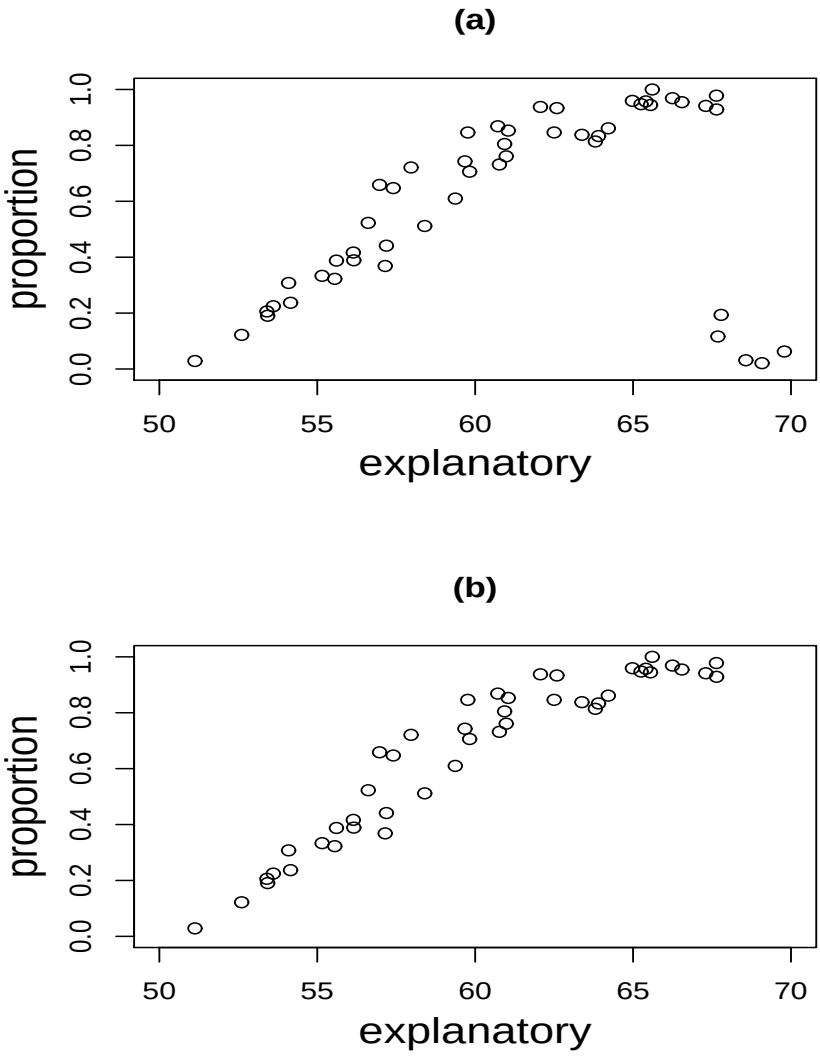

Figure 3.12: (a) Scatter plot of sample of 50. (b) The union plot of 6 good subsamples selected in the procedure of proposal I for the model (3.19).

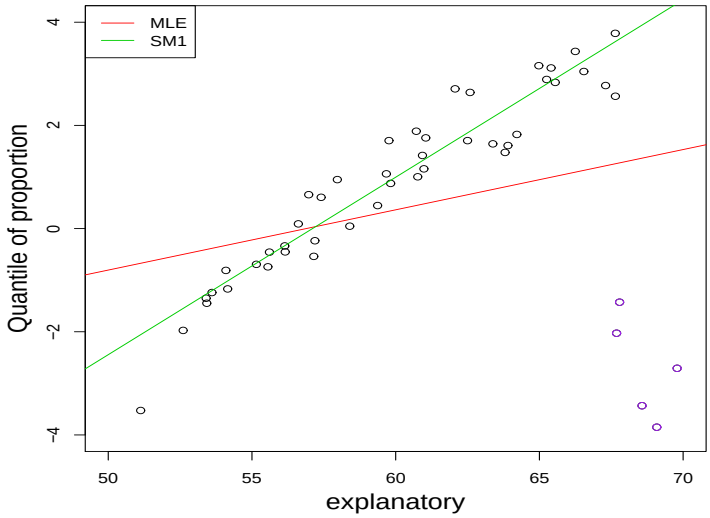

Figure 3.13: The fitted lines for the model (3.19): maximum likelihood method fitted lin in red and subsampling method fitted line in green.

AIC of the subsampling fit is much smaller than the MLE. The good fit given by the subsampling method can also be seen in the plot of the fitted lines in Figure 3.13. The subsampling line (in green) fits the sample much better. The MLE line (in red) is pulled down by the outliers on the right and fits poorly. □

3.3 Comparison with other methods

We described the implementation and application of the subsampling method in the last section. In the present section, we report findings of a simulation study on comparing our subsampling method with other robust methods using Examples 3.3 and 3.4 above. To get a sense on the absolute performance of the robust methods being compared, we also included in our simulation study the "optimal method" representing the method of maximum likelihood applied to the good data only. Such an "optimal method" is not practical but here it does provide a meaningful yardstick

for measuring the absolute performance of the robust methods.

Table 3.4: Estimates of 50 data generated by the model (3.19) with s.e. in brackets

	True value	MLE	SM1
$\hat{\beta}_0$	-19.95	-6.645 (0.596)	-19.644 (0.995)
$\hat{\beta}_1$	0.348	0.117 (0.009)	0.344 (0.017)
AIC	-	967.78	209.63

We considered the six combinations of the total sample size N, number of outliers m and number of good data points n shown in the Table 3.5. These represent different sample sizes (moderate $N = 30$ and moderately large $N = 50$) and degrees of contaminations (none $m = 0$, small $m = 3$, moderate $m = 5$ and moderately heavy $m = 8$). These use of these six combination is also convenient as the corresponding

Table 3.5: Six combinations of sample size and contamination (N, m, n).

(N, m, n)		Degree of Contamination		
		None	Small	moderate
Sample size n_s	Moderate	$(30, 0, 30)$	$(30, 3, 27)$	$(30, 5, 25)$
	Moderately large	$(50, 0, 50)$	$(50, 5, 45)$	$(50, 8, 42)$

parameter values for the subsampling algorithm SA1(n_s, r^*, k) are readily available from Table 3.1. These are used in our computation for the subsequent discussions.

Example 3.3 Continued: For robust estimation of the multiple linear model in Example 3.3, we compared our subsampling method (SM1) with the MM method of (Yohai, 1987) (applied to the entire contaminated sample of N) and the optimal least-squares method (applied to only the good data set of $N - m$; hence it is denoted as LS-). The latter establishes a performance reference point as it is optimal when

there are no outliers. At each combination of (N, m, n), we randomly generate 1000 contaminated samples of size N with m outliers as described in Example 3.3. We then estimated the model parameters using all three methods for each contaminated sample and obtained for each method 1000 estimates of the parameters. With these 1000 estimates, we computed the absolute bias (which is the difference between the average of the 1000 estimates and the true value) and the estimated standard error. Tables 3.6 and 3.7 contain the biases and standard errors of all three methods.

In Tables 3.6 and 3.7, we can see that even though the sample sizes and the number of outliers are different, all estimates are very accurate. The biases for all three methods are small in both absolute and relative terms. The bias of our subsampling estimate is slightly but consistently smaller than that of the MM method. The standard error of the subsampling estimate is comparable to that of the MM estimate. All considered, the subsampling method is competitive to the MM method and both robust methods are accurate (in terms of the bias) and efficient (in terms of the standard error) when compared to the optimal LS- method as their biases and standard errors are comparable to that of the LS-.

We now consider inference based on the subsampling method. We note that the subsampling estimates are inherently nonlinear functions of the data and there is no analytic expression for the subsampling estimators. So the exact distribution theory for the subsampling estimators is difficult to obtain. Fortunately, as we have observed before that when we set the efficiency to a high level of say 99%, the subsampling algorithm can recover the good data effectively. Hence we may view the combined sample S_g as the entire set of good data, and consequently inference methods associated with the non-robust method Π (which may be approximate methods anyway based on some asymptotic theory) may be applied to the subsampling estimates. Table 3.8 summarizes the observed efficiencies for 1000 simulations of all six combinations of

Table 3.6: The estimates of SM1, MM and LS- when $N = 30$

		$m = 0$			$m = 3$			$m = 5$		
		SM1	MM	LS-	SM1	MM	LS-	SM1	MM	LS-
$\hat{\beta}_0$	average	-15.6452	-15.6395	-15.6519	-15.7356	-15.7363	-15.6766	-15.7227	-15.7289	-15.7204
	bias	0.0548	0.0605	0.0481	0.0356	0.0363	0.0234	0.0227	0.0289	0.0204
	s.e.	3.1778	2.8662	2.7378	2.9189	2.8485	2.7052	3.1703	3.1397	3.1053
$\hat{\beta}_1$	average	16.8870	16.8931	16.8762	16.8632	16.8554	16.8694	16.8745	16.8766	16.8737
	bias	0.0170	0.0231	0.0062	0.0068	0.0146	0.0006	0.0045	0.0066	0.0037
	s.e.	0.3123	0.2815	0.2742	0.2976	0.2830	0.2679	0.3179	0.3111	0.3078
$\hat{\beta}_2$	average	18.9841	18.9803	18.9864	19.0020	19.0024	18.9986	18.9968	18.9952	18.9987
	bias	0.0159	0.0197	0.0136	0.0020	0.0024	0.0014	0.0032	0.0048	0.0013
	s.e.	0.3899	0.3458	0.3270	0.3613	0.3624	0.3376	0.3754	0.3739	0.3692
$\hat{\beta}_3$	average	10.9883	10.9881	10.9894	11.0165	11.0247	11.0125	10.9969	10.9953	10.9971
	bias	0.0117	0.0119	0.0106	0.0165	0.0247	0.0125	0.0031	0.0047	0.0029
	s.e.	0.3213	0.2900	0.2727	0.2979	0.2931	0.2747	0.3060	0.3035	0.3030

Table 3.7: The estimates of SM1, MM and LS- when $N = 50$

		$m = 0$			$m = 5$			$m = 8$		
		SM1	MM	LS-	SM1	MM	LS-	SM1	MM	LS-
$\hat{\beta}_0$	average	-15.7335	-15.7716	-15.7131	-15.6564	-15.6233	-15.6641	-15.7196	-15.6734	-15.6819
	bias	0.0335	0.0716	0.0131	0.0546	0.0767	0.0359	0.0196	0.0266	0.0181
	s.e.	2.3337	2.3231	2.1035	2.3292	2.3271	2.2418	2.3322	2.3340	2.3155
$\hat{\beta}_1$	average	16.8730	16.8741	16.8693	16.8708	16.8691	16.8695	16.8734	16.8746	16.8723
	bias	0.0030	0.0041	0.0007	0.0008	0.0009	0.0005	0.0034	0.0046	0.0023
	s.e.	0.2238	0.2228	0.2056	0.2230	0.2254	0.2157	0.2247	0.2244	0.2229
$\hat{\beta}_2$	average	19.0039	19.0063	19.0002	18.9909	18.9805	18.9914	19.0044	19.0064	19.0026
	bias	0.0039	0.0063	0.0002	0.0091	0.0195	0.0086	0.0044	0.0064	0.0026
	s.e.	0.2796	0.2834	0.2512	0.2904	0.2912	0.2792	0.2816	0.2799	0.2769
$\hat{\beta}_3$	average	10.9983	11.0033	11.0020	10.9943	10.9897	10.9956	10.9934	10.9931	10.9930
	bias	0.0017	0.0033	0.0020	0.0057	0.0103	0.0044	0.0066	0.0069	0.0054
	s.e.	0.2307	0.2339	0.2105	0.2285	0.2223	0.2163	0.2171	0.2166	0.2151

Table 3.8: The observed efficiency of the multiple regression model

	N=30			N=50		
	m=0	m=3	m=5	m=0	m=5	m=8
average size of S_g	28.239	26.378	24.686	48.265	43.886	41.367
observed efficiency	94.13%	97.69%	98.74%	96.53%	97.52%	98.49%

(N, m, n). The nominal efficiency $E_F(S_g)$ was set to 99% for all the simulations. The average observed/achieved efficiency ranges from 94% to 98%. So the recovery rate of the good data points is quite good. This supports our statement that existing inference methods associated with Π may be applied to the subsampling estimators.

In this particular example of multiple linear regression model, the non-robust method Π we used is the least-squares method whose estimators and MSEs have well-known properties. To look for further evidence supporting the statement, we need to demonstrate for this example that the subsampling method shares such properties of the least-squares method. To this end, we need to examine (1) the distribution of the MSE of the subsampling fit based on S_g (that of the least-squares method has a χ^2 distribution) and (2) the distribution of subsampling estimators of the model parameters (that of the least-square method is asymptotically normal). We now consider these two points for the case of $(N, m, n) = (50, 5, 45)$.

To examine point (1), in Figure 3.14(a) we plotted the histogram of the 1000 χ^2 statistics $(n-4)MSE/4$ obtained after fitting the least-squares line to the combined sample S_g. Although the size of these 1000 combined samples S_g vary slightly depending on the observed efficiency, the resulting χ^2 statistics do have a χ^2-like histogram. Figure 3.14(b) shows the QQ-plot of the χ^2 statistics using a theoretical degrees of freedom of $n - p = 45 - 4 = 41$ and it is rather straight indicating good overall agreement with the χ^2 distribution with degrees of freedom 41.

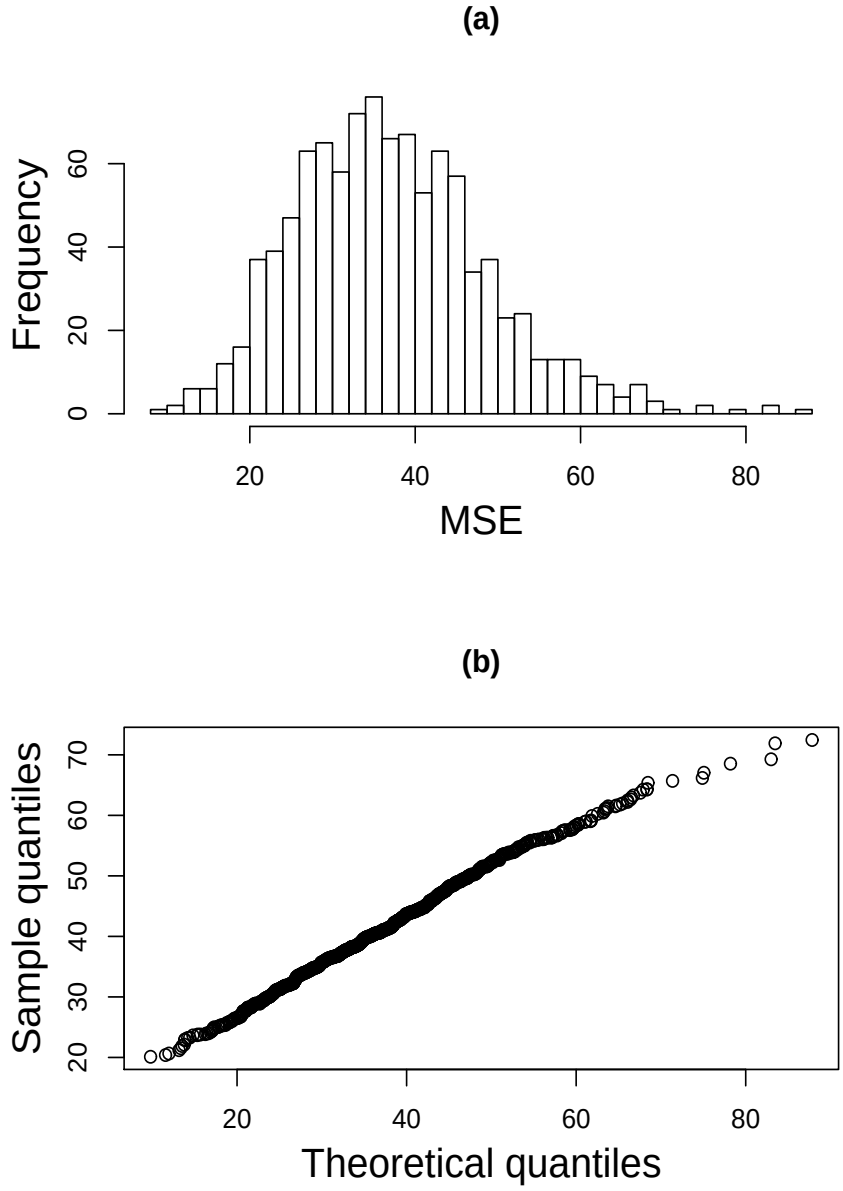

Figure 3.14: (a) the histogram of chisquare statistic $\frac{(n-4)\text{MSE}}{4}$. (b) the QQ - plot of the chisquare statistics.

