EXTENDED BOOTLIER PROCEDURE FOR DETECTION OF OUTLIERS IN UNIVARIATE SAMPLES AND LINEAR REGRESSION ANALYSIS

BY YI XIA

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Statistics and Biostatistics
Written under the direction of
Professor Minge Xie
And approved by

____

____

____

____

New Brunswick, New Jersey
October, 2014

ABSTRACT OF THE DISSERTATION

Extended Bootlier Procedure for Detection of Outliers in Univariate Samples and Linear Regression Analysis

By Yi Xia

Dissertation Director: Professor Minge Xie

Determining if a dataset has one or more outliers is a fundamental and challenging problem in statistical analysis. This dissertation introduces a statistical framework that addresses two well-known problems in the outlier analysis. The first problem (Problem 1) is to detect outliers in independent and identically distributed univariate samples, which is the basic setting of outlier problem. The second problem (Problem 2) is to detect outliers and influential observations in the linear regression analysis, which is a major topic in linear regression model diagnostics and represents a more complete setting.

The proposed framework is motivated by a graphic outlier detection method proposed recently for Problem 1. It is observed in bootstrapping that some bootstrap samples contain outliers while others do not, when outliers are present in a sample. Based on this observation, the method discovers that a bootstrap sample statistic (termed “mean – trimmed mean”) is sensitive to outliers, and particularly its histogram is multimodal in the presence of outliers. Consequently outliers are detected by plotting and visually checking the histogram. Considering that method captures the essence of outliers that the researches often call, the proposed framework further develops it to a complete inference procedure by constructing a formal statistical test based on a quantitative index that measures the degree of outliers effect. The proposed framework is first developed to address Problem 1. A procedure with a formal test is detailed and the large sample theory is developed to support the proposed procedure. Then, the procedure is extended to linear regression to address Problem 2. The measures for outliers and influential observations, including several residuals and a square-root version of Cook's distance, are discussed, and large sample theory is developed for such non-independent case. In addressing both problems, the simulation studies are conducted and real data examples are explored to show the wide-range application of the proposed framework. In particular, the comparison with other commonly used methods in the simulation studies demonstrates the overall advantage of the proposed framework.

Acknowledgements

First and foremost I would like to express my sincere appreciation to my advisor Professor Minge Xie for his insightful guidance in my research through the years. Without his continuous support and encouragement, my completion of this dissertation would not have been accomplished.

I would like to thank our graduate director, Professor John Kolassa, for his tremendous help during my study in Rutgers. I would also like to thank Professor Lee Dicker and Dr. Dayong Li for their time and effort to serve on my dissertation committee.

I am also thankful to Professor Kesar Singh, who passed away two years ago, for his valuable suggestions for my research. It is sad that I could not share this accomplishment with him. Last but not least, my thanks go to all faculty and staff at Department of Statistics and Biostatistics of Rutgers University for their support through these years.

Dedication

To my wife, Bin Xue, and lovely sons, Daniel and Dylan

Table of Contents

CHAPTER 1

INTRODUCTION

Outliers present a fundamental problem in statistical data analysis. Although many statistical theories and methods for detection and handling of outliers have been proposed, it still remains an interesting research topic. Among the recent developments, Singh and Xie (2003) propose a non-parametric and graphic method for the detection of outliers in the independent and identically distributed (i.i.d.) univariate sample, the basic setting of the outlier problem. Considering that method captures the essence of outliers that researchers often call, this dissertation proposes a statistical framework that extends the graphic method to a more complete inference procedure by constructing a formal test based on a quantitative measure to detect outliers in the i.i.d. univariate sample. Another more complicated and well-known outlier problem is to detect outliers and influential observations in linear regression analysis, which Barnett and Levis (1994) describe as a "structured data" case. With the modifications, the proposed statistical framework is extended to the regression setting to detect outliers and influential observations by analyzing different residuals and influential measures. In summary, this dissertation is aiming to solve the following two well-defined problems.

- **Problem 1**. To detect outliers in the independent and identically distributed univariate data
- **Problem 2**. To detect outliers and influential observations in linear regression analysis

Detection of outliers in i.i.d. univariate data is a basic problem in the analysis of outliers. The methods commonly used in the real data analysis are developed mostly in the last century. Examples of these outliers detection methods include the well-known box-plot and interquartile range (IQR) method with a classification of outlier as mild/extreme ones (NIST/SEMATECH, 2012), the modified IQR method (Barbato et al.

2011), Grubbs' test (Grubb 1969), a general extreme studentized deviate (ESD) test (Rosner 1983), and Dixon's Q test (Dixon 1950, 1951). Barbato et al. (2011) provide a complete review of these outlier detection methods. Although these methods are often used in the exploratory setting without considering the underlying distribution of the sample, they are developed under the assumption of normally distributed data. Therefore they often exhibit higher false positive rate when normality assumption does not hold. Other limitations include not taking sample size into account or only applying to small samples. For example, IQR method is not adjusted to sample size and Dixon's Q test is appropriate for sample size no more than 40. Among the outlier detection methods proposed in the recent years, Bootlier plot (Singh and Xie 2003) provides a non-parametric and graphic way to detect outliers using the bootstrapping technique. Singh and Xie (2003) prove that the limiting distribution of a bootstrap sample statistic "mean – trimmed mean" is expressed as a mixture of normal distributions with multiple modes when the sample has outliers. Therefore detecting outliers is equivalent to checking multimodality in the density plot (a bootstrap histogram) of that bootstrap sample statistic. By plotting and checking the bumpiness of the density plot, one can infer the presence of outliers in the sample. A quantitative measure "Bootlier index" is introduced by Singh and Xie (2003) to measure the degree of bumpiness of the density plot, and is used to screen multiple plots to identify those bumpy ones. The empirical thresholds are suggested, but no further utilization of this index is discussed.