To examine point (2), we need to look at the distribution associated with the subsampling estimator for β_i and to ensure it is consistent with that for the simple least-squares method. It is well-known that if the model errors ε_i are independently and normally distributed with mean 0 and variance σ^2, the sampling distribution of $(\hat{\beta}_i - \beta_i)/s.e.(\hat{\beta}_i)$ is a t distribution with $n - p$ degrees of freedom, where $\hat{\beta}_i$ is the simple least-squares estimator based on the n good data. Does the subsampling estimator for β_i also have this property? We explore this question numerically by plotting such a t ratio for the 1000 subsampling estimates for the case of $(N, m, n) = (50, 5, 45)$. First note that the degrees of freedom for t ratio of the least-squares estimator is $n - p = 41$, which is quite large. Hence the histograms of the t ratios for the least-squares estimators will look like normal distribution histograms. Figure 3.15 shows the histograms for the t ratios of subsampling estimators for the four regression coefficients. They also look like normal distribution histograms. Further, Figure 3.16 shows normal QQ-plot of the t ratios for the subsampling estimators. They all follow straight lines. Hence, based on numerical exams of the MSEs and estimators, the sampling method behaves like the simple least-squares method. This is not surprising given the high efficiency that the subsampling algorithm has achieved. Indeed, at the 100%, the subsampling method is identical to the least-squares applied to the good data. Provided the efficiency is not lower than 90%, empirically the subsampling method is expected to behave like the simple least-squares method on a good sample.

Having justified the use of inference methods for least-squares estimators, we can now construct subsampling confidence intervals for the model parameters and compare them with confidence intervals given by other methods. Table 3.9 reports the 90% confidence intervals given by the subsampling method, the MM method and the optimal least-squares method applied to the good data. The subsampling confidence interval (SM1) are just least-squares confidence intervals on the combined

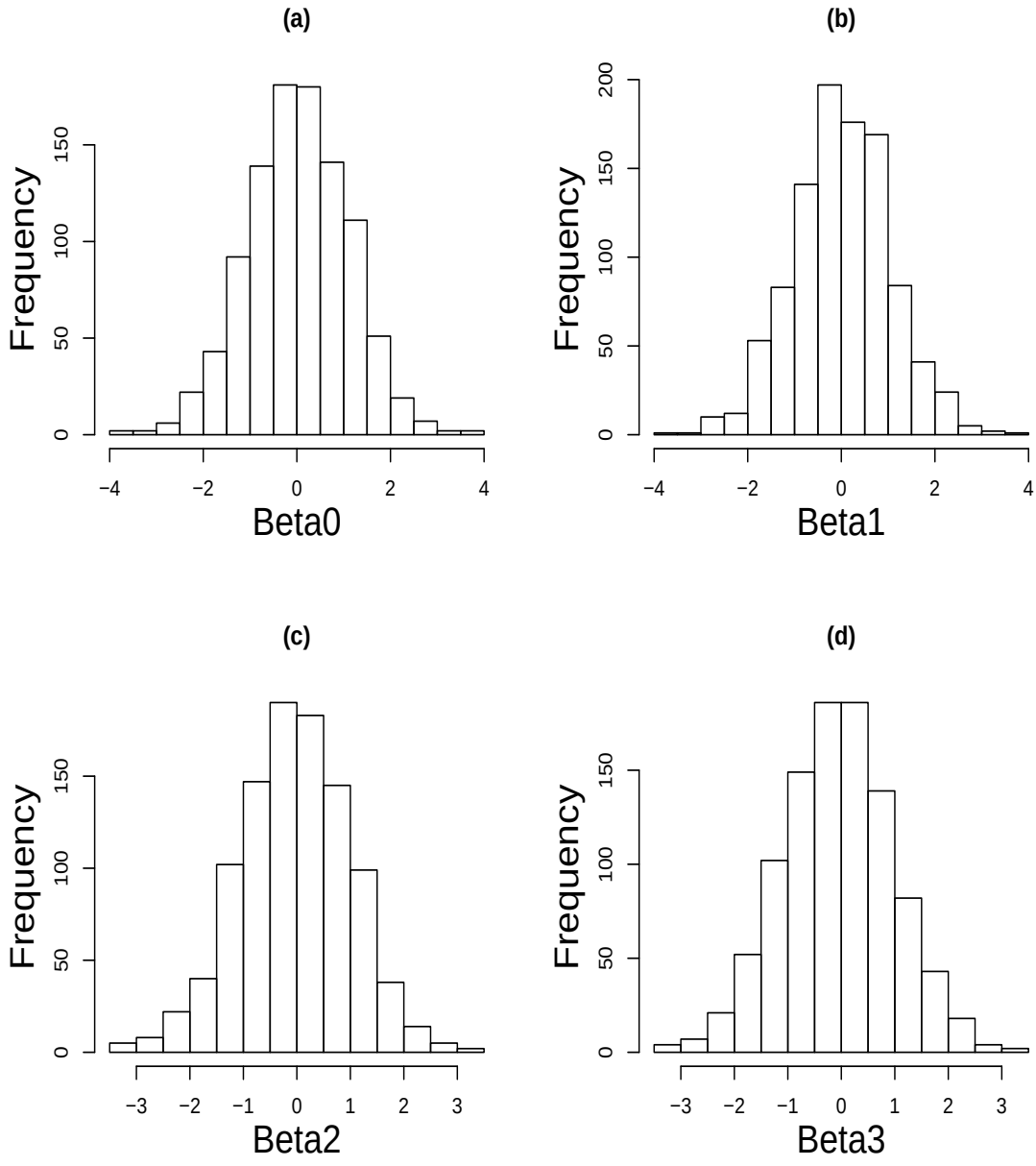

Figure 3.15: Histograms of 1000 subsampling t ratios $(\hat{\beta}_i - \beta_i)/s.e.(\hat{\beta}_i)$, $i = 0, 1, 2, 3$.

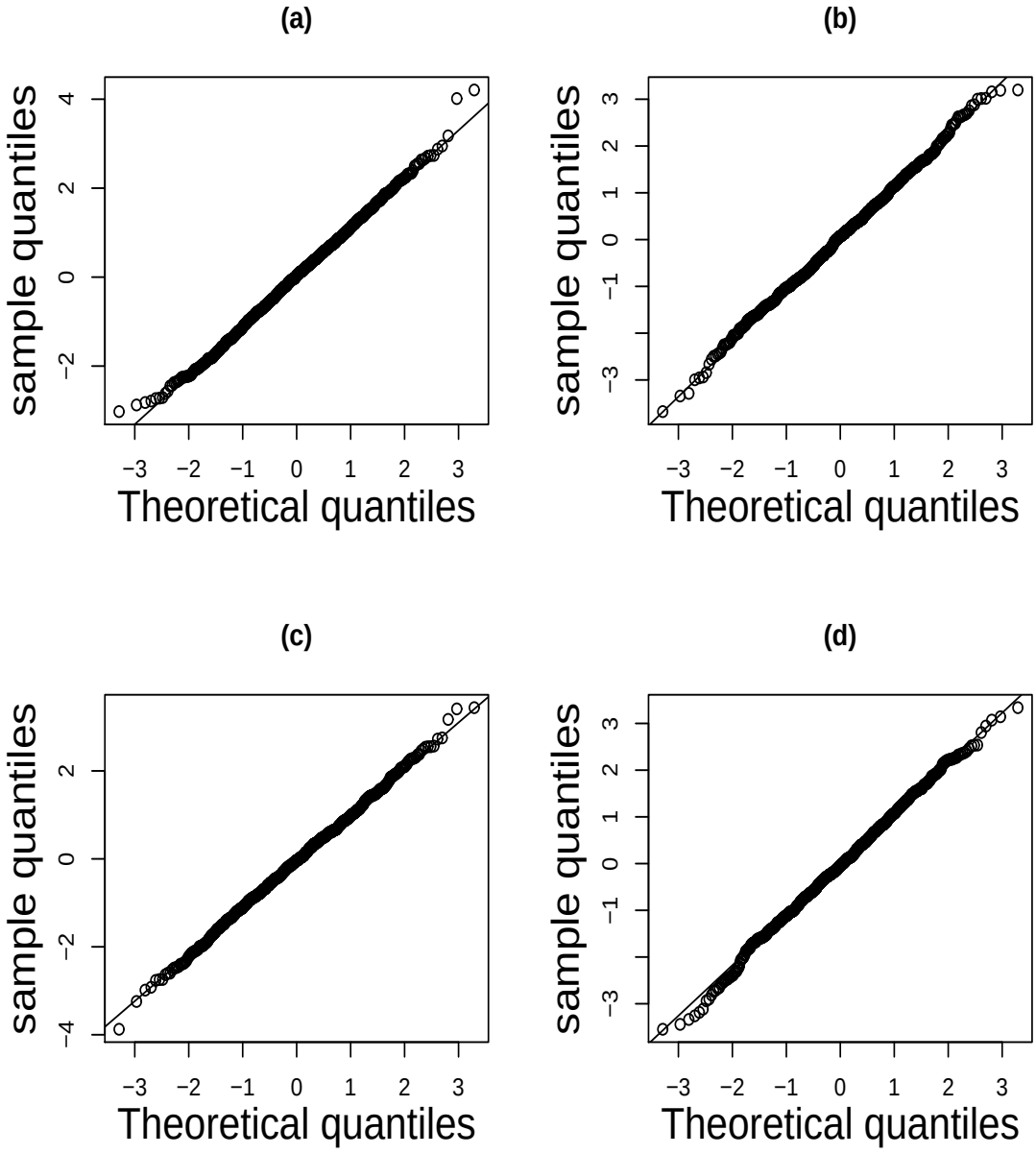

Figure 3.16: Normal QQ-plots for 1000 subsampling t ratios $(\hat{\beta}_i - \beta_i)/s.e.(\hat{\beta}_i)$, $i = 0, 1, 2, 3$.

sample S_g. Each "coverage" entry in the table is the percentage of intervals out of the 1000 generated that contain the true value. Each "average length" is just the average length of the 1000 intervals. Overall, the performance of the subsampling confidence intervals are quite close that of the optimal method LS-; the coverage level is slightly lower than that of the LS- but the average length is also a bit shorter. Comparing to the MM method, the subsampling intervals have a consistently more accurate coverage rate (more accurate by about 1%) but the average length of the subsampling intervals is also above 1% longer than that of the MM intervals. Hence the subsampling method is quite competitive to the MM method in terms of confidence intervals for model parameters. □

Example 3.4 Continued: We now compare the performance of our subsampling method on the logistic regression model (3.19) in Example 3.4 to that of an M-estimation based method by Cantoni and Ronchetti (2001). For computation of the latter method, we have used program *glmRob* in R. The "optimal method" included in this comparison is the maximum likelihood method applied to the good data only (hence it is denoted by MLE-).

Table 3.10 gives the observed efficiency based on 1000 simulation at each combinations (N, m, n); each average size of S_g is that of 1000 simulated S_g's and the observed efficiency is just this average divided by n. The underlying nominal efficiency level is 99% for all cases. The table shows that the observed efficiencies are all above 94%. For those columns with outliers ($m = 3, 5, 8$), the observed efficiencies are higher than 97%, which lend strong support to the use of simple maximum likelihood inference for the subsampling estimators. For cases where there are no outliers, the efficiencies turned out to be a bit lower. This is partly due to the small number of subsample k that we have used for cases of $m = 0$ (see table 3.1). In practice, we always assume a default percentage of outliers of 10% which will lead to a much larger k value. Hence

Table 3.9: The summary of 90% confidence intervals for the SM1, MM and LS- estimates

			SM1	MM	LS-	SM1	MM	LS-	SM1	MM	LS-
N=30			m=0			m=3			m=5		
	$\hat{\beta}_0$	coverage	88.7%	87.3%	89.1%	87.5%	86.2%	88.2%	88.9%	86.1%	90%
		average length	9.1781	9.3861	9.1621	9.3018	9.2364	9.7565	10.1504	9.5629	10.2401
	$\hat{\beta}_1$	coverage	91.6%	88.7%	91.9%	87.6%	86.1%	88.2%	88.1%	85.9%	89.9%
		average length	0.9131	0.9351	0.9111	0.9162	0.8967	0.9602	1.0157	0.9704	1.0244
	$\hat{\beta}_2$	coverage	91.7%	89%	91.7%	87.5%	86.3%	88.1%	89.8%	87%	90.9%
		average length	1.1396	1.1682	1.1370	1.1567	1.1416	1.2133	1.2566	1.1967	1.2681
	$\hat{\beta}_3$	coverage	89.3%	87.3%	89.1%	86.8%	85.3%	88%	89.1%	87.8%	89.9%
		average length	0.9065	0.9339	0.9044	0.9249	0.9050	0.9691	1.0041	0.9541	1.0137
N=50			m=0			m=5			m=8		
	$\hat{\beta}_0$	coverage	91.8%	90%	91.7%	88.9%	87.4%	90.2%	88.8%	87.4%	89.5%
		average length	7.1334	7.0045	6.86073	7.170538	7.141036	7.228027	7.55497	7.377988	7.603617
	$\hat{\beta}_1$	coverage	89.7%	89%	89.8%	87.7%	86.2%	89.2%	89.4%	88%	89.5%
		average length	0.7072	0.0.6955	0.6801	0.7128298	0.7120258	0.7183204	0.7492251	0.7326589	0.7541819
	$\hat{\beta}_2$	coverage	90.5%	90.1%	90.8%	88.6%	88.5%	89.4%	88.7%	87.6%	89.2%
		average length	0.8838	0.8672	0.8508	0.8888926	0.8843975	0.8958066	0.9294513	0.9089452	0.9358113
	$\hat{\beta}_3$	coverage	91.1%	91%	90.8%	88.5%	88.4%	89.3%	90.7%	89.5%	91.2%
		average length	0.7058	0.6960	0.6792	0.707966	0.7029538	0.7133879	0.7450151	0.7242658	0.7502068

Table 3.10: The observed efficiency of the logistic regression model

	N=30			N=50		
	m=0	m=3	m=5	m=0	m=5	m=8
average size of S_g	28.372	26.238	24.875	48.055	43.923	41.401
observed efficiency	94.57%	97.17%	99.5%	96.11%	97.6%	98.57%

the observed efficiency for the no outliers should also improve.

Table 3.11 compares the subsampling estimates (SM1) with the M-estimates (M) and the "optimal" MLE- estimates. We note that the observations made in comparisons for the multiple regression model also apply here: for cases with outliers in the sample, the subsampling estimates enjoy a much smaller bias when compared to the M-estimates but they also tend to have bigger standard error than the latter. For cases without outliers, the two robust methods are compatible. Overall, the subsampling estimates are close to the "optimal" MLE- estimate, again due to the insignificant differences between S_g and the entire good data set because of high efficiency.

In this case, we compare the SM1 estimates with other two methods, a robust M-estimates and maximum likelihood estimates based on the all good data in the original data set (MLE-). The robust M-estimator analyzed in this case is introduced by Cantoni and Ronchetti (2001) and can be simply implemented by R software. In Tables 3.11, we can see that even though the sample sizes and the number of outliers is different, all estimates are consistent. Furthermore, the estimates of the SM1 all have much smaller bias compared with the M-estimates and the bias of the MLE-estimates are the smallest. We still note that all biases are smaller than 0.5 and most of them are smaller than 0.05. Thus, the estimates of all methods are robust and the subsampling method proposal I has better unbiasness property than the M-

estimators. We also can see that most of the standard errors of the SM1 estimates are a little bit larger than those of the M-estimates, as we may drop few good data points in the original data set when we use the subsampling method proposal I in Step 5 and we use the whole original data set in the M-estimation.

Before we construct subsampling confidence intervals for the model parameters, let us again first examine the empirical distribution of the subsampling estimates to ensure that the maximum likelihood based inference applies to the subsampling estimators. For the case of $(N, m, n) = (50, 5, 45)$, Figure 3.17(a) and (b) show, respectively, the histogram and χ^2 QQ-plot of the deviances resulting from 1000 simulations. We see that histogram has a χ^2 shape and the QQ-plot looks like a straight line, both confirming that the deviance of the subsampling fit behaves like that of maximum likelihood fit to an outlier free data set. Figure 3.18 examines the distributions of the subsampling estimators for the logistic regression parameters β_0 and β_1. The histograms of the estimators showing in plots (a) and (b) are normal like, agreeing with the asymptotic distribution of the maximum likelihood theory. The normal QQ-plots of the estimates are given in plots (c) and (d) both are nice straight lines further supporting that the subsampling estimators are normally distributed.

Finally, we compare subsampling confidence intervals based on the normal assumption of the subsampling estimators with that of the other two methods. Table 3.12 contains the coverage and average length of these intervals. The nominal level used is 90%. Here we see that for cases with outliers, the subsampling confidence interval outperforms the M-estimation interval in terms of coverage level by a large margin; the coverage level of the subsampling interval, though lower than the nominal level, is still acceptable but that of the M-estimation interval is too low to be of any value. The poor performance of the latter is most likely due to its relatively large bias *and* small standard deviation. Overall, the subsampling method outperformed

Table 3.11: The estimates of SM1, M and MLE- for the logistic regression model

			SM1	M	MLE-	SM1	M	MLE-	SM1	M	MLE-
			$m=0$			$m=3$			$m=5$		
N=30	$\hat{\beta}_0$	average	-20.0712	-20.0857	-20.0403	-19.5954	-18.7103	-19.9781	-19.9046	-17.2825	-19.9811
		bias	0.1212	0.1357	0.0903	0.3546	1.2397	0.0281	0.0454	2.6675	0.0311
		s.e.	1.3946	1.1961	1.1598	1.2766	1.2810	1.2285	1.5716	1.8835	1.2586
	$\hat{\beta}_1$	average	0.3501	0.3593	0.3496	0.3417	0.326	0.3485	0.3472	0.3004	0.3485
		bias	0.0021	0.0113	0.0016	0.0077	0.022	0.0005	0.0008	0.0476	0.0005
		s.e.	0.024	0.0206	0.02	0.0346	0.0218	0.0204	0.0275	0.0338	0.0218
			$m=0$			$m=5$			$m=8$		
N=50	$\hat{\beta}_0$	average	-19.9354	-19.9774	-19.95	-20.0063	-18.7894	-20.0004	-19.6049	-17.6435	-19.9952
		bias	0.0146	0.0274	0	0.0563	1.1606	0.0504	0.3451	2.3065	0.0452
		s.e.	1.0908	0.9005	0.8709	0.9823	0.9766	0.9472	1.4228	1.1123	1.1123
	$\hat{\beta}_1$	average	0.3485	0.3488	0.348	0.3489	0.3274	0.3481	0.3419	0.3071	0.3489
		bias	0.0005	0.0008	0	0.0009	0.0206	0.0001	0.0079	0.0471	0.0009
		s.e.	0.0187	0.0149	0.0154	0.0169	0.0168	0.0163	0.0292	0.0194	0.017

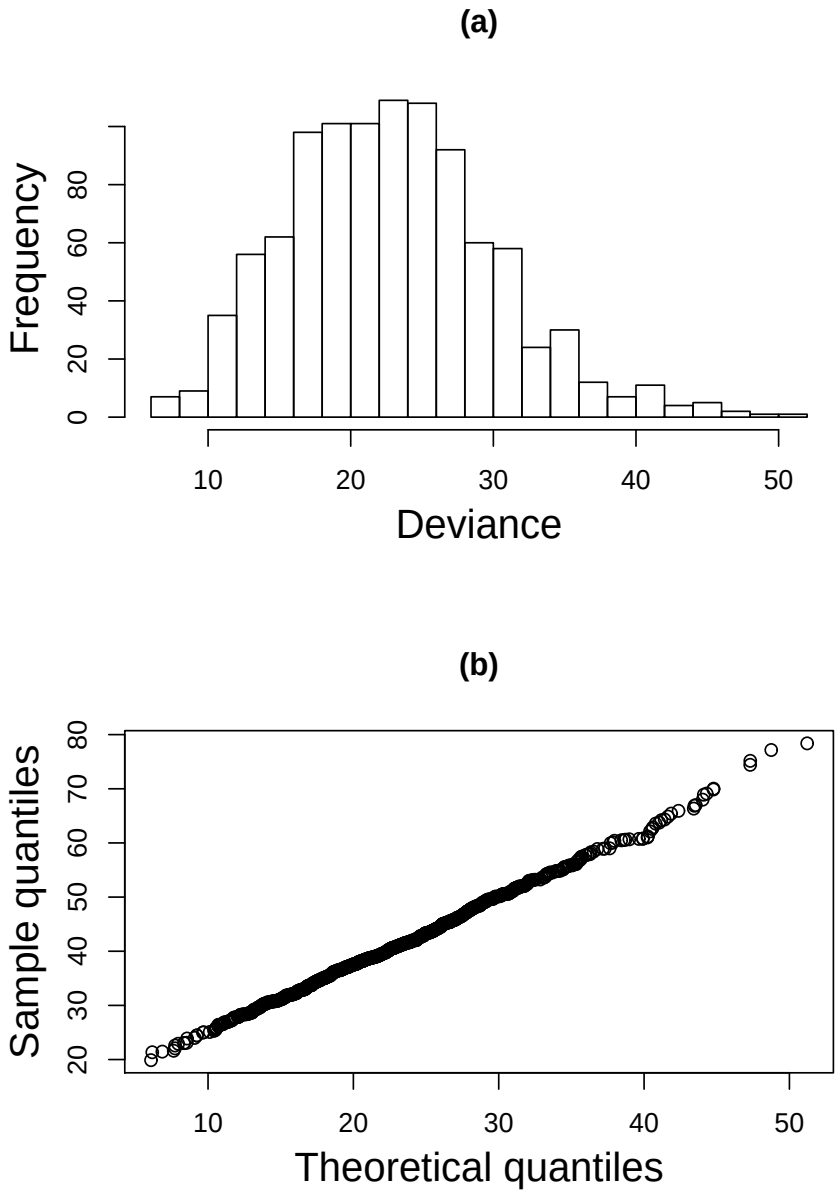

Figure 3.17: (a) the histogram of deviance. (b) the QQ - plot of the deviance.

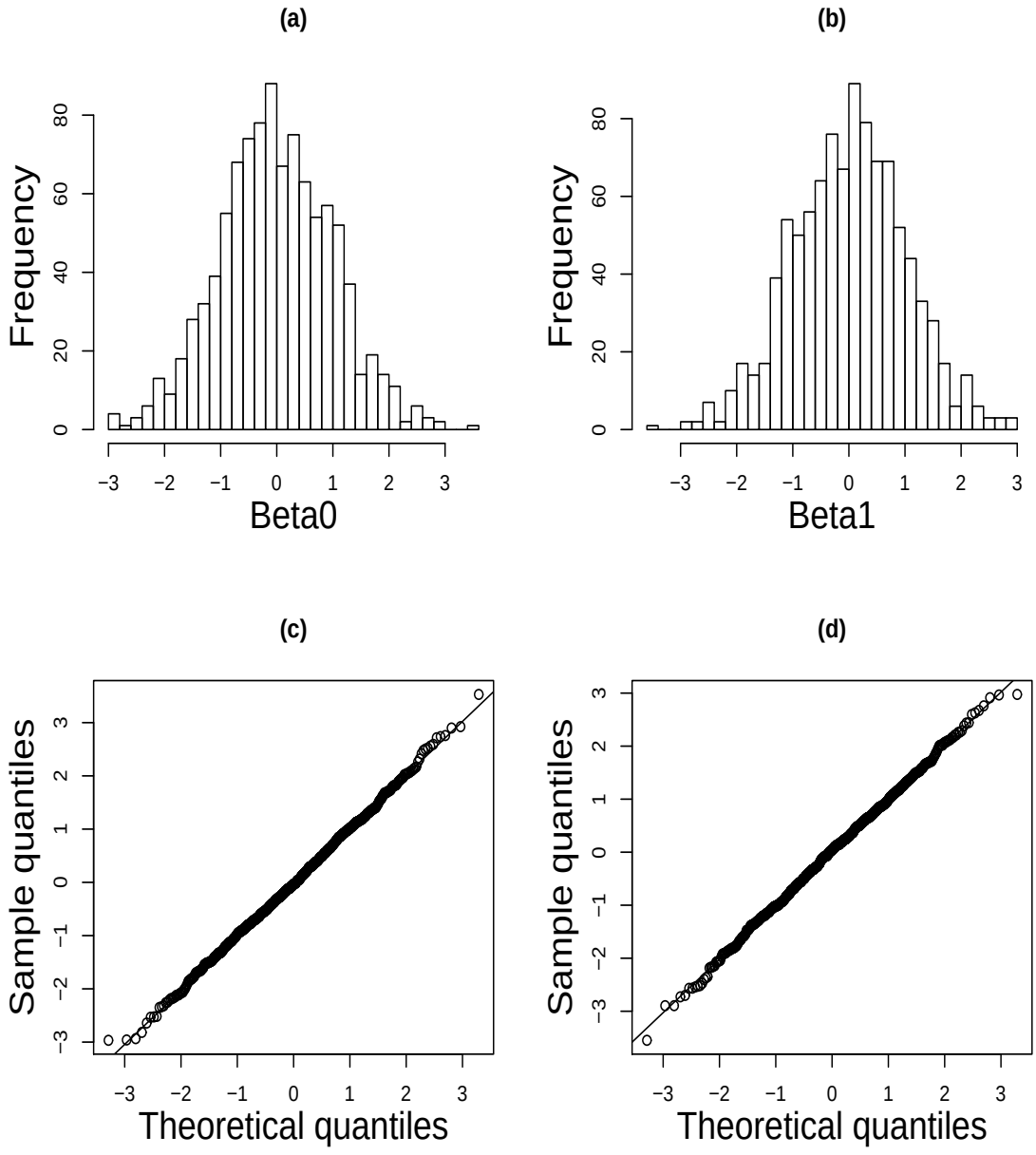

Figure 3.18: (a) and (b) histograms for each estimate, β_0, β_1. (c) and (d) QQ - plots for each estimate testing normality.

the M-estimation method for this example.

3.4 Concluding remarks

We have demonstrated through the applications and comparisons that the subsampling method is a valuable addition to robust methods for regression models. The straightforward manner in which the subsampling method extends common non-robust methods to deal with outliers through the subsampling algorithm makes it a very general method, applicable to all regression problems where well established non-robust methods exist. The advantages of the subsampling method-proposal I are that (1) it can be applied to many regression models and (2) it may be (nearly) unbiased conditioning on that the combined sample is nearly the good data set and that method Π is unbiased.

The subsampling algorithm SA1(n_s, r^*, k) that we presented here is the most basic version of a general subsampling algorithm-I that Dr. Tsao and Dr. Zhou have been working on. Although it worked well for our examples, improvements are necessary and some of which are easily accomplished. For example, noting that in all three examples that we have considered, there were more good subsamples than the $r^* = 6$ that we took union of to form the combined sample S_g, we can improve the efficiency by including more such good subsamples in S_g. To this end, Step 5^* may be used instead of Step 5. For Example 1, if we are to take γ_C to be $2\gamma_1$, then the combined sample S_g will include all 9 good subsamples instead of only 6 which was the number used in Example 1. Further research on ways to improve the basic subsampling algorithm SA1(n_s, r^*, k) is still ongoing. See Chapter 5 for a brief discussion.

Table 3.12: The summary of the 90% confidence intervals for the SM1, M and MLE- estimates

			SM1	M	MLE-	SM1	M	MLE-	SM1	M	MLE-
N=30			m=0			m=3			m=5		
	$\hat{\beta}_0$	coverage	84.6%	83.7%	90.3%	84.3%	66.8%	90.5%	88.5%	53.6%	90.1%
		average length	3.9492	3.9249	3.8042	3.9829	3.7012	4.006	4.1951	3.454	4.1741
	$\hat{\beta}_1$	coverage	84.6%	83%	90%	84.7%	66.2%	91.1%	88.6%	56.8%	89.9%
		average length	0.068	0.0674	0.0645	0.0687	0.0635	0.0691	0.0726	0.0591	0.0723
N=50			m=0			m=5			m=8		
	$\hat{\beta}_0$	coverage	84.6%	84%	90.2%	89.1%	67.6%	89.5%	81.5%	59.9%	89.6%
		average length	3.0549	3.002	2.9176	3.1175	2.8531	3.0777	3.1716	2.7023	3.1929
	$\hat{\beta}_1$	coverage	85.2%	84.7%	90.9%	89.1%	68.7%	89.1%	80.8%	58.1%	89.1%
		average length	0.0526	0.0502	0.0517	0.0034	0.003	0.0033	0.0548	0.0463	0.0552

Chapter 4

Subsampling Method - Proposal II

In Chapter 3, we discussed subsampling method-proposal I for identifying good data points from a contaminated sample and then making inferences based on the identified good data points. This method is robust against model outliers. In this chapter, we discuss another implementation of the subsampling idea: *subsampling method-proposal II* or SM2. The objective of this method is also to identify good data points from the contaminated sample for estimation and inference using the model based subsampling. But it differs from proposal I in how the set of good data points are identified; whereas proposal I finds such a set by taking union of r^* good subsamples S_g, proposal II requires only one good subsample and adds additional good data points to the subsample to form the set of good data points.