The identifying outliers and influential observations in linear regression analysis, as a major topic of linear regression model diagnostics, is a more complicated problem in the outlier analysis. There is, of course, a vast literature on the detection of outliers and influential observations in linear regression analysis; see Beckman and Cook (1983), and Barnett and Levis (1994). The topic is present in almost every textbook of linear regression model. The current methods are generally classified into two groups, namely the graphical and the analytical methods. The graphic methods usually display different statistics measuring the degree of departure of outliers or influential observations from other data points. These graphic methods include scatter plot, residual plot, box-plot and normality plot. The analytical methods include various discordancy tests based on different residuals and influence measures. Chatterjee and Hadi (1986) have a complete review of the measures and discordancy tests based on those measures. Some discordancy tests are discussed in Barnett and Levis (1994).

In this dissertation, we develop a statistical inference framework, named extended Bootlier procedure, to address the above two problems.

- To address the first problem, we obtain the density plot of the bootstrap sample statistic “mean – trimmed mean” and Bootlier index of the density plot (referred as sample Bootlier index hereafter) from the given data. We then construct a formal statistical test for the presence of outliers using sample Bootlier index as the test statistic. Assuming the underlying distribution of the data is known, we estimate the null distribution of sample Bootlier index by a simulation method, and obtain *P*-value for the test. When the underlying distribution is unknown, the test is repeated assuming data is from several representative distributions to provide the reference lines. We illustrate the wide-range applications of extended Bootlier procedure by simulation studies and two real data examples. In particular, the comparison with other commonly used outlier detection methods in the simulation studies demonstrates the overall advantage of extended Bootlier procedure. The large sample theory that generalizes the results of Singh and Xie (2003) is developed to support the proposed framework. The general results also answer a question for the choice of bootstrap sample size when the sample has multiple outliers, which is not addressed in the original results. These developments are presented in Chapter 2.

- To address the second problem, we extend the proposed procedure in Chapter 2 to the residuals and influential measures from regression model fitting. The outliers and influential observations are detected by analyzing the outliers in the residuals and influential measures. In the analysis of outliers, the ordinary residuals, studentized residuals, and studentized deletion results are suggested, while a square-root version of Cook’s distance (SRCD) is proposed to analyze influential observations. While the large sample theory that explains the association between the sample Bootlier index and the outliers is provided for i.i.d. univariate data in Chapter 2, we develop similar results for the residuals and SRCD because they are dependent. The proposed framework is illustrated through simulation studies and a real data example to show its usefulness. In particular, we show that the extended Bootlier procedure has a lower false positive rate compared with other commonly used methods when the distribution of error terms deviates from normal with high probability in its tail. These developments are presented in Chapter 3.

- To facilitate the use of proposed extended Bootlier procedure, we develop R/C functions in Chapter 4.

CHAPTER 2

EXTENDED BOOTLIER PROCEDURE – A BOOTSTRAP-BASED METHOD FOR OUTLIERS DETECTION

ABSTRACT

Detecting outliers in a sample is a fundamental and challenging problem in statistical analysis. This chapter introduces a bootstrap-based statistical framework to detect outliers in an independent and identically distributed univariate sample. The framework extends a graphic outlier detection method, the Bootlier plot, to a complete inference procedure by constructing a formal statistical test based on a quantitative index that measures the degree of outlier effect. The large sample theory for general case of multiple outliers is developed as the support of the proposed framework, which also addresses the issue of bootstrap sample size selection. The proposed framework is illustrated through simulation studies for various scenarios which include multiple outliers, heavy-tailed distributions and large samples to show its wide-range applications. In particular, the comparison with other common outlier detection methods in the simulation studies shows the overall advantages of proposed framework. Two real data examples which include temperature data of space shuttle Challenger and natality data are explored to show its usefulness.

Keywords: outlier; bootstrap; density estimation; mean – trimmed mean; large sample theory, Bootlier index

2.1 INTRODUCTION

This chapter develops a bootstrap-based statistical framework for detecting outliers in an independent and identically distributed (i.i.d.) univariate sample. Suppose we have a sample $Y = \{Y_1, Y_2, \cdots, Y_n\}$ from a distribution F with finite variance ($\sigma^2 < \infty$) that may or may not contain outliers. A basic observation in bootstrapping is that, when outliers are present in Y, some bootstrap samples contain outliers while others do not. Motivated by this observation, Singh and Xie (2003) discover that a bootstrap sample statistic presents multimodality in its density plot (a bootstrap histogram), and propose a graphic outliers detection method (Bootlier plot) by drawing the density plot of the statistic and checking its multimodality to infer the presence of outliers. Using the fact that this method captures the essence of outliers, this chapter further develops a formal testing procedure based on a quantitative index that measures multimodality of the density plot for detection of outliers. The large sample theory is developed for general case of multiple outliers, which generalizes the results developed by Singh and Xie (2003). These are the main focuses of this chapter.