The rest of this chapter is organized as follows. In Section 4.1, we present and discuss the algorithm for proposal II. In Section 4.2, we give two examples to illustrate the use of proposal II and to compare it with proposal I, the MM-method and the method of the least squares. We then investigate the influence functions and breakdown points for the two subsampling methods in Section 4.3. In Section 4.4, we investigate the simulation behavior of SM2. We conclude this chapter with some

discussions on the subsampling methods in Section 4.5.

4.1 Subsampling method-proposal II

Subsampling methods proposals I and II have many elements in common. For consistency, we will use the same notation for such elements in our presentation of proposal II below. We assume that there is a sample of $N = n + m$ observations

$$\mathcal{S}_N = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, \boldsymbol{z}_{n+1}, \ldots, \boldsymbol{z}_{n+m}\}$$

from regression model (3.1), where n is the number of good data and m is the number of bad data points (outliers). The objective is to construct a good subsample S_g^2 of $\mathcal{S}_N$ which contains only and as many as possible good data points. To do so, we consider a random subsample of size n_s, S_{n_s}, taken without replacement from $\mathcal{S}_N$. As in proposal I, we assume that $m < n_s \leq n$ to ensure that S_{n_s} cannot be consisted entirely of bad data points and that there exists at least one S_{n_s} containing only good data points. We use some simple non-robust method Π to fit the regression model to this random subsample. Let Γ be a quantitative goodness-of-fit criterion associated Π which may be indicative of the presence of outliers in S_{n_s} and suppose a small γ value means a good fit (absence of outliers). We compute the good subsample S_g^2 for proposal II from S_{n_s} and γ through the following algorithm.

Algorithm SA2(n_s, k): Subsampling algorithm—proposal II

Step 1: Randomly draw a subsample $S_{n_s}^1$ without replacement from the original sample $\mathcal{S}_N = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, \boldsymbol{z}_{n+1}, \ldots, \boldsymbol{z}_{n+m}\}$.

Step 2: Fit the regression model (3.1) to the subsample obtained in Step 1 using method Π and compute the corresponding goodness-of-fit measure γ_1.

Step 3: Repeat Steps 1 and 2 for $j = 1, 2, \ldots, k$ times. Each time record $(S_{n_s}^j, \gamma_j)$, the subsample taken and the associated goodness-of-fit measure at the jth repeat.

Step 4: Sort the k subsamples by the size of their associated γ values; denote by $\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(k)}$ the increasingly ordered values of γ_j, and by $S_{n_s}^{(1)}, S_{n_s}^{(2)}, \ldots, S_{n_s}^{(k)}$ the correspondingly ordered subsamples.

Step 5: Use the method Π to fit the regression model to the subsample $S_{n_s}^{(1)}$-the best subsample in k subsamples.

Step 6: Denote by $\bar{S}_{n_s}^{(1)}$ the complement of $S_{n_s}^{(1)}$ from $\mathcal{S}_N$. So $\bar{S}_{n_s}^{(1)}$ contains $N - n_s$ data points. Use a criterion τ and the fitted model in Step 5 to test each point $\boldsymbol{z_i}$ in $\bar{S}_{n_s}^{(1)}$ to see if it is a good data point. Suppose there are l good points in $\bar{S}_{n_s}^{(1)}$, say, $\boldsymbol{z}_{i_1}, \boldsymbol{z}_{i_2}, \ldots, \boldsymbol{z}_{i_l}$. Then the final good subsample S_g^2 is given by

$$S_g^2 = S_{n_s}^{(1)} \cup \{\boldsymbol{z}_{i_1}, \boldsymbol{z}_{i_2}, \ldots, \boldsymbol{z}_{i_l}\}$$

We make the following remarks on the algorithm and terminology:

1. The criterion τ in Step 6 is used to test the individual "outlyingness" of the $N - n_s$ points in $\bar{S}_{n_s}^{(1)}$ with respect to the fitted model. As such, it can be different from the Γ criterion which measures the goodness-of-fit of the model to a subsample. For the linear regression model, for example, we compute the residuals for all the points in $\bar{S}_{n_s}^{(1)}$ using the fitted model in Step 5 and define criterion τ using the residuals as the following: a point in $\bar{S}_{n_s}^{(1)}$ is a declared good point if and only if its residual is within $\pm 2.5\hat{\sigma}$.

2. The first four steps are the same for SA1(n_s, r^*, k) and SA2(n_s, k). They differ in Step 5 and there is an extra step in SA2. In SA1(n_s, r^*, k), r^* good subsamples are selected to form the good subsample S_g. In SA2(n_s, k), only one good

subsample $S_{n_s}^{(1)}$ is selected from the k subsamples. Those points in $\bar{S}_{n_s}^{(1)}$ which are consistent (in term of the τ criterion) with the estimated model based on $S_{n_s}^{(1)}$ are added to $S_{n_s}^{(1)}$ to form the S_g^2.

3. The theoretical efficiency of algorithm SA1(n_s, r^*, k) can be set to a desired level through the choice of r^*. See the theoretical considerations discussed in Chapter 3. For the implementations of SA1 in examples throughout Chapter 3, we have set this efficiency to 99%. But the efficiency of algorithm SA2(n_s, k) is more difficult to control. It depends on the accuracy of the fitted model as well as the criterion τ. Hence we will not pursue a theoretical investigation here. Nevertheless, we can still use the observed efficiency E_{obs} given by

$$E_{obs} = \frac{\text{the number of points in } S_g^2}{n} \times 100\%,$$

to examine the efficiency of SA2(n_s, k) empirically when n is known.

4. A main advantage of Proposal II over proposal I is that proposal II is in general computationally less demanding. Indeed, the main computational effort for both SA1(n_s, r^*, k) and SA2(n_s, k) is in generating the required number of good subsamples, which is $r^* > 1$ for SA1(n_s, r^*, k) and 1 for SA2(n_s, k). At some fixed probability (of having the required least number of good subsamples) p^*, the total number of good subsamples k needed to generate 1 good subsample is much less than that needed for r^* good subsamples. Hence with the same p^*, SA2(n_s, k) requires far fewer subsamples and is faster to run than SA1(n_s, r^*, k). But since our objective is not maximizing saving in computation, for implementation of SA2(n_s, k), we generally set p^* to a considerably higher value than the typical value of 0.99 for SA1(n_s, r^*, k). See further discussion below.

To use the algorithm SA2(n_s, k), we need to choose the parameter values for n_s

and k which depend on the number of outliers m or the percentage of outliers in $\mathcal{S}_N$. In the absence of any information on this percentage, we use the default value of 10% which means the corresponding default value for m is $0.1N$. Parameter n_s must satisfies $n_s > m$ so that a subsample consisting entirely of outliers will not be undetected. Usually we set $n_s = 0.5N+1$ to ensure that the subsampling method has a high finite sample breakdown point. Parameter k can be determined as a special case in SA1(n_s, r^*, k) with $r^* = 1$. In SA2(n_s, k), we want to set k such that there is a probability of $p^* = 0.9999$ of having at least one good subsample in $S^1_{n_s}, S^2_{n_s} \ldots, S^k_{n_s}$. Using the result in (3.7), we can determine k by setting $p_1 > p^*$. To illustrate the determination of k for algorithm SA2(n_s, k), we computed the k values for various cases in Table 4.1. The subsample size is $n_s = 0.5N + 1$ for all cases. Since we only need one good subsample in SA2(n_s, k), we can afford to have a higher p^* than that in SA1(n_s, r^*, k). Thus, the chance of no good subsample in k subsamples is very small. Comparing Tables 4.1 with 3.1, we can see that even though p^* in Table 4.1 is much higher than that in Table 3.1, the values of k in Table 4.1 are smaller than those in Table 3.1. So we can set p^* to a high value to ensure that there is at least one good subsample in k subsamples when we use SA2(n_s, k).

Table 4.1: The number of subsamples k required to achieve a $p^* = 0.9999$.

Sample Size N	Number of good data n	Number of Outliers m	Subsample size n_s	r^*	Number of subsamples k
$N = 30$	$n = 30$	$m = 0$	$n_s = 16$	$r^* = 1$	$k = 1$
	$n = 27$	$m = 3$	$n_s = 16$	$r^* = 1$	$k = 99$
	$n = 25$	$m = 5$	$n_s = 16$	$r^* = 1$	$k = 651$
$N = 50$	$n = 50$	$m = 0$	$n_s = 26$	$r^* = 1$	$k = 1$
	$n = 45$	$m = 5$	$n_s = 26$	$r^* = 1$	$k = 455$
	$n = 42$	$m = 8$	$n_s = 26$	$r^* = 1$	$k = 6719$

Once the good subsample S^2_g is computed by the algorithm SA2(n_s, k), we apply

method Π to fit the regression model to S_g^2. The subsampling method-proposal II or SM2 consists of two steps: (1) using algorithm SA2(n_s, k) to obtain S_g^2 and (2) applying method Π to S_g^2 for estimation and inference.

4.2 Application of proposal II

In this section, we use some real data sets to compare the subsampling methods proposal I, II, the MM method and the method of least-squares.

Example 4.1. *This example uses the well-known stackloss data set presented by Brownlee (1965), which has been examined by a great number of statisticians by means of several methods. The data describe the operation of a plant for the oxidation of ammonia to nitric acid and consist of* $N = 21$ *four-dimensional observations (list in Table 4.2). The stackloss* (y) *is explained by the rate of operation* (x_1)*, the cooling water inlet temperature* x_2*, and the acid concentration* (x_3)*.*

A multiple linear regression model is proposed with three explanatory variables and one response variable, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. We apply SM1, SM2, MM and the least-squares methods to estimate regression parameters, β_0, β_1, β_2 and β_3. To apply SA2(n_s, k) to find the good subsample S_g^2, we need to set the percentage of outliers first. Since we know nothing about the number of outliers, we set the percentage to the default value of 10% for a default $m = 0.1N \approx 2$. Also we set the subsample size $n_s = 12$ $(> 0.5N)$. Now with $(N, m, n, n_s) = (21, 2, 19, 12)$, we can use (3.5) and (3.7) to find k satisfying $p_1 > p^*$. For $p^* = 0.9999$, it gave $k = 49$. We ran SA2$(n_s = 12, k = 49)$ and obtained a S_g^2 containing 16 good data points. The regression estimates based on S_g^2 are presented in Table 4.3. To compare the results with other estimators, Table 4.3 also shows the results for the LS estimates, the MM-estimates and the SM1 estimates. The SM1 estimates are computed at the 99%

Table 4.2: Stackloss Data

index (i)	Rate (x_1)	Temperature (x_2)	Acid Concentration (x_3)	Stackloss (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

efficiency level with $p^* = 0.99$. The parameter values for SA1(n_s, r^*, k) are $n_s = 12$, $r^* = 5$ and $k = 64$. The resulting combined data set S_g contains 19 good data points. Residual plots are constructed in Figure 4.1 to check for outliers. Observations with residuals outside of $\pm 2.5\hat{\sigma}$ are outliers. In Figure 4.1(b), (c) and (d), the residual plots for SM1, SM2, and MM-estimator show that there are outliers in the stackloss data. Both SM1 and MM indentify 2 outliers, while SM2 identifies 5 outliers. Rousseeuw and Leroy (1987) also identify 5 outliers. Therefore, SM2 with $m = 2$ works well for this data set and its estimates for σ is the smallest. SM1 and MM estimates are similar and they pick up the two very extreme outliers. The residual plot for LS in Figure 4.1(a) does not identify any outliers and the LS estimates are highly influenced by the outliers.

Table 4.3: Estimates of Stackloss Data with s.e. in brackets assuming $m = 2$.

	LS estimates (s.e.)	SM1 estimates (s.e.)	SM2 estimates (s.e.)	MM-estimates (s.e.)
$\hat{\beta}_0$	-39.9197 (11.8960)	-42.9453 (7.0149)	-35.4078 (3.9582)	-41.5246 (5.2978)
$\hat{\beta}_1$	0.7156 (0.1349)	1.0101 (0.0955)	0.8462 (0.0580)	0.9389 (0.1174)
$\hat{\beta}_2$	1.2953 (0.3680)	0.5701 (0.2507)	0.4453 (0.1442)	0.5796 (0.2630)
$\hat{\beta}_3$	-0.1521 (0.1563)	-0.1425 (0.0940)	-0.0924 (0.0513)	-0.1129 (0.0699)
$\hat{\sigma}$	3.2430 (2.0599)	1.895 (1.7194)	1.025 (0.9167)	1.912 (3.3332)

To check for the sensitivity of the value of m for SM1 and SM2, we set $m = 5$ to re-analyze this data set. The parameter values in SA1(n_s, r^*, k) and SA2(n_s, k) are computed and given in Table 4.4. Again $p^* = 0.9999$ is used to find k in SA2 and $p^* = 0.99$ in SA1. Applying algorithm SA1 and SA2, we obtain a good subsample S_g

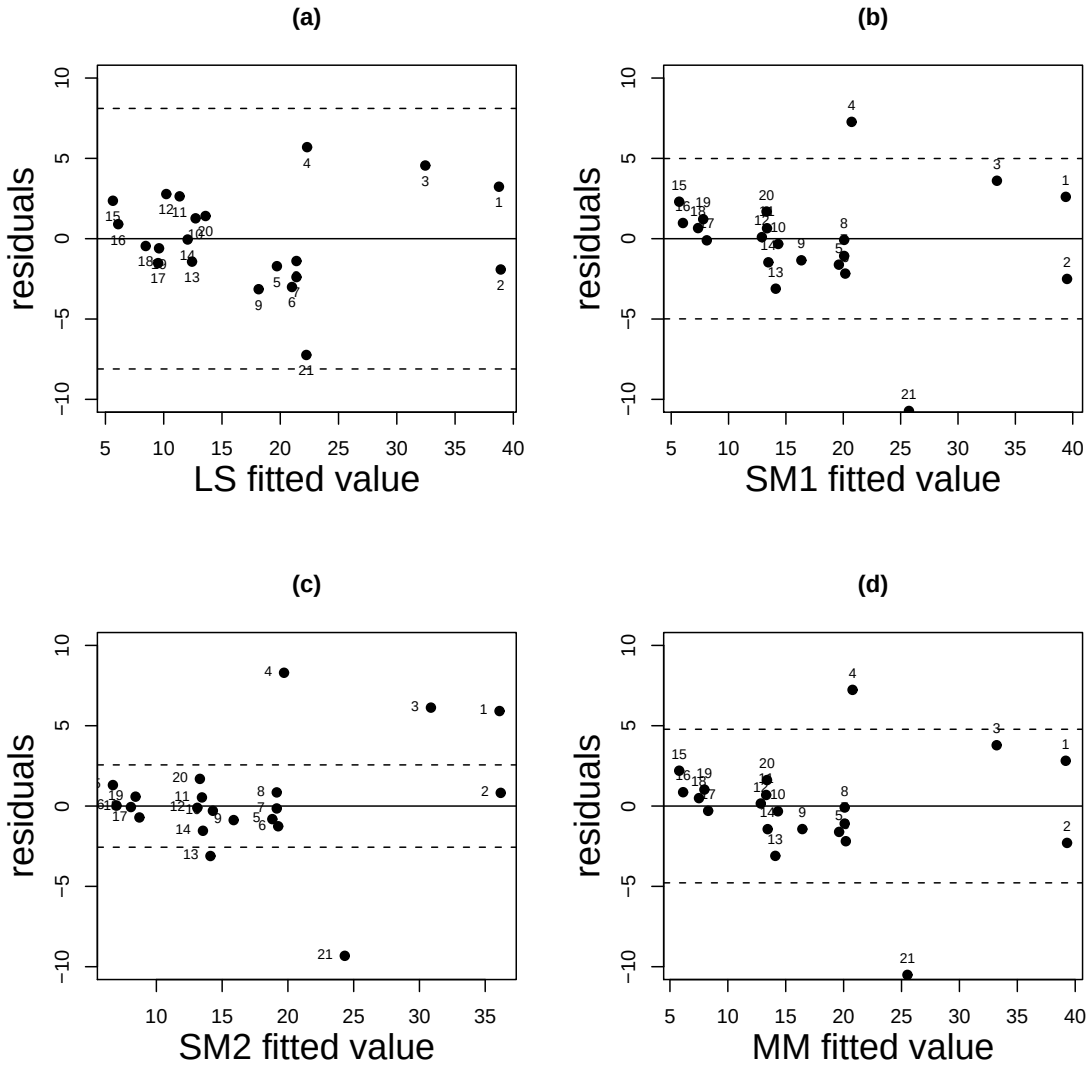

Figure 4.1: Residual plots: (a) LS residuals; (b)SM1 residuals; (c) SM2 residuals; (d)MM residuals. The dashed lines are $\pm 2.5\hat{\sigma}$.

containing 16 good points and S_g^2 containing 16 good data points too. In fact S_g and S_g^2 are the same. Furthermore, the S_g^2 for $m = 2$ is the same as S_g^2 for $m = 5$ in the method of SM2. Thus, for $m = 5$ the SM1 and SM2 estimates are the same as SM2 estimates for $m = 2$ in Table 4.3, and the residual plots for SM1 and SM2 are the same as the one for SM2 under assumption $m = 2$ in Figure 4.1(c). Both SM1 and SM2 identify the 5 outliers correctly. In addition, the s.e.'s from SM1 and SM2 are smaller than those from MM-estimator.

Table 4.4: Setting of different cases for Example 4.1

	No. of outliers (m)	Subsample size (n_s)	r^*	No. of subsamples (k)
SM1	$m = 2$	$n_s = 12$	$r^* = 5$	$k = 64$
	$m = 5$	$n_s = 12$	$r^* = 4$	$k = 1619$
SM2	$m = 2$	$n_s = 12$	$r^* = 1$	$k = 49$
	$m = 5$	$n_s = 12$	$r^* = 1$	$k = 1483$

From this example, we can see that SM2 is not sensitive to the value we set for m. The estimates for β_0, β_1, β_2 and β_4, are the same for $m = 2$ and $m = 5$. However, SM1 seems to be sensitive to the value of m, because the estimates are different for $m = 2$ and $m = 5$. This is related to the number of subsamples k used in SA1(n_s, r^*, k) and k is not large enough to get r^* good subsamples when m is not specified correctly. As a practical guideline for SM1, we could use a large k if the computation is not a problem. From Table 4.4, $k = 64$ for $m = 2$ seems too small. When $m = 5$, we actually need $k = 1619$. Or we can set m to be large, so k has to be large. For example, if we want the algorithm to work for the data sets with maximum proportion of contamination as ϵ_0, then $m = \epsilon_0 N$.

Example 4.2. *(Montgomery and Peck, 2006, p.70). A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in*

predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time (y) *are the number of cases of product stocked* (x_1) *and the distance walked by the route driver* (x_2)*. The engineer has collected* $N = 25$ *observations on delivery time, which are shown in Table 4.5.*

A multiple linear regression model is used to analyze the relationship between the two explanatory variables (x_1 and x_2) and the response variable (y). The underlying regression model is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. We want to find the best estimate for β_0, β_1 and β_2. Similar to Example 4.1 and set the number of outliers $m = 2$ and $m = 5$ to do the analysis. With $n_s = 14$, parameter values in SA1(n_s, r^*, k) and SA2(n_s, k) are computed in Table 4.6.

After applying the algorithms SA1(n_s, r^*, k) and SA2(n_s, k) for $m = 2$, we obtain a good sample S_g with 24 good data points and S_g^2 with 22 good data points. The estimates for regression parameters are presented in Table 4.7. It is clear that SM1, SM2 and MM-estimates are very similar, but LS estimates are quite different. The residual plots in Figure 4.2 show that there is one extreme outlier (observation 9) in delivery time data set. SM1, SM2 and MM-estimators all identify this outlier. SM2 also identifies two mild outliers (observation 11 and 22) which do not have much influence on the regression estimates. The LS method does not identify any outliers, which explains why LS estimates are so different.

For $m = 5$, two algorithms SA1(n_s, r^*, k) and SA2(n_s, k) yield good subsample S_g with 24 good data points and S_g^2 with 22 good data points. In fact, they are the same as for $m = 2$. Thus, the regression estimates are exactly the same for $m = 5$ and $m = 2$. In this example, both SA1 and SA2 are not sensitive to the value of m,

Table 4.5: Delivery time data

Observation Number	Delivery Time, y (min)	Number of Cases, x_1	Distance, x_2 (ft)
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

Table 4.6: Setting of different cases for Example 4.2

	No. of outliers (m)	Subsample size (n_s)	r^*	No. of subsamples (k)
SM1	$m = 2$	$n_s = 14$	$r^* = 5$	$k = 60$
	$m = 5$	$n_s = 14$	$r^* = 4$	$k = 1152$
SM2	$m = 2$	$n_s = 14$	$r^* = 1$	$k = 46$
	$m = 5$	$n_s = 14$	$r^* = 1$	$k = 1055$

Table 4.7: Estimates of Delivery Time Data with s.e. in brackets assuming $m = 2$

	LS estimates (s.e.)	SM1 estimates (s.e.)	SM2 estimates (s.e.)	MM-estimates (s.e.)
$\hat{\beta}_0$	2.3412	4.4472	4.8143	4.4718
	(1.0967)	(0.9525)	(0.8763)	(0.6979)
$\hat{\beta}_1$	1.6159	1.4977	1.4452	1.4718
	(0.1707)	(0.1302)	(0.1199)	(0.1401)
$\hat{\beta}_2$	0.0144	0.0103	0.0098	0.0108
	(0.0036)	(0.0029)	(0.0026)	(0.0044)
$\hat{\sigma}$	3.2590	2.4300	2.2000	2.0230
	(3.1207)	(2.3219)	(2.0977)	*

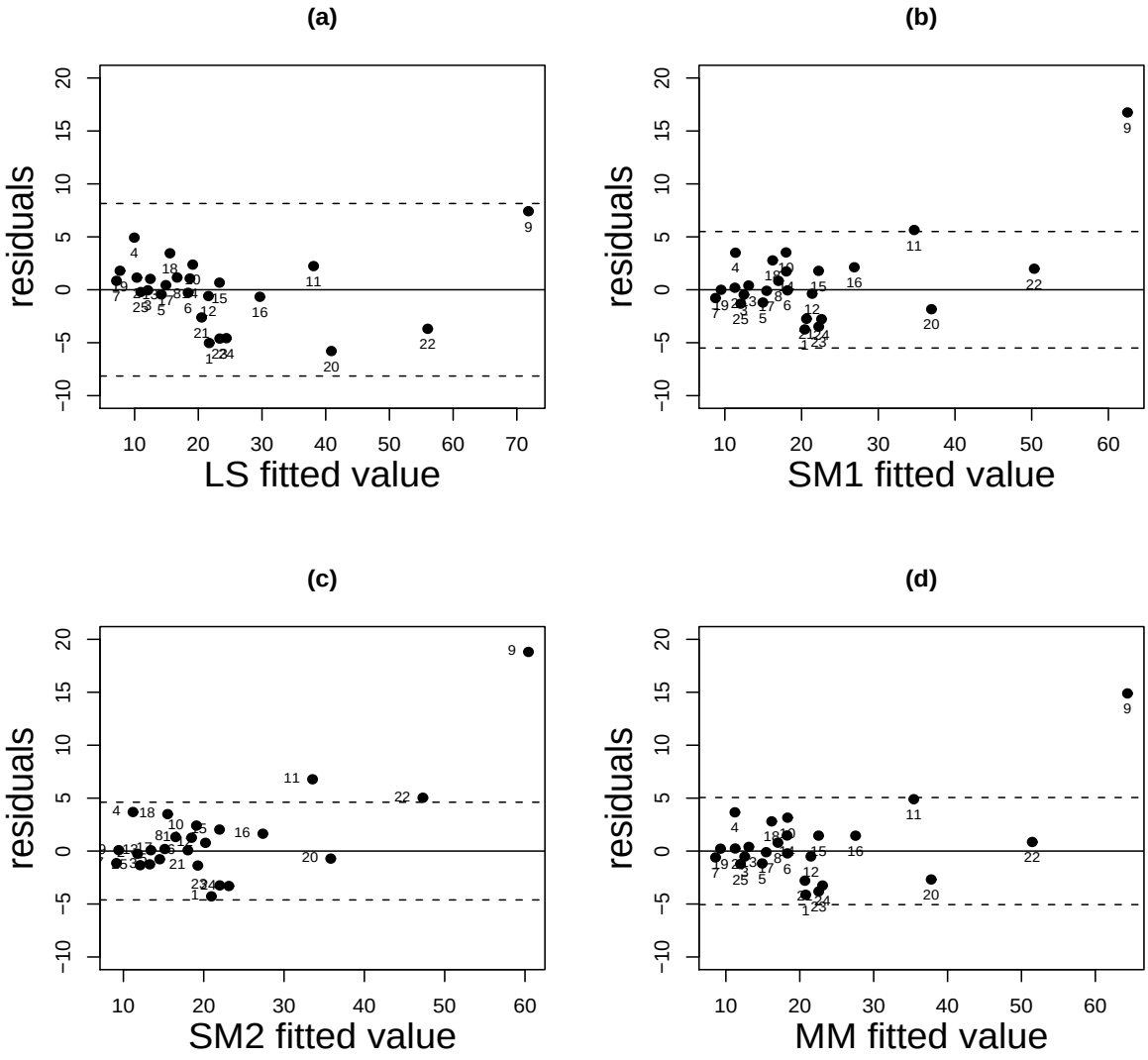

Figure 4.2: Residual plots: (a) residuals from LS; (b) residuals from SM1; (c) residuals from SM2; (d) residuals from MM. The dashed lines are $\pm 2.5\hat{\sigma}$.

because there is only one extreme outlier in the delivery time data and the values of m are set greater than 1. Both SM1 and SM2 work well. For Example 4.1, when $m = 5$ and $k = 1619$, it takes about 20 seconds to compute SM1 estimates on a laptop with CPU T7700, while it takes about 15 seconds when $k = 1483$ for SM2 on the same laptop. For Example 4.2, the computation time is very similar.

4.3 Influence function and breakdown point for SM1 and SM2

In this section we briefly examine the robustness of SM1 and SM2 using the empirical influence function and the finite sample breakdown point introduced in Chapter 2.

Consider Example 2.2 where the simple linear regression model $E(y) = \beta_0 + \beta_1 x$ was used to generate $N = 20$ data points with true values $\beta_0 = 20$ and $\beta_1 = -5$. To apply SM1 and SM2, we set the default values $m = 2$ and $n_s = 11$. With $p^* = 0.99$ and 99% efficiency, we found $r^* = 5$ and $k = 58$ for SM1, and with $p^* = 0.9999$, we got $k = 44$ for SM2. To construct the empirical influence function, we moved one observation in the horizontal direction to form an x-outlier and by vertical direction to form an y-outlier. We then plotted the empirical influence functions for $\hat{\beta}_0$ and $\hat{\beta}_1$ with the x-outlier and y-outlier in Figure 4.3. It is clear that all the influence functions are bounded. So both SM1 and SM2 are robust against the x-outlier and the y-outlier. Figure 4.4 shows the fitted lines for SM1 and SM2 with one x-outlier and one y-outlier.

To find out how many outliers SM1 and SM2 can handle, we check for the finite sample breakdown points. Theoretically, the breakdown points are not larger than

$$\frac{N - n_s}{N} = \frac{20 - 11}{20} = \frac{9}{20}.$$

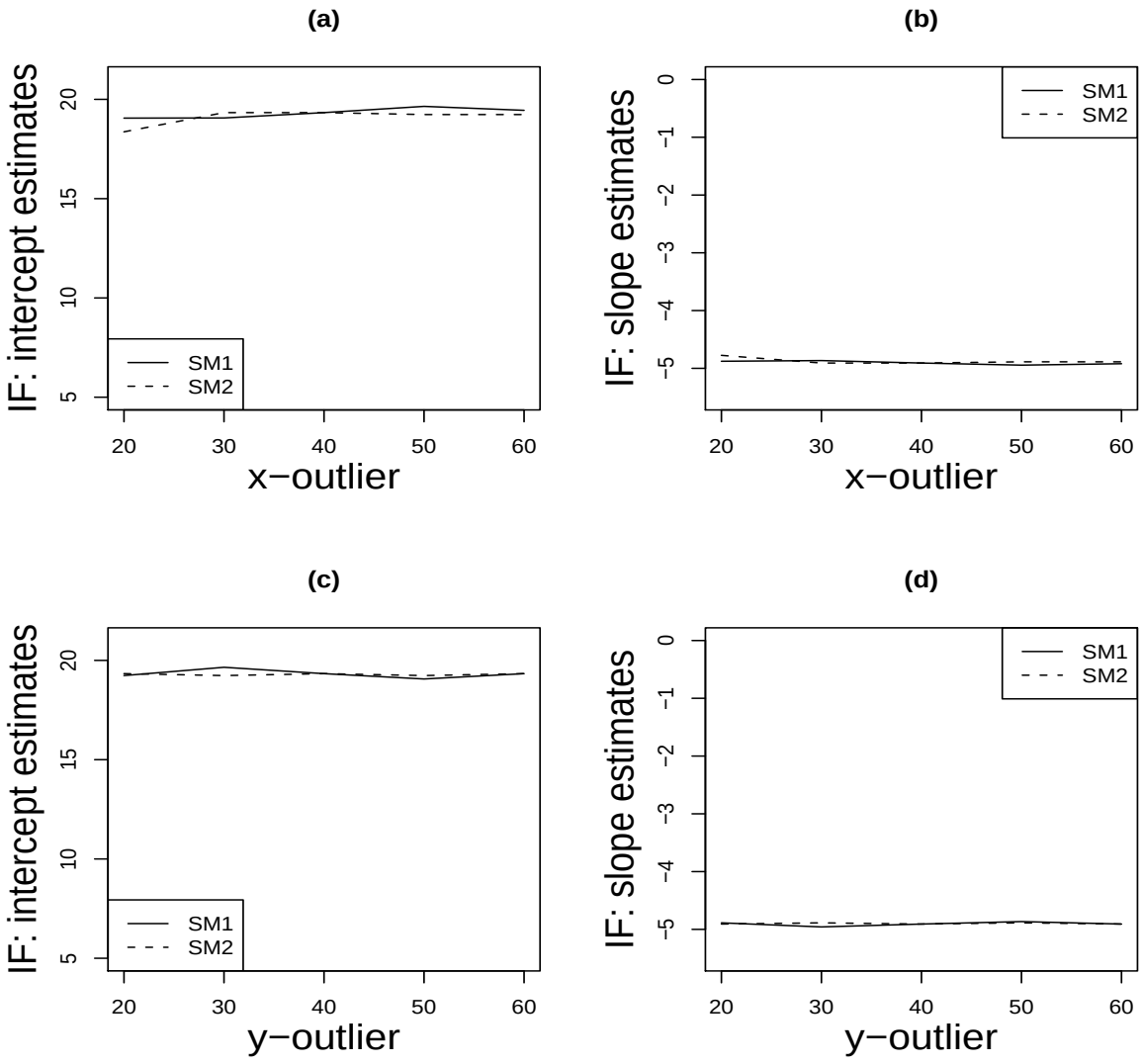

Figure 4.3: The empirical influence functions: (a) intercept for one x-outlier; (b) slope for one x-outlier; (c) intercept for one y-outlier; (d) slope for one y-outlier. The solid lines are the empirical influence functions for SM1 and the dashed lines are the empirical influence functions for SM2.