Detection of outliers is a fundamental problem in statistical analysis. While the statistical theories and methodologies aiming at drawing valid inference irrespective of the presence of outliers, as one of two main focuses of handling of outliers according to Barnett and Levis (1994), evolve remarkably in the past few years, those for detection of outliers, as the other focus, do not have such rapid growth. The methods commonly used in the real data analysis nowadays are developed mostly in the last century. Examples of these outlier detection methods include the well-known box-plot and interquartile range (IQR) method with a further classification of outlier as mild/extreme ones (NIST/SEMATECH 2012), the modified IQR method (Barbato et al. 2011), Grubbs' test (Grubb 1969), a general extreme studentized deviate (ESD) test (Rosner 1983), and Dixon's Q test (Dixon 1950, 1951). Barbato et al. (2011) provides a complete review of these outlier detection methods. Although these outlier detection methods are often used in the exploratory setting without considering the underlying distribution of the sample, they are developed assuming that the data are normal. Therefore they often exhibit higher false positive rate when the normality assumption does not hold. Some of these methods (for example, the IQR method) do not taking sample size into account, and others (for example, Dixon's Q test) are appropriate for sample sizes no more than 40.

Among the outlier detection methods proposed in the recent years, the Bootlier plot (Singh and Xie 2003) provides a non-parametric and graphic way to detect outliers in the i.i.d. univariate sample utilizing the bootstrap technique. Suppose $\{Y_1^*, Y_2^*, \cdots, Y_n^*\}$ is a bootstrap sample from $\{Y_1,\ Y_2, \cdots,\ Y_n\}$. The Bootlier plot method discovers that a sample statistic of $\{Y_1^*, Y_2^*, \cdots, Y_n^*\}$, named "mean – trimmed mean" (referred to as MTM hereafter), presents multimodality (or bumpiness in layman's language) in its density plot when the data has outliers, and we draw a number of bootstrap samples. Therefore by plotting and checking the bumpiness of the density plot, one can infer the presence of outliers in the sample. Singh and Xie (2003) prove that the limiting distribution of MTM can be expressed as a mixture of normal distributions when a single outlier is present in the data. However, the general case of multiple outliers is not discussed. A quantitative measure "Bootlier index" is introduced by Singh and Xie (2003) to measure the degree of bumpiness of a density, and is used to screen multiple density plots to identify those bumpy ones, but no further utilization of this index is discussed.

Considering that Bootlier index provides a good measure of multimodality of the density plot, in this research we propose a statistical inference framework, named extended Bootlier procedure, using Bootlier index of the density plot of MTM (referred as sample Bootlier index) as the test statistic to detect outliers in the sample. By evaluating observed sample Bootlier index by its distribution under the hypothesis of no outliers, we obtain P-values that quantify the significance of outliers. Together with the density plot, we can conclude whether there are outliers in the sample. The simulation studies show the good testing power of this procedure and the comparison with other commonly used outlier detection methods demonstrates the overall advantage of the extended Bootlier procedure. While large sample theory which explains multimodality of the density plot caused by outliers is proven when a single outlier is present in the sample by Singh and Xie (2003), in this research we extend the results to general case of multiple outliers. The general results also answer a question for the choice of bootstrap sample size when the data has multiple outliers, which is not addressed in the original results.

The rest of this chapter is organized as follows. Section 2.2 briefly reviews the Bootlier plot method. The proposed extended Bootlier procedure is introduced in Section

2.3 and the large sample theory is developed in Section 2.4. In section 2.5, several simulation studies are conducted to illustrate the performance of the extended Bootlier procedure for various scenarios, such as the samples with multiple outliers, the samples from heavy-tailed distributions and large samples. The testing power of extended Bootlier procedure and the comparison with other commonly used outlier detection methods are also present in Section 2.5. Section 2.6 presents two real data examples. Finally, concluding remarks are present in Section 2.7.

2.2 A REVIEW OF BOOTLIER PLOT METHOD (Singh and Xie 2003)

The Bootlier plot method is illustrated as follows.

Let $\left\{Y_{(1)}^*, Y_{(2)}^*, \cdots, Y_{(n)}^*\right\}$ be the order statistics of a bootstrap sample $\{Y_1^*, Y_2^*, \cdots, Y_n^*\}$ from $\{Y_1,\ Y_2, \cdots,\ Y_n\}$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $\bar{Y}_n^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*$. MTM of the bootstrap sample with trimming size k_0 is obtained as,

$$\text{Upper-sided trimming: MTM} = \bar{Y}_n^* - \frac{1}{n-k_0}\sum_{i=1}^{n-k_0} Y_{(i)}^* \tag{2.1}$$

$$\text{Lower-sided trimming: MTM} = \bar{Y}_n^* - \frac{1}{n-k_0}\sum_{i=k_0+1}^{n} Y_{(i)}^* \tag{2.2}$$

$$\text{Two-sided trimming: MTM} = \bar{Y}_n^* - \frac{1}{n-2k_0}\sum_{i=k_0+1}^{n-k_0} Y_{(i)}^* \tag{2.3}$$

Singh and Xie (2003) prove that the limiting distribution of MTM is expressed as a mixture of normal distributions with multiple modes when the sample has outliers. Therefore detecting outliers in $\{Y_1,\ Y_2, \cdots,\ Y_n\}$ is equivalent to checking multimodality in the density plot of MTM. Two factors in the above notations need to explain. First, the trimming direction is determined by which side(s) of outlier one likes to investigate. The MTM by (2.1), (2.2) and (2.3) is sensitive to the outliers in the upper side, lower side, and both sides respectively. For two-sided trimming (2.3), the same number of data points is trimmed for simplicity. Second, the trimming size k_0 is considered as a smoothing factor for the density of MTM. Singh and Xie (2003) prove that the separation of two possible largest modes in the density of MTM is approximately proportional to $\frac{1}{k_0}$ when the sample has outliers and provide the some empirical suggestions, for example $k_0 = 2$ for two-sided trimming.

Singh and Xie (2003) also introduce a "Bootlier index" to measure the degree of bumpiness of a density. If M denotes the global mode of a density $g(x)$ for a distribution, Bootlier index $BI(g)$ is defined as,

$$BI(g) = \int_{M}^{+\infty}\left(sup_{y\geq x}g(y) - g(x)\right)dx + \int_{-\infty}^{M}\left(sup_{y\leq x}g(y) - g(x)\right)dx \qquad (2.4)$$

Bootlier index measures the valley area of a density with multiple modes. An example is illustrated in Figure 2.1 for a density function as a mixture of three normal densities, $g(x) = \frac{1}{2}\phi(x) + \frac{1}{3}\phi(x-3) + \frac{1}{6}\phi(x-9)$, where $\phi(\cdot)$ denotes the standard normal density function. Clearly Bootlier index is 0 for unimodal density functions, while the densities with multiple modes have non-zero values depending on the degree of separation between modes. The original idea for Bootlier index is to screen out bumpy-free plots when one is examining multiple density plots to pinpoint those bumpy ones. As suggested by Singh and Xie (2003), a Bootlier index of 0.1 can be considered as a bumpy plot, and a Bootlier index between 0.01 and 0.1 is considered as borderline cases.

[Insert Figure 2.1]

Considering that Bootlier index provides a good measure of multimodality of a density, in this research we propose a statistical inference framework based on Bootlier index to detect outliers. In what follows we introduce the framework.

2.3 EXTENDTED BOOTLIER PROCEDURE

In this section, we extend the Bootlier plot method to a complete inference procedure with formal testing. First we follow what Bootlier plot suggests to draw a large number of bootstrap samples and obtain a set of MTM. Second we obtain the density plot using the kernel density estimation method and obtain the Bootlier index of the density plot (i.e., sample Bootlier index). Third we construct a test for the presence of outliers using the sample Bootlier index as the test statistic given F is known. When F is unknown, the test is repeated assuming sample is from several representative distributions to provide the reference lines.

Suppose m bootstrap samples are drawn from $\{Y_1, Y_2, \cdots, Y_n\}$. The set of MTM is obtained as $\{MTM_1, MTM_2, \cdots, MTM_m\}$. The density plot is expressed as $\hat{f}_{MTM}(x)$ using the kernel density estimation method,

$$\hat{f}_{MTM}(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x-MTM_i}{h}\right) \tag{2.5}$$

where $K(.)$ is a kernel function and h is the bandwidth. The kernel density estimation is a non-parametric method introduced by Rosenvblatt (1956) and Parzen (1962), and widely used for density estimation problem; see Silverman (1986) and Sheather (2004). For our method, $K(.)$ is chosen to be the standard normal density function, and h is chosen to be Silverman's bandwidth, which is one of default selections in the R statistical software package. Although Singh and Xie (2003) pointed out, "it is clear that multimodality in a bootlier plot is a feature caused by outliers, not by selection of bandwidth", the impact of bandwidth to sample Bootlier index is apparent. For example, the bandwidth suggested by Jones et al., (1996) may oversmooth the curve to yield a smaller value of sample Bootlier index. The impact of bandwidth to the inference (which is discussed below) is an interesting area for future study. It would be interesting to compare results for other commonly used bandwidth selection methods (see Sheather, 2004) when the data comes from different distributions. The sample Bootlier index is calculated as $BI(\hat{f}_{MTM})$ using a numerical method.

While the Bootlier plot (Singh and Xie 2003) detects outliers by plotting $\hat{f}_{MTM}(x)$ and visually checking the bumpiness of the curve, we construct a statistical formal test for the presence of outliers. Given F is known, the test is formulated as the following hypothesis,

H_0: $\{Y_1, Y_2, \cdots, Y_n\}$ is from F without outliers

H_1: $\{Y_1, Y_2, \cdots, Y_n\}$ is from F with outliers

To test the hypothesis, we consider sample Bootlier index $BI(\hat{f}_{MTM})$ as the test statistic that is function of $\{Y_1, Y_2, \cdots, Y_n\}$. We denote the distribution function of $BI(\hat{f}_{MTM})$ under H_0 (null distribution) by F_0. The P-value for this test is expressed as $Prob = 1 - F_0(BI(\hat{f}_{MTM}))$. Since it is difficult to obtain the closed form of F_0, we propose a simulation approach to estimate it instead. Suppose N independent samples with each sample having n independent values are drawn from F. The sample Bootlier index is obtained for each independent sample using the same step as $BI(\hat{f}_{MTM})$, and is denoted by BI_i for $i = 1,2,\ldots,N$. The P-value is then computed as $\widehat{Prob} = \frac{1}{N}\sum_{i=1}^{N} 1_{(BI(\hat{f}_{MTM}))\leq BI_i)}$. While F is unknown, we propose to obtain P-values assuming $\{Y_1, Y_2, \cdots, Y_n\}$ from several representative distributions that range from short-tailed distributions to heavy-tailed distributions, which include uniform, normal, student t_6, exponential and Cauchy distribution. The purpose is to provide the reference lines so that one can make the conclusion right away, for example, when none of P-values is significant or all of P-values are significant, or investigate the data more to draw firm conclusion.

Combining the above steps we obtain an integrated statistical method, named “extended Bootlier procedure”. The procedure can be summarized as follows.