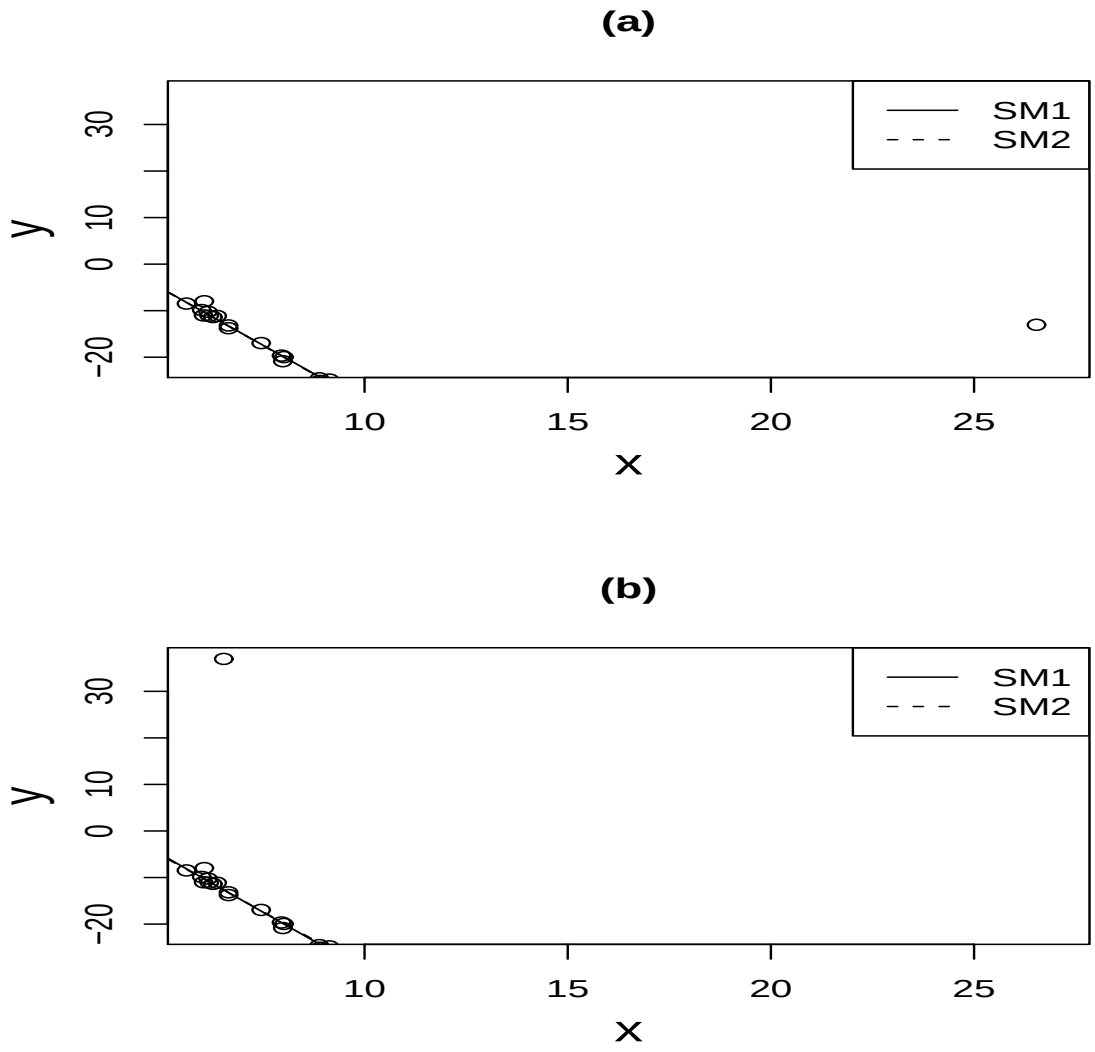

Figure 4.4: The fitted lines: (a) with one x-outlier; (b) with one y-outlier. The solid lines are the fitted lines for SM1 and the dashed lines are the fitted lines for SM2.

Assuming $m = 5$ in SA1(n_s, r^*, k) and SA2(n_s, k), we get $r^* = 4$ and $k = 1233$ in SA1 and $k = 1129$ in SA2. The estimated regression lines for various number of outliers in the data are shown in Figure 4.5. We can see that the SM1 breaks down with 6 outliers and the SM2 breaks down with 8 outliers. Hence, the breakdown points are $5/20 = 0.25$ for SM1 and $7/20 = 0.35$ for SM2.

To see if the breakdown point depends on the value of m we set in SA1(n_s, r^*, k) and SA2(n_s, k), we repeat the above analysis for $m = 2$. The estimated regression lines for various number of outliers are shown in Figure 4.6. In this case, the SM1 breaks down with 3 outliers and the SM2 breaks down with 5 outliers. Correspondingly, the breakdown points are $2/20 = 0.1$ for SM1 and $4/20 = 0.2$ for SM2. Thus, the breakdown points tend to depend on the value of m set in algorithms SA1 and SA2, and the SM2 has higher breakdown points than the SM1. Further, the larger the m is, the higher the breakdown point is. However, if k is set to be very large, the breakdown points for SM1 and SM2 will be close to $(N - n_s)/N$ in theory.

4.4 Simulation study for SM2

We described the implementation and application of the subsampling method-proposal II in the previous sections. In this section, we conduct a simulation study to compare our subsampling methods, SM1 and SM2, with MM-estimator. In particular, we will examine the bias, standard error and efficiency for those methods. Similar to the simulation study in Chapter 3, we still use "optimal method" as the method of least-squares applied to the good data only (LS-).

The design of the simulation study is as follows:

Regression model: The multiple linear regression model with 3 regressors in Example 3.3 is used. Two error distributions are considered: [1] symmetric error

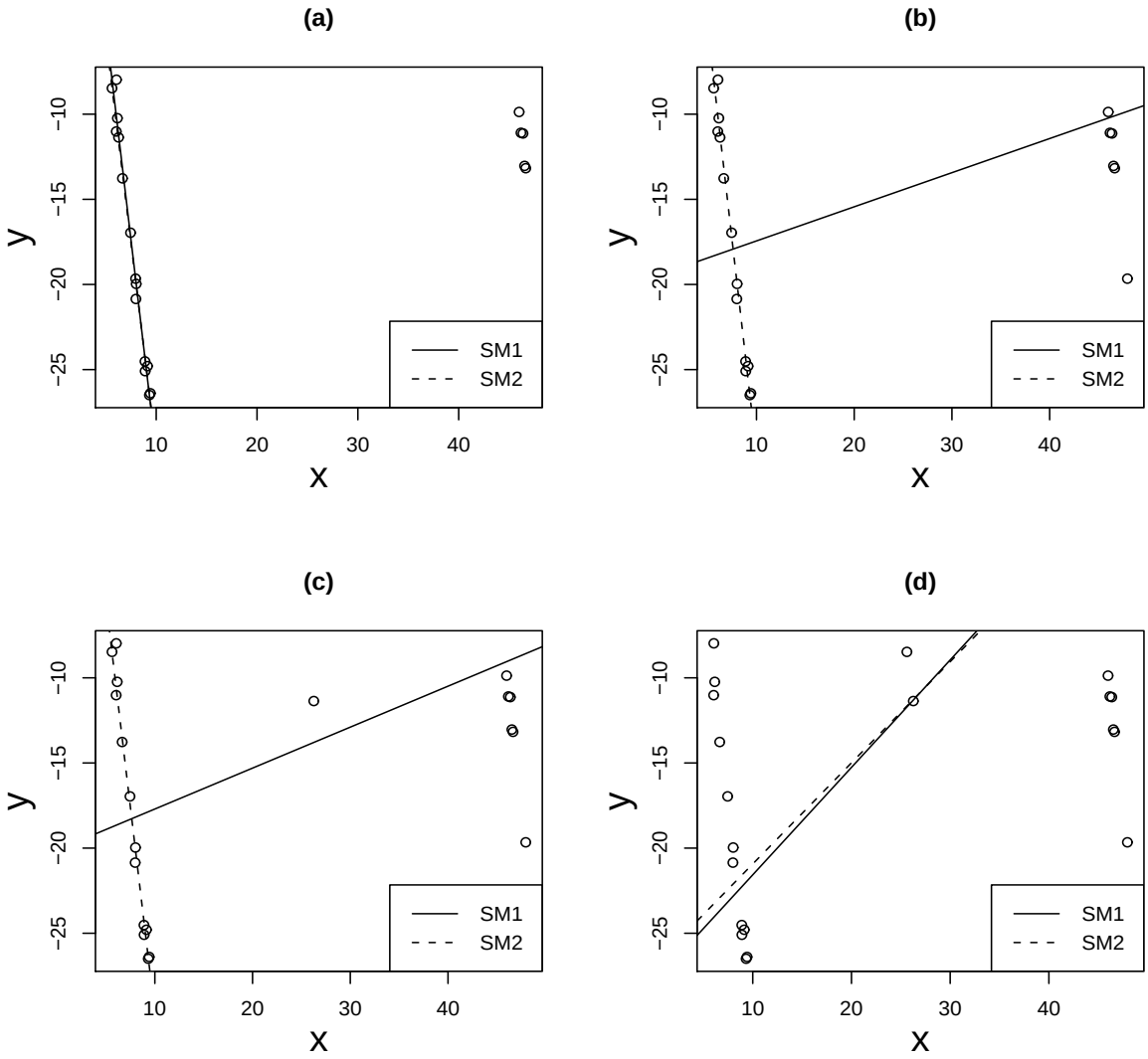

Figure 4.5: The fitted line (solid for the SM1, dashed for the SM2) for various numbers of outliers: (a) 5 outliers; (b) 6 outliers; (c) 7 outliers; (d) 8 outliers.

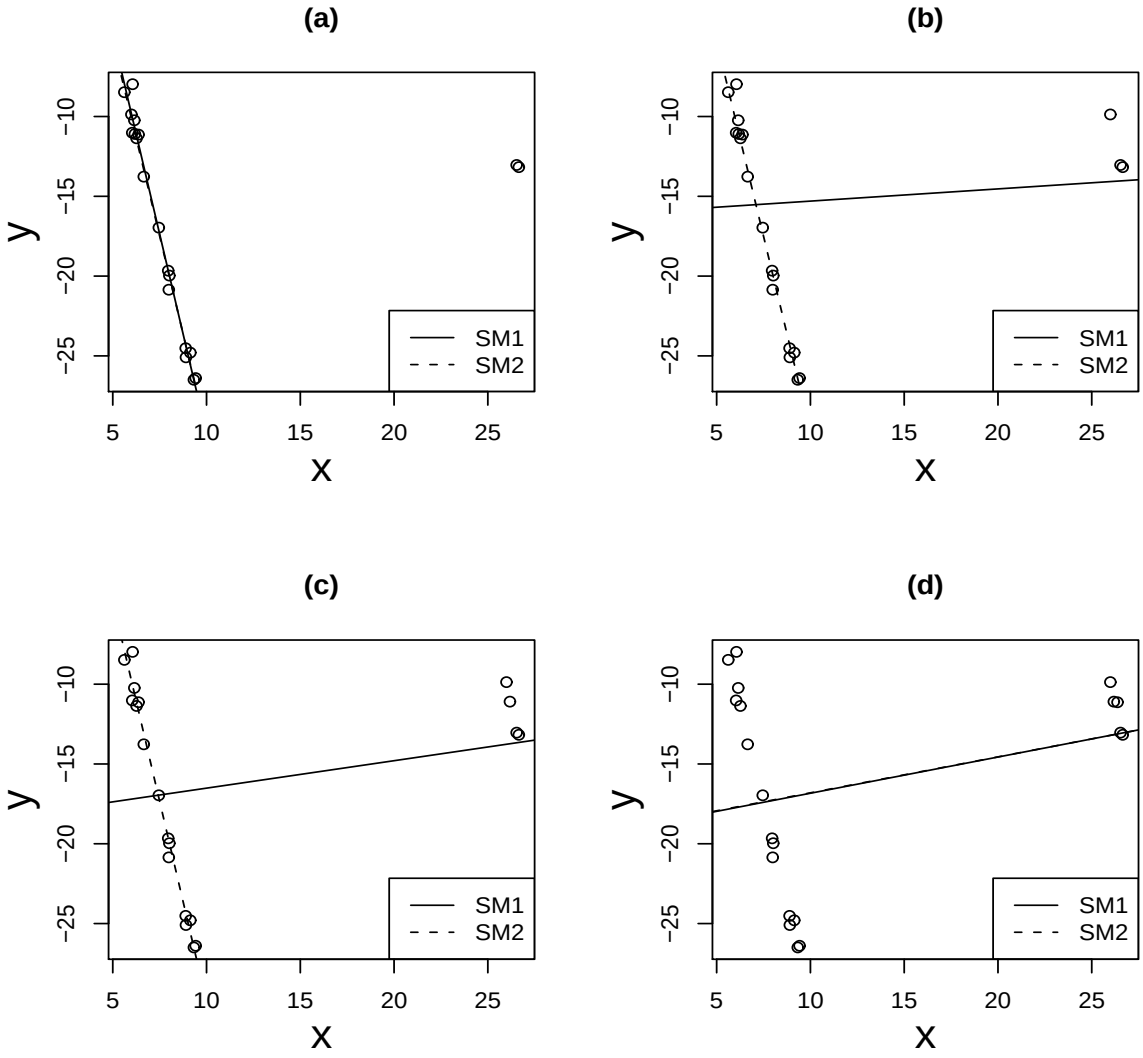

Figure 4.6: The fitted line (solid for the SM1, dashed for the SM2) for various numbers of outliers: (a) 2 outliers; (b) 3 outliers; (c) 4 outliers; (d) 5 outliers.

distribution $N(0, \sigma = 2)$ and [2] asymmetric error distribution $\chi^2_2 - 2$.

Sample size: $N = 30$ and 50.

True number of outliers: $m = 0$ or $m = 0.1N$.

Table 4.8: The number of subsamples k for each combination.

N	$\hat{m}$	n_s	r^*	k
$N = 30$	$m = 3$	$n_s = 16$	$r^* = 5$	$k = 126$
	$m = 5$	$n_s = 16$	$r^* = 5$	$k = 823$
$N = 50$	$m = 5$	$n_s = 26$	$r^* = 5$	$k = 575$
	$m = 8$	$n_s = 26$	$r^* = 5$	$k = 8468$

Table 4.9: Ten combinations of contaminations (N, m, n), distributions and τ.