1. Draw m bootstrap samples from data $\{Y_1, Y_2, \cdots, Y_n\}$ with each bootstrap sample of size n, and compute the sample statistic MTM as $\{MTM_1, MTM_2, \cdots, MTM_m\}$.

2. Obtain $\hat{f}_{MTM}$ and $BI(\hat{f}_{MTM})$.

3. Estimate the null distribution for $BI(\hat{f}_{MTM})$ and obtain P-value(s).

   (a) If F is known, draw N independent samples with each sample having n independent values drawn from F.

   (b) Obtain BI_i for $i = 1,2,\ldots,N$ for each bootstrap sample by repeating 1&2.

   (c) Obtain P-value as $\frac{1}{N}\sum_{i=1}^{N} 1_{(BI(\hat{f}_{MTM})\leq BI_i)}$.

(d) If F is unknown, repeat 3(a) to 3(c) to obtain P-values assuming the data from several reference distributions that include uniform, normal, student t_6, exponential, Cauchy distribution, and bimodal distribution (for example, a mixture of two normal distributions with a density $g(x) = \frac{1}{2}\phi(x-1.5) + \frac{1}{2}\phi(x+1.5)$) .

2.4 LARGE SAMPLE THOERY

In this section, we provide the results for limiting distribution of MTM for the general case of multiple outliers, which is the extension of the results proven by Singh and Xie (2003) when one outlier is present in the sample. The main purpose is to explain why the density plot of MTM is sensitive to outliers in terms of multimodality and to provide support for the extended Bootlier procedure. Under further imposed conditions for outliers, the results also address the problem of the choice of bootstrap sample size when multiple outliers are present in the sample.

Suppose there are no outliers in $\{Y_1, Y_2, \cdots, Y_n\}$ from a distribution F with finite variance ($\sigma^2 < \infty$), and L outliers, $\xi = \{\xi_1, \xi_2, \cdots, \xi_L\}$, are added in upper side to the sample. Without loss of generality, we assume variance $\sigma^2 = 1$.

Let $\{Y_1^*, Y_2^*, \cdots, Y_{n+L}^*\}$ be a bootstrap sample from $\{Y_1, Y_2, \cdots, Y_n, \xi_1, \xi_2, \cdots, \xi_L\}$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $\bar{Y}_{n+L}^* = \frac{1}{n+L}\sum_{i=1}^{n+L} Y_i^*$. We let $G_n(x) = P(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \leq x)$, and denote the number of times for any elements from ξ appearing in $\{Y_1^*, Y_2^*, \cdots, Y_{n+L}^*\}$ by $m(\xi)$.

For any fixed $i \geq 1$, let $S_i(\xi) = \{S_{i,j}(\xi) | S_{i,j}(\xi) = (\xi_{j_1}, \xi_{j_2}, \cdots \xi_{j_i}), j = 1,2,..,L^i\}$ be the set of all configurations of choosing i elements from $\{\xi_1, \xi_2, \cdots, \xi_L\}$ with replacement. We denote the mean for each configuration by $\overline{S_{i,j}}(\xi) = \frac{1}{i}\sum_{m=1}^{i} \xi_{j_m}$ for $j = 1,2,..,L^i$. For notation purpose, we set $\overline{S_{0,1}}(\xi) = 0$.

The following theorem and corollary show that the limiting distribution of bootstrap sample mean is not sensitive to outliers.

THEOREM 2.1

Under the above setting, the cumulative distribution function $G_n(x)$ can be expressed as the mixture

$$\sum_{i=0}^{K} \frac{e^{-L}}{i!} \sum_{j=1}^{L^i} \Phi(x - \frac{i\overline{S_{\iota,J}}(\xi)}{\sqrt{n+L}}) + r_{n,K}, \tag{2.6}$$

where $\limsup_n |r_{n,K}| \le P(\mathcal{P}_L > K)$, a.s., K is a positive integer, and $\mathcal{P}_L$ is a Poisson random variable with mean L.

Proof of Theorem 2.1 is provided in Appendix A.2.

COROLLARY 2.2

If $\xi_{(L)}/\sqrt{n} \to 0$ as $n \to \infty$, we have the limiting result $G_n(x) = \Phi(x) + o(1)$ a.s.. That is, unless $\xi_{(L)}$ increases at a rate of $\sqrt{n}$ or faster as $n \to \infty$, its effect on the distribution of the normalized bootstrap mean vanishes in limit.

Theorem 2.1 and Corollary 2.2 are multiple-outlier versions of the limiting distribution of bootstrap sample mean. It reveals that the bootstrap sample mean converges to standard normal distribution even when there are outliers in the sample.

Next we turn to the Bootlier statistic MTM with the upper trimming with size 1. One can find out that,

$$\frac{1}{n+L-1} \sum_{i=1}^{n+L-1} Y_{(i)}^* - \bar{Y}_{n+L}^* = \frac{1}{n+L-1} \left(\bar{Y}_{n+L}^* - Y_{(n+L)}^*\right)$$

It is sufficient to study the distribution of normalized $\bar{Y}_{n+L}^* - Y_{(n+L)}^*$.

Let $T_n = \sqrt{n+L}(\bar{Y}_{n+L}^* - Y_{(n+L)}^* - \bar{Y}_n)$, and let $H_n(x) = P(T_n \le x)$ that is further expressed as the mixture

$$H_n(x) = \lambda_{n,0} H_{n,0}(x) + \lambda_{n,1} H_{n,1}(x),$$

where $\lambda_{n,0} = 1 - \lambda_{n,1} = e^{-L} + o(1)$, $H_{n,0}(x) = P(T_n \le x | m(\xi) = 0)$, and $H_{n,1}(x) = P(T_n \le x | m(\xi) \ge 1)$.

By above notations, Bootlier statistic MTM has a mixture of distributions $H_{n,0}$ and $H_{n,1}$, where $H_{n,0}$ is free of outliers and $H_{n,1}$ involves of outliers. The next theorem shows the representation of $H_{n,0}$ and $H_{n,1}$ as sample size increases to infinity. The mixture of distributions presents multimodality driven by the outliers.

THEOREM 2.3

Under the above setting, the following representations hold for any positive integer K with $n > K$:

$$\text{(a)}\quad H_{n,0}(x) = \sum_{i=1}^{K}(e^{-i+1} - e^{-i})\,\Phi\left(x + \sqrt{n+L}Y_{(n-i+1)}\right) + s_{n,K}, \tag{2.7}$$

where $\limsup_n |s_{n,K}| \le e^{-K}$.

(b) $H_{n,1}(x)$

$$= \sum_{i=1}^{K} \frac{L}{(e^{L}-1)i!} \sum_{j=1}^{L}\left(\left(\frac{j}{L}\right)^{i} - \left(\frac{j-1}{L}\right)^{i}\right) \sum_{k=1}^{L^{i-1}} \Phi\left(x + \sqrt{n+L}\xi_{(j)} - \frac{\xi_{(j)} + (i-1)w_{i-1,k}}{\sqrt{n+L}}\right) + t_{n,K}\ , \tag{2.8}$$

where $\limsup_n |t_{n,K}| \le \frac{1}{1-e^{-L}} P(\mathcal{P}_L > K)$ and $w_{i-1,k} = \overline{S_{\iota-1,k}}\left(\{\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(j)}\}\right)$ for $k = 1,2,..,L^{i-1}$.

Proof of Theorem 2.3 is provided in Appendix A.2. This result reveals the impact of outliers to the limiting distribution of MTM and the possible modes caused by outliers. However the representations are complicate. The messages are much clear if we impose additional conditions on ξ in the following corollary.

COROLLARY 2.4

Under the same setting as Theorem 2.3 and assuming $\xi_{(L)}/\sqrt{n} \to 0$, Theorem 2.3 (a) holds and (b) has the following representation for any positive integer K with $n > K$:

$$\text{(b)}\quad H_{n,1}(x) = \sum_{i=1}^{L} \frac{e^{i} - e^{i-1}}{e^{L}-1}\,\Phi\left(x + \sqrt{n+L}\xi_{(i)}\right) + t_{n,K}\ , \tag{2.9}$$

where $\limsup_n |t_{n,K}| \le \frac{1}{1-e^{-L}} P(\mathcal{P}_L > K)$.

Corollary 2.4 is the immediate result from Theorem 2.3. When there are no outliers increasing at rate $\sqrt{n}$ or faster, which is true if ξ is drawn from some distribution with finite variance, the possible modes caused by outliers are $\sqrt{n+L}\xi_{(i)}$ for $i = 1,2,\dots,L$, and those are separated from the possible modes in $H_{n,0}$, $Y_{(n-i+1)}$ for $i = 1,2,\dots,K$.

For a more special case, $\sqrt{n}(Y_{(n)} - Y_{(n-i)}) \to 0$, a.s. for any fixed i, which holds for short-tailed distributions (Singh and Xie 2003), Theorem 2.3 (a) can be simplified as,

$$H_{n,0}(x) \to \Phi\left(x + \sqrt{n+L}Y_{(n)}\right) . \tag{2.10}$$

Then $H_n(x)$ has the representation as,

$$H_n(x) \to e^{-L}\Phi\left(x + \sqrt{n+L}Y_{(n)}\right) + (1 - e^{-L}) \sum_{j=1}^{L} \frac{e^j - e^{j-1}}{e^L - 1} \Phi\left(x + \sqrt{n+L}\xi_{(j)}\right). \tag{2.11}$$

From (2.11), the message is clear that the normalized statistic T_n has a limiting distribution of a mixture of standard normal distributions. The bumpiness of the density plot of MTM and thus Bootlier index of the density plot increase as the magnitude of outliers increases. The (2.11) also reveals a fact that the density plot gets smoother when there are multiple outliers. If we let $\xi_1 = \cdots = \xi_L = \eta$, (2.11) is further simplified as

$$H_n(x) \to e^{-L}\Phi\left(x + \sqrt{n+L}Y_{(n)}\right) + (1 - e^{-L})\Phi(x + \sqrt{n+L}\eta). \tag{2.12}$$

When there is a single outlier, the weights of $H_{n,0}(x)$ and $H_{n,1}(x)$ in the mixture are 0.37 and 0.63 respectively, which easily results in the bumpiness of density plot. While the number of outliers L increases, the weight of $H_{n,0}(x)$ decreases rapidly. For a moderate L, say $L = 5$, the weight of $H_{n,0}(x)$ is only $e^{-5} \approx 6.7 \times 10^{-3}$ so that $H_{n,1}(x)$ dominates the mixture. The density plot is nearly smooth in this case. This phenomenon has been noticed by Singh and Xie (2003) and they propose a simple remedy approach to reduce the bootstrap sample size to a fraction of sample size n, say $[rn]$, where $0 < r \leq 1$. If we repeat the same arguments for Theorem 2.3, Corollary 2.4 and the results we have so far, the following holds,

$$H_n(x) \to e^{-rL}\Phi\left(x + \sqrt{n+L}Y_{(n)}\right) + (1 - e^{-rL})\Phi\left(x + \sqrt{n+L}\eta\right). \tag{2.13}$$

The choice of r is suggested as $\frac{\log(2)}{L}$ by (2.13), where $\log(\cdot)$ standards for natural logarithm. If the number of outliers is unknown for a given problem, one can apply the extended Bootlier procedure using different bootstrap sample size of $[rn]$. Any density plot that exhibits significant modality indicates the presence of outliers.