	Distribution of ε	(N, m, n)	τ
Case 1	$\varepsilon \sim N(0, \sigma = 2)$	$(30, 0, 30)$	$\pm 2.5\hat{\sigma}$
			$\pm 3.5\hat{\sigma}$
		$(30, 3, 27)$	$\pm 2.5\hat{\sigma}$
			$\pm 3.5\hat{\sigma}$
Case 2	$\varepsilon \sim N(0, \sigma = 2)$	$(50, 0, 50)$	$\pm 3.5\hat{\sigma}$
		$(50, 5, 45)$	$\pm 3.5\hat{\sigma}$
Case 3	$\varepsilon \sim \chi^2_2 - 2$	$(30, 0, 30)$	$\pm 3.5\hat{\sigma}$
		$(30, 3, 27)$	$\pm 3.5\hat{\sigma}$
Case 4	$\varepsilon \sim \chi^2_2 - 2$	$(50, 0, 50)$	$\pm 3.5\hat{\sigma}$
		$(50, 5, 45)$	$\pm 3.5\hat{\sigma}$

Estimators: Basically, four estimators, SM1, SM2 and MM and LS-, are studied. In SA1(n_s, r^*, k), $n_s = 0.5N+1$ and $r^* = 5$ are used for all cases. At the same time, k is computed with two estimates of m, $\hat{m} = 0.1N$ and $\hat{m} \approx 0.2N$, by (3.14). The values of n_s, $\hat{m}$, r^* and k are given in Table 4.8. In SA2(n_s, k), the set of n_s and k is the same as in SA1(n_s, r^*, k) so that the cost of the computation is

about the same for SM1 and SM2. Also in SA2, two versions of criterion τ are examined. One uses cutoff points $\pm 2.5\hat{\sigma}$ to declare outliers, and the other one uses $\pm 3.5\hat{\sigma}$.

Table 4.9 summarizes all the combinations of error distribution, contaminations (N, m, n) and criterion τ in the simulation study. The simulation results for the bias, the standard error and observed efficiency (E_{obs}) and computed based on 1000 simulation runs, and they are presented in Tables 4.10-4.14. From the results, we have the following conclusions for the comparison of the four estimators.

1. When the error term is symmetrically distributed, we can first see that the estimates of SM1 and SM2 seem to be unbiased from the results of Tables 4.10-4.12. Moreover, when we increase the sample size N to 50, the biases of the estimates for SM1 and SM2 seem to be smaller. Concerning the efficiency of the estimates, we can find that the efficiency decreases as the $\hat{m}$ increases for both SM1 and SM2. Also, SM2 is less efficient than SM1 with the regular conservative choice of τ ($\pm 2.5\hat{\sigma}$). However it is not surprising because we have higher breakdown point for SM2. In addition, the efficiency of SM2 can be improved by setting a larger cutoff, ($\pm 3.5\hat{\sigma}$). One reason for choosing a larger cutoff is the underestimation of $\hat{\sigma}$. Next, when comparing MM, SM1 and SM2, we can find that MM owns the highest efficiency due to the smallest standard errors of the estimates. Overall, the estimation of LS- seems to be optimal with high efficiency and unbiasness.

2. When the error term is asymmetrically distributed, we can first find out that the estimates of the intercept are biased for SM1, SM2 and MM from Tables 4.13 and 4.14. However, the estimates of the rest are unbiased for these methods. Concerning the efficiency of the estimates, SM2 with $\pm 3.5\hat{\sigma}$ has similar efficiency

Table 4.10: The summary of SM1, SM2, MM and LS- estimates: $N = 30$, $m = 0$ and $\varepsilon \sim N(0, \sigma = 2)$

		$\hat{m} = 3$			$\hat{m} = 5$			MM	LS-
		SM1	SM2		SM1	SM2			
			(τ: $\pm 2.5\hat{\sigma}$)	(τ: $\pm 3.5\hat{\sigma}$)		(τ: $\pm 2.5\hat{\sigma}$)	(τ: $\pm 3.5\hat{\sigma}$)		
$\hat{\beta}_0$	average	-15.6363	-15.5859	-15.7903	-15.7597	-15.7344	-15.7703	-15.6032	-15.6864
	bias	0.0637	0.1141	0.0903	-0.0597	0.0344	0.0703	0.0968	0.0136
	s.e.	3.3852	4.2914	3.5841	3.8566	4.9774	4.5085	3.2275	2.8399
$\hat{\beta}_1$	average	16.8583	16.8531	16.8759	16.8584	16.8651	16.8733	16.8599	16.8611
	bias	0.0117	0.0169	0.0059	0.0116	0.0049	0.0033	0.0101	0.0089
	s.e.	0.3273	0.4253	0.3449	0.3684	0.4788	0.4227	0.3127	0.2737
$\hat{\beta}_2$	average	19.0021	18.9979	18.9987	19.0165	19.0244	18.9967	19.0142	19.0041
	bias	0.0021	0.0021	0.0013	0.0165	0.0244	0.0033	0.0142	0.0041
	s.e.	0.3962	0.5244	0.4431	0.4550	0.6027	0.5217	0.3796	0.3455
$\hat{\beta}_3$	average	11.003	11.0017	11.0091	11.0167	11.0016	11.0084	11.0112	11.0069
	bias	0.0030	0.0017	0.0091	0.0167	0.0016	0.0084	-0.0112	0.0069
	s.e.	0.3252	0.4078	0.3345	0.3648	0.4698	0.4227	0.3109	0.2750
$\hat{n}$	average	27.6470	25.2810	27.9410	25.9750	22.7250	25.7420		
	E_{obs}	92.16%	84.27%	93.14%	86.58%	75.75%	85.81%		
	s.e.	1.1161	1.6039	1.3898	1.5365	1.5485	1.8091		

Table 4.11: The summary of SM1, SM2, MM and LS- estimates: $N = 30$, $m = 3$ and $\varepsilon \sim N(0, \sigma = 2)$

		$\hat{m} = 3$			$\hat{m} = 5$			MM	LS-
		SM1	SM2		SM1	SM2			
			(τ: $\pm 2.5\hat{\sigma}$)	(τ: $\pm 3.5\hat{\sigma}$)		(τ: $\pm 2.5\hat{\sigma}$)	(τ: $\pm 3.5\hat{\sigma}$)		
$\hat{\beta}_0$	average	-15.5405	-15.4739	-15.5173	-15.4953	-15.3729	-15.4531	-15.4993	-15.5142
	bias	0.1595	0.2261	0.1827	0.2047	0.3271	0.2469	0.2007	0.1858
	s.e.	3.0510	3.7310	3.1595	3.5335	4.3734	3.6793	2.9904	2.9565
$\hat{\beta}_1$	average	16.866	16.8685	16.867	16.8604	16.8554	16.8565	16.8633	16.8652
	bias	0.0040	0.0015	0.0030	0.0096	0.0146	0.0135	0.0067	0.0048
	s.e.	0.3048	0.3745	0.3185	0.3546	0.4417	0.3663	0.2961	0.2931
$\hat{\beta}_2$	average	18.9863	18.9787	18.983	18.982	18.9746	18.9847	18.987	18.9872
	bias	0.0137	0.0213	0.0170	0.0180	0.0254	0.0153	0.0130	0.0128
	s.e.	0.3766	0.4603	0.3975	0.4214	0.5392	0.4566	0.3654	0.3613
$\hat{\beta}_3$	average	10.9821	10.9704	10.9786	10.9843	10.9721	10.98	10.9782	10.9783
	bias	0.0179	0.0296	0.0214	0.0157	0.0279	0.0200	0.0218	0.0217
	s.e.	0.2851	0.3463	0.2954	0.3409	0.4175	0.3503	0.2787	0.2758
$\hat{n}$	average	26.4790	25.0360	26.4710	24.8680	22.6160	25.0390		
	E_{obs}	98.07%	92.73%	98.04%	92.10%	83.76%	92.74%		
	s.e.	0.6768	1.2555	0.7441	1.0127	1.4686	1.3672		

Table 4.12: The summary of SM1, SM2, MM and LS- estimates: $N = 50$ and $\varepsilon \sim N(0, \sigma = 2)$

		m=0						m=5					
		$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-	$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-
		SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$			SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$		
$\hat{\beta}_0$	ave.	-15.7788	-15.7310	-15.7696	-15.7185	-15.7874	-15.7720	-15.8223	-15.7537	-15.7516	-15.7512	-15.7964	-15.7950
	bias	0.0788	0.0310	0.0696	0.0185	0.0874	0.0720	0.1223	0.0537	0.0516	0.0512	0.0964	0.0950
	s.e.	2.5337	2.4121	2.6127	2.5448	2.2743	2.0673	2.1965	2.2348	2.4573	2.3813	2.1798	2.1789
$\hat{\beta}_1$	ave.	16.8710	16.8683	16.8673	16.8668	16.8714	16.8703	16.8794	16.8765	16.8710	16.8753	16.8794	16.8794
	bias	0.0010	0.0017	0.0027	0.0032	0.0014	0.0003	0.0094	0.0065	0.0010	0.0053	0.0094	0.0094
	s.e.	0.2571	0.2399	0.2647	0.2589	0.2550	0.2052	0.2225	0.2220	0.2371	0.2342	0.2158	0.2152
$\hat{\beta}_2$	ave.	19.0113	19.0051	19.0072	19.0071	19.0139	19.0026	19.0041	19.0018	19.0037	18.9983	19.0044	19.0017
	bias	0.0113	0.0051	0.0072	0.0071	0.0139	0.0026	0.0041	0.0018	0.0037	0.0017	0.0044	0.0017
	s.e.	0.3107	0.3009	0.3226	0.3093	0.2805	0.2586	0.2673	0.2738	0.2974	0.2971	0.2679	0.264
$\hat{\beta}_3$	ave.	11.0037	11.0035	11.0107	11.0027	11.0061	11.0015	11.0074	10.9997	11.0051	11.0027	11.0043	11.0005
	bias	0.0037	0.0035	0.0107	0.0027	0.0061	0.0015	0.0074	0.0003	0.0051	0.0027	0.0043	0.0005
	s.e.	0.2504	0.2439	0.2593	0.2558	0.2273	0.2097	0.2243	0.2260	0.2453	0.2408	0.2200	0.2172
$\hat{n}$	ave.	45.7440	47.9670	44.7810	47.4330			44.0920	44.5820	42.5820	43.9020		
	E_{obs}	91.49%	95.93%	89.56%	94.87%			97.98%	99.07%	94.63%	97.56%		
	s.e.	1.4099	1.5127	1.4377	1.6680			0.8765	0.6795	1.6726	1.1161		

Table 4.13: The summary of SM1, SM2, MM and LS- estimates: $N = 30$ and $\varepsilon \sim \chi^2_2 - 2$

		m=0						m=3					
		$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-	$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-
		SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$			SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$		
$\hat{\beta}_0$	ave.	-16.1754	-16.2477	-16.3139	-16.3264	-16.2482	-15.7523	-15.8402	-15.8435	-15.9095	-15.904	-15.8538	-15.732
	bias	0.4754	0.5477	0.6139	0.6264	0.5482	0.0523	0.1402	0.1435	0.2095	0.204	0.1538	0.0320
	s.e.	2.3991	2.3731	2.3464	2.4714	2.1814	2.7989	2.7443	2.6777	2.4675	2.518	2.4257	2.8802
$\hat{\beta}_1$	ave.	16.8641	16.8686	16.8679	16.8689	16.866	16.8666	16.8796	16.8708	16.8714	16.8642	16.8672	16.8821
	bias	0.0641	0.0686	0.0679	0.0689	0.0660	0.0666	0.0796	0.0708	0.0714	0.0642	0.0672	0.0821
	s.e.	0.2327	0.2298	0.2291	0.2534	0.2131	0.2701	0.2730	0.2691	0.2509	0.2526	0.2395	0.2903
$\hat{\beta}_2$	ave.	19.024	19.0228	19.0239	19.0265	19.0219	19.0261	18.9905	18.9847	18.9826	18.9793	18.9842	18.9992
	bias	0.0240	0.0228	0.0239	0.0265	0.0219	0.0261	0.0095	0.0153	0.0174	0.0207	0.0158	0.0008
	s.e.	0.2897	0.2924	0.2990	0.3188	0.2714	0.3378	0.3394	0.3321	0.3108	0.3120	0.2941	0.3525
$\hat{\beta}_3$	ave.	11.0055	11.0037	11.0021	10.9957	11.0089	11.0028	10.9949	10.9907	10.9835	10.9832	10.9484	10.9914
	bias	0.0055	0.0037	0.0021	0.0043	0.0089	0.0028	0.0051	0.0093	0.0165	0.0168	0.0516	0.0086
	s.e.	0.2307	0.2359	0.2320	0.2502	0.2237	0.2811	0.2722	0.2661	0.2501	0.2451	0.2379	0.2867
$\hat{n}$	ave.	26.9450	26.8140	25.7820	25.4520			26.2070	25.8920	25.0630	24.7650		
	E_{obs}	89.82%	89.38%	85.94%	84.84%			97.06%	95.89%	92.82%	91.72%		
	s.e.	1.0550	1.3591	1.1789	1.5072			0.7132	0.8838	0.8740	1.1032		

Table 4.14: The summary of SM1, SM2, MM and LS- estimates: $N = 50$ and $\varepsilon \sim \chi^2_2 - 2$

		m=0						m=5					
		$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-	$\hat{m} = 5$		$\hat{m} = 8$		MM	LS-
		SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$			SM1	SM2 $\pm 3.5\hat{\sigma}$	SM1	SM2 $\pm 3.5\hat{\sigma}$		
$\hat{\beta}_0$	ave.	-16.1431	-16.0695	-16.3425	-16.2683	-16.2107	-15.6764	-15.922	-16.0256	-16.118	-16.1675	-16.1469	-15.7901
	bias	0.4431	0.3695	0.6425	0.5683	0.5107	0.0236	0.2220	0.3256	0.4180	0.4675	0.4469	0.0901
	s.e.	1.8020	1.5565	1.2727	1.6532	1.4199	2.2489	2.0431	1.8988	1.8262	1.7784	1.7122	2.1835
$\hat{\beta}_0$	ave.	16.8892	16.8734	16.8848	16.8676	16.8619	16.8656	16.8758	16.8767	16.8779	16.8781	16.8773	16.8752
	bias	0.0892	0.0734	0.0848	0.0676	0.0619	0.0656	0.0758	0.0767	0.0779	0.0781	0.0773	0.0752
	s.e.	0.1509	0.1237	0.1499	0.1773	0.1347	0.2122	0.2010	0.1869	0.1748	0.1740	0.1672	0.2146
$\hat{\beta}_0$	ave.	19.0108	19.0068	19.0098	19.0075	19.0105	19.0062	19.0137	19.0104	19.0007	19.0067	19.0007	19.0136
	bias	0.0108	0.0068	0.0098	0.0075	0.0105	0.0062	0.0137	0.0104	0.0007	0.0067	0.0007	0.0136
	s.e.	0.1832	0.1353	0.1613	0.1375	0.1148	0.2068	0.2637	0.2422	0.2267	0.2209	0.2148	0.2794
$\hat{\beta}_0$	ave.	11.0026	10.9791	11.0044	11.0312	10.9987	10.9964	11.0044	11.0035	11.0018	11.001	11.0005	10.9991
	bias	0.0026	0.0209	0.0044	0.0312	0.0013	0.0036	0.0044	0.0035	0.0018	0.001	0.0005	0.0009
	s.e.	0.1040	0.1186	0.0872	0.1292	0.1150	0.1640	0.2025	0.1863	0.1799	0.1726	0.1709	0.2148
$\hat{n}$	ave.	44.6780	45.3510	42.6740	43.6030			43.8440	43.5830	42.0080	42.0480		
	E_{obs}	89.36%	90.7%	85.35%	87.21%			97.43%	96.85%	93.35%	93.44%		
	s.e.	0.8432	1.4142	1.1005	1.4181			0.8924	0.9241	0.9241	1.3158		

as that of SM1. Like in the symmetric error term case above, the MM owns the highest efficiency among the three methods. Finally, although the estimates of LS- are still unbiased, it has the lowest efficiency compared with those of other methods.