Theorem 2.3 and Corollary 2.4 study the case of trimming size 1. When the trimming size is more than 1, say $k_0 > 1$, there are too many terms involved in the mixture normal representation of the limiting distribution under the setting of Theorem 2.3. But if we add the additional constrains as discussed above, say $\sqrt{n}(Y_{(n)} - Y_{(n-j)}) \to 0$ a.s. and $\eta/\sqrt{n} \to 0$, the limiting distribution of MTM has a simplified form that allows us to study the impact of trimming size k_0 to the multimodality.

Considering MTM with the upper trimming with size $k_0 > 1$, we have,

$$\frac{1}{n+L-k_0}\sum_{i=1}^{n+L-k_0} Y_{(i)}^* - \bar{Y}_{n+L}^* = \frac{k_0}{n+L-k_0}\left(\bar{Y}_{n+L}^* - \frac{1}{k_0}\sum_{i=n+L-k_0+1}^{n+L} Y_{(i)}^*\right).$$

Then $T_n = \sqrt{n+L}(\bar{Y}_{n+L}^* - \frac{1}{k_0}\sum_{i=n+L-k_0+1}^{n+L} Y_{(i)}^* - \bar{Y}_n)$, and $H_n(x) = P(T_n \le x)$.

The following theorem reveals the representation of H_n as the sample size increases to infinity, similar to (2.13).

THEOREM 2.5

Under the same setting as Theorem 2.3, and further assuming $\sqrt{n}(Y_{(n)} - Y_{(n-j)}) \to 0$ a.s. for any fixed j, $\xi_1 = \xi_2 = \cdots = \xi_L = \eta$ and $\eta/\sqrt{n} \to 0$, $H_n(x)$ has the following limiting distribution for trimming size of k_0,

$$H_n(x) \to \sum_{i=0}^{k_0-1}\frac{e^{-L}L^i}{i!}\Phi\left(x + \sqrt{n+L}\,\frac{(k_0-i)Y_{(n)}+i\eta}{k_0}\right) + (1 - \sum_{i=0}^{k_0-1}\frac{e^{-L}L^i}{i!})\Phi\left(x + \sqrt{n+L}\eta\right). \tag{2.14}$$

The proof of Theorem 2.5 follows the same arguments as Theorem 2.3 by partitioning the space $(T_n \le x)$ into $(T_n \le x) \cap (m(\xi) = i)$ for $i = 0,1,\cdots,(k_0-1)$, and $(T_n \le x) \cap (m(\xi) \ge k_0)$. We do not repeat here.

By (2.14), the potential modes of $H_n(x)$ are $\sqrt{n+L}\frac{(k_0-i)Y_{(n)}+i\eta}{k_0}$ for $i=0,1,\cdots,k_0$. They spread evenly between $\sqrt{n+L}Y_{(n)}$ and $\sqrt{n+L}\eta$ by an equal distance of $\frac{\sqrt{n+L}}{k_0}\left(\eta-Y_{(n)}\right)$ between two adjacent modes. When k_0 is very large, the multimodality of $H_n(x)$ will vanish because the distance between modes gets too small. When $k_0=1$, the mixture distribution is the same as (2.12). Therefore k_0 is considered as a smoothing factor and is associated with how much separation between the outliers and the rest of samples we would assume. Although the original Bootlier plot recommends "$k_0=3$ or 4 for sample size 15 to a few hundreds" based on empirical calibration, we notice that the extended Bootlier procedure is less sensitive to k_0 because the sample Bootlier index is evaluated by the null distribution obtained using the same trimming size. Thus, we use $k_0=2$ in the simulation studies and real data examples. The modes are not solely determined by $\sqrt{n+L}\frac{(k_0-i)Y_{(n)}+i\eta}{k_0}$ for $i=0,1,\cdots,k_0$. The coefficient of each normal component, $\frac{e^{-L}L^i}{i!}$ for $i=0,1,\cdots,k_0-1$ and $(1-\sum_{i=0}^{k_0-1}\frac{e^{-L}L^i}{i!})$, is another factor contributing to the mixture distribution. As the discussion following (2.13), some key components will have little effect to the mixture distribution due to small coefficient value, which results in the smooth density. The remedy approach proposed for (2.13), i.e. reducing the bootstrap sample size to $[rn]$, applies for $k_0>1$ as well. Repeating the same arguments, we have

$$H_n(x)\to$$

$$\sum_{i=0}^{k_0-1}\frac{e^{-rL}(rL)^i}{i!}\Phi\left(x+\sqrt{n+L}\frac{(k_0-i)Y_{(n)}+i\eta}{k_0}\right)+(1-\sum_{i=0}^{k_0-1}\frac{e^{-rL}(rL)^i}{i!})\Phi\left(x+\sqrt{n+L}\eta\right). \tag{2.15}$$

Because what we are interested in is the separation of outliers with the rest of sample, we suggest $r=\frac{\log(k_0+1)}{L}$ to let the coefficient of $\Phi\left(x+\sqrt{n+L}Y_{(n)}\right)$ about $\frac{1}{k_0+1}$. Practically one could examine the potential number of outliers first, for example, checking the largest gap among the extreme observations, to get the good estimation of L and then apply to the extended Bootlier procedure.