4.5 Discussion

In this chapter, we presented subsampling method-proposal II focusing on the subsampling algorithm II and its implementation. We applied both proposals to two examples of multiple linear regression models. The SM2 worked very well and is more effective to identify outliers than SM1. In addition, we needed fewer subsamples to obtain one good subsample in SA2(n_s, k). These results are based on the assumption that there exists an effective criterion τ in step 6 of SA2 to test if a point is a good point or not. For linear regression models, the criterion that we use in remark 1 seems to be very effective. For other regression models such as logistic regression model, it may be very hard to find an effective criterion to test for each data point. In that situation, the use of SM2 is limited. In general, SM1 has wider applications than SM2. For the examples, we have used mostly the default parameter settings for SA1. The relative performance of SM1 may be further improved by more carefully setting the parameter values.

The empirical influence function and finite sample breakdown point are examined through one example in section 4.3. The results show that SM1 and SM2 are robust against x-outliers and y-outliers. But for fixed k, SM2 has higher breakdown point than SM1. As k increases, the breakdown points for SM1 and SM2 are also increase. In practice, we can always set a large value of k to get high breakdown points for SM1 and SM2 when the computation is not a problem. Note that by the very construction

of the subsampling algorithm, there is always a small chance that SM1 and SM2 may fail when this is one or more outliers. But with a high p^* value, this possible failure is not probable as we have seen in this example. So for all practical purposes, the influence functions of both methods are bounded.

With our parameter settings in the discussion above, SM1 and SM2 seem to be sensitive to the value of m. In practice, we may try several values of m to check for the real number of outliers. If m is set to be larger than the true number of outliers, then both SM1 and SM2 provide reliable estimates. Another way to avoid the issue of setting m is to set k to be very large without specifying m in the algorithms.

Finally, like the subsampling algorithm SA1(n_s, r^*, k) studied in Chapter 3, the subsampling algorithm SA2(n_s, k) examined in this chapter is also the most basic version of a general subsampling algorithm-II that Dr. Tsao and Dr. Zhou have been working on. For a brief discussion on the general algorithm, see Chapter 5.

Chapter 5

Conclusion and Discussion

This thesis is a pilot study of an ongoing research project on the subsampling approach for robust inference for regression models that my supervisors, Dr. Min Tsao and Dr. Julie Zhou, have been working on. In this thesis, we have examined the most basic versions of subsampling proposals I and II. We will comment on the more elaborate versions that Dr. Tsao and Dr. Zhou have been working on in Section 5.2. But we first summarize the main findings of this thesis in the next section.

5.1 Summary of this thesis

The subsampling methods have the following advantages: (1) the underlying idea is simple and easy to apply in practice; (2) they are flexible and can be applied to various regression models including the linear regression models, the non-linear regression models and the generalized linear regression models; (3) it provides (nearly) unbiased estimates for examples that we have examined; (4) the estimation method Π in both SM1 and SM2 is usually the least-squares method or the maximum likelihood method and no complicated estimation method is required. Point (3) is a particularly good empirical property of the subsampling estimators and it typically holds whenever S_g

or S_g^2 is a good approximation to the set of good data points.

It should be noted that the subsampling methods assume that the regression model is correct. This justifies the subsampling algorithms which rely on the goodness-of-fit of the model to determine the presence of outliers. Without the assumption, a poor fit could be the consequence outliers or simply wrong model. The notion of outliers to the model is also not well defined. Hence the subsampling methods are recommended for situations where the model correctness is not a question.

When we apply the algorithms SA1(n_s, r^*, k) and SA2(n_s, k) to do the subsampling analysis, the criterion Γ needs to be selected carefully and it has to be a sensitive to the presence of outliers. In this thesis, we chose MSE as a criterion Γ in the analysis of the linear regression models and it worked well in examples and simulations in Chapters 3 and 4. Nevertheless, different criterion may need to be used for other types of regression models. Further, SM2 depends on an extra criterion τ to detect outliers and this may limit its use to some regression models.

For the subsampling method proposal I and II, there are several parameters, n_s, r^* and k, in the corresponding algorithms SA1(n_s, r^*, k) and SA2(n_s, k) which need to be determined first. Parameter n_s is the subsample size of each $S_{n_s}^1, \ldots, S_{n_s}^k$ and we often set it to be slightly larger than half of the total sample size which ensures relatively high breakdown points for SM1 and SM2. Parameter k is the total number of subsamples in the first step of each algorithm and it is determined by the value of r^* and the probability p^* which is associated with the number of outliers m and ensures at least r^* good subsamples in all k subsamples. In Chapter 3, we gave a function (3.14) to solve k. Parameter r^* in SA1(n_s, r^*, k) is the number of good subsamples to ensure the 99% efficiency. Through the determination of parameters of the algorithms, we can see that the number of outliers m plays an important role since the determination of r^* and k in SA1(n_s, r^*, k) highly depends on it. The subsampling

proposal II is also sensitive to the assumption of m, but it is not as sensitive as the proposal I when p^* is set to be very high, say $p^* = 0.9999$. In practice, since we do not know the exact number of outliers in a data set, we can use a couple of values of m to do the analysis, say $m = 0.1N$ and $m = 0.2N$, especially for SM1. If the results are the same or very similar, then each value of m is fine. If the results are different, we need to choose a larger m to run the algorithms again. If m is set to be larger than the actual number of outliers, the SM1 works well.

Both SM1 and SM2 are robust against x-outliers and y-outliers and they have high (theoretical) breakdown points (corresponding to $k = \infty$). In this thesis, the SM1 and SM2 are also compared with the robust MM-estimator for the linear regression models. The latter is highly efficient with high breakdown points. Our numerical results indicate that SM1 and SM2 are comparable to the MM-estimator. Since the MM-estimator for the intercept is biased and it may be difficult to interpret whereas the SM1 and SM2 estimators are unbiased from the simulation results, SM1 and SM2 provide very competitive robust estimation methods for practice applications.

5.2 Ongoing research on subsampling methods for robust inference

This thesis has presented the basic ingredients of the subsampling methods and demonstrated numerically that the subsampling estimators are attractive robust estimators. It is equally important to examine their theoretical properties such as unbiasedness, asymptotic distributions and quantitative robustness. Dr. Tsao and Dr. Zhou have been working on such properties but instead of focusing on their ongoing work on theoretical aspects of the subsampling estimators, we now describe the more general subsampling methods that they have been working on. To simplify

our discussion, we suppose we have the entire sequence of $k^* = \binom{n+m}{n_s}$ different subsamples. This corresponds to taking infinitely many ($k = \infty$) subsamples and then eliminating duplicating subsamples from this infinite set of subsamples.

5.2.1 General subsampling method proposal-I

The basic assumption of the subsampling method proposal-I in Chapter 3 is that the γ-ordered sequence of subsamples

$$S_{n_s}^{(1)}, S_{n_s}^{(2)}, \ldots, S_{n_s}^{(k^*)}$$

will have those subsamples free of outliers at the beginning of the sequence, and thus they may be taken union of to form the combined good sample. Unfortunately, while a small γ is a necessary condition for a subsample being free of outliers it is in general not a sufficient condition. Hence it is possible that outliers are present in the first several subsamples. An extreme example that illustrates this point is the case where the outliers are generated by the same type of model as the good data but with different parameter values. If n_s is small enough, then $S_{n_s}^{(2)}$, say, may contain only outliers and $\gamma_{(2)}$ is small because the model fits the outliers well.

To resolve this issue and to better define the notation of a good subsample, we need further criteria than the γ score. We now give an example based on the estimated parameters. Denote by $\hat{\beta}_i$ the estimated parameter vector based on the ith subsample in the γ-ordered sequence. Consider dividing $S_{n_s}^{(1)}, S_{n_s}^{(2)}, \ldots, S_{n_s}^{(k^*)}$ into two or more subsequences/groups using a secondary criterion, say, a distance measure for $\hat{\beta}_i$. If their γ scores as well as associated $\hat{\beta}_i$ values are all close, the sequence will not be divided. We can then take their union knowing that they are all consistent in their goodness-of-fit and their estimated $\hat{\beta}_i$ values. If γ scores and the associated $\hat{\beta}_i$ values

vary substantially from subsample to subsample, we may have either outliers in the original sample or a genuine mixture model situation where the data set come from two (or more) models which are the same in structure but differ in parameter values. In the former case, all those subsamples similar to $S_{n_s}^{(1)}$ in terms of both γ score and $\hat{\beta}_i$ may be taken union of to form a combined sample for final model estimation. In the latter case, further investigation of the data set may be necessary.

The secondary criterion can also help to maximize efficiency of the combined sample as it takes union of all subsamples that are similar to $S_{n_s}^{(1)}$ instead of a fixed number of r^* good subsamples. Note that the combined sample of a fixed number of r^* good subsamples may not be a random subsample of the good data but the combined sample of all those subsamples similar to $S_{n_s}^{(1)}$ may be more appropriately viewed as a random sample from the good data. This is important as the former combined sample is known to produce biased estimates due to the fact that it tends to pick those points of the good data that are most favorable to the model. Finally, we note that it is important that the combined sample be of high efficiency so that the resulting estimated model will be close to that based on all good data and hence the true underlying model. If the combined sample is only a small subsample of the good data, the estimated model based on the combined sample may be very different from the true model even when it contains no outliers.

5.2.2 General subsampling method proposal-II

The basic version of subsampling method-proposal II that we examined in Chapter 4 suffers from two problems. First, it may be conservative in picking up good data points and this may result in low efficiency. Second, it requires a criterion for deciding whether a single point is an outlier, and this is not always easily available. A more elaborate proposal-II that Dr. Tsao and Dr. Zhou have been investigating is described

below.

Step 1: Find the subsample with the smallest γ score, $S_{n_s}^{(1)}$, as before.

Step 2: Estimate model parameters with all points, in *and* outside of $S_{n_s}^{(1)}$, and compare the resulting estimated parameter values with those based on $S_{n_s}^{(1)}$ only. If the two sets of estimated values are close, declare that there are no outliers and use the estimates based on the entire data set.

Step 3: Suppose substantial difference in estimated parameters is found in step [2]. Then order the remaining points (those outside of $S_{n_s}^{(1)}$) using, say, their residuals given by the estimated model based on $S_{n_s}^{(1)}$; points with smaller ranks are more likely to be good points. Divide these remaining points into, say, two groups of say equal size, group 1 and group 2. Combine points in $S_{n_s}^{(1)}$ and those in group 1 and call it combined sample 1.

Step 4: Refit the model to the combined sample 1 and compare the estimated parameter values based on data points in $S_{n_s}^{(1)}$ with those based on combined sample 1. If no substantial difference in estimated parameter values is found, declare group 1 all good data points and combined sample 1 becomes the new larger good subsample. Subdivide group 2 into two groups using the rank and repeat steps 3 and 4.

Step 5: If substantial difference in estimated parameter values is found in step 4, declare group 2 are all outliers and remove them from further consideration. Subdivided group 1 into two groups and combine the first of the subgroup with $S_{n_s}^{(1)}$ and repeat step 4.

Step 6: Iterate the above steps until either the combined good subsample reaches a certain percentage, say, 90% of the original sample size or some other stopping

criterion is reached.

Like the basic version of the proposal-II, the above proposal is also anchored in a single good subsample $S_{n_s}^{(1)}$. But it also takes into consideration the secondary criterion based on the estimated parameter $\hat{\beta}_i$ which gives greater assurance of the consistency of the data points been added to $S_{n_s}^{(1)}$. It may also be more efficient than the testing a single point at a time approach of the basic proposal-II. So far, we have assumed that subsample $S_{n_s}^{(1)}$ contains good data only and that the estimated model based on this subsample is close to the true model that generated the data. This latter assumption may not hold especially when the subsample size n_s is small, even if $S_{n_s}^{(1)}$ contains only good data. When this latter assumption fails, proposal-II will fail. Research is ongoing to find ways to deal with this problem. One partial solution is a careful selection of n_s value which may reduce the chance of such a proposal-II failure due to a faulty $S_{n_s}^{(1)}$.

To conclude, Dr. Tsao and Dr. Zhou continue to study these and other subsampling proposals which make use of additional criteria (not just the γ score) to improve the consistency and efficiency of the basic proposals that we have examined in this thesis. With their improved subsampling proposals, the subsampling method may become a promising alternative for robust estimation of regression models.

References

Analytical Methods Committee (1989), Robust statistics - How not to reject outliers, *Analyst*, **114**, 1693-1702.

Beaton A.E. and Tukey J.W. (1974), The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics*, **16**, 147-185.

Berk, Richard A. (2004), *Regression Analysis: A Constructive Critique*. Sage Publications.

Bond, N.W. (1979), Impairment of shuttlebox avoidance-learning following repeated alcohol withdrawal episodes in rats, *Pharmacology, Biochemistry and behavior*, **11**, 589-591.

Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering*, 2nd ed., New York: Wiley.

Burnham, K.P., and Anderson, D.R. (2002), *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer.

Cantoni, E. and Ronchetti, E. (2001), Robust Inference for Generalized Linear Models. *Journal of American Statistical Association* **96**(455), 1022-1030.

Cressie, N. (1996), Change of support and the modiable areal unit problem, *Geographical Systems*, **3**, 159-180.

Dodge, Y. (1987), *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, North-Holland, Amsterdam

Donoho, D.L. and Huber, P.J. (1983), The notion of breakdown point, *A Festschrift for Erich L. Lehmann*, 157-184.

Freedman, David A. (2005), *Statistical Models: Theory and Practice*. Cambridge University Press.

Hampel, F.R. (1975), Beyond location parameters: Robust concepts and methods, *Bulletin of the International Statistical Institute*, **46**, 375-382.

Hampel, Frank R., Ronchetti, Elvezio M., Rousseeuw, Peter J., and Stahel, Werner A. (1986), *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hocking, R.R. (1985), *The Analysis of Linear Models*. Pacific Grove, California: Brooks/Cole.

Huber, P.J. (1964), Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, **35**,73-101.

Huber, P.J. (1981), *Robust Statistics*. New York: Wiley.

Jolliffe, I.T. (1986), *Principal Component Analysis*, New York: Springer- Verlag.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006), *Robust Statistics : Theory and Methods*. New York: Wiley.

McCullagh P. and Nelder, J.A. (1989), *Generalized Linear Models*. London, Chapman and Hall.

Montgomery D.C., Peck E.A. and Vining G.G. (2006), *Introduction to linear regression analysis*. 4th ed., New York: Wiley.

Pukelsheim, Friedrich (1994), The three sigma rule, *The American Statistician* **48**, 234 - 251.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Rousseeuw, P.J. and Yohai, V.J. (1984), Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series*, **26**, 256-272.

Salibian-Barrera, M. (2003), Fast and stable bootrap methods for robust estimates, *Computing Science and Statistics*, **34**, 346-359.

Salibian-Barrera, M. and Yohai, V.J. (2006), A fast algorithm for S-regression estimates, *Journal of Computational and Graphical Statistics*, in press.

Shao, Jun (2003), *Mathematical statistics*. Berlin: Springer-Verlag.

Stigler, S.M. (1977), Do robust estimators deal with real data? *The Annals of Statistics*, **5**, 1055-1098.

Wikipedia, (2009), http://en.wikipedia.org/wiki/Robustness.

Wolberg, J. (2005), *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*. Springer.

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