2.5 SIMULATION STUDIES

In this section, we conduct four simulation studies to illustrate the performance and features of the extended Bootlier procedure on various data cases including samples with multiple outliers, samples from heavy-tailed distributions and large samples. In particular, we compare the performance of the extended Bootlier procedure with other commonly used outlier detection methods for each simulation study to demonstrate the overall advantage of extended Bootlier procedure. The details of those commonly used outlier detection method are provided in Appendix A.1. At the end of this section, another simulation study is performed to access the testing power of Bootlier procedure.

In the simulation study #1, a sample of size 25 is simulated from standard normal distribution with three severe outliers, 3.4, 3.5 and 3.6, added to the sample. The total number of data points is 28. The sample data is plotted in Figure 2.2 (a). The density plot, sample Bootlier index and P-value for normal distribution are presented in Figure 2.2 (b) for investigating the outliers in upper side. To evaluate the performance, we also apply the extended Bootlier procedure with a single outlier 3.5 added to the sample. The results are presented in Figure 2.3.

P-value is 0.012 and 0.008 for the first and the second case respectively, and clear multimodality in the density plot presents in both cases. Therefore the conclusion of presence of outliers in upper side is established by the strong evidence for both cases.

[Insert Figure 2.2]

[Insert Figure 2.3]

Not all other outlier detection methods (see Appendix A.1) give same results though. The intervals to identify mild and extreme outlier are $(-2.047, 3.414)$ and $(-4.095, 5.461)$ respectively for interquartile range method. Two of the three added data points are then considered as mild outliers. For modified interquartile range, the intervals to identify the mild and extreme outlier are $(-2.257, 3.625)$ and $(-4.517, 5.883)$ respectively, so that the three added data points are not outliers. P-value for Grubbs' test is 0.218, thus we do not reject the null hypothesis of presence of outliers. For Dixon's Q test, P-value is 0.063, which only suggests borderline outliers.

When there are some intermediate data points filling in the gap between the outliers and other data points, the outlier effect is mitigated such that the original outliers are rather considered as data points from some heavy-tailed distribution than outliers. As seen in Section 2.3, for such case, $Y_{(n)}$ and ξ do not have clear distinctions and the multimodality will vanish. The extended Bootlier procedure performs better than other outlier detection methods in this case, which is illustrated by the following simulation study.

The simulation study #2 has a sample of size 100 simulated from standard normal distribution, and four points, 2.8, 3.2, 3.6 and 4, are added to the sample. The results of the extended Bootlier procedure are present in Figure 2.4. The density plot is nearly smooth and P-value is 0.586, which suggest no outliers in the sample. This is due to the intermediate points filling the gap between the large values and rest of data points.

[Insert Figure 2.4]

Other outlier detection methods have different results. The interval to identify mild outlier is $(-2.730, 3.062)$ for interquartile range method, the largest three values are considered as mild outliers. By modified interquartile range, the interval to identify the mild outlier becomes $(-3.239, 3.570)$. Again the largest two values are considered as mild outliers. P-value for Grubbs' test is 0.023, which suggests the outliers in the data. Dixon's Q test is not suitable for this study.

When the sample comes out from a heavy-tailed distribution, the probability of having large values is higher. Most commonly used outlier detection methods are developed under the assumptions of normal distribution, therefore they often exhibit higher false positive rate when the sample is from heavy-tailed distributions. The extended Bootlier procedure usually performs better on such samples. This is also observed and explained in simulation #2. The following simulation study illustrates a case when the entire sample is from a heavy-tailed distribution.

In simulation study #3, a sample of size 25 is simulated from exponential distribution with 7 added to the sample. The total number of data points is then 26. The sample data is plotted in Figure 2.5 (a). The density plot, sample Bootlier index and P-value for exponential distribution are presented in Figure 2.5 (b). The nearly smooth density plot in Figure 2.5 (b) and P-value (0.534) suggest no outlier in data if the data is from exponential distribution. Even if we do not know the distribution, the moderate sample Bootlier index (0.02244) may only suggest very mild outliers. In fact P-value using normal distribution as reference distribution is 0.101, which is not an evidence of outliers. Besides the normal and exponential distributions, P-value is 0.242 and 0.711 for student t_6 and Cauchy distribution respectively, which also suggest no outliers in the data given that the sample is from other heavy-tailed distributions. Although P-value is 0.001 and 0.024 for uniform distribution and a bimodal distribution with a density $g(x) = \frac{1}{2}\phi(x-1.5)+\frac{1}{2}\phi(x+1.5)$ respectively, it is easy to rule out these distributions.

[Insert Figure 2.5]

Other outlier detection methods all provide strong evidence of the presence of outliers though. The interval to identify extreme outlier is $(-2.388, 3.887)$ for interquartile range method, the three largest values are considered as extreme outliers. By modified interquartile range, the interval to identify the extreme outlier becomes $(-2.645, 4.144)$, again the three largest values are considered as extreme outliers. P-value for Grubbs' test is 0.003, which is a strong evidence of presence of outliers in the data. For Dixon's Q test, P-value is 0.056, which suggests borderline outliers in the data.

Without any distribution assumptions, the inference using the empirical threshold of Bootlier index as suggested by Singh and Xie (2003) or based on the distribution of $BI\left(\hat{f}_{MTM}\right)$ using normal distribution as reference may increase the false positive rate. However it is still a good strategy if the purpose of analysis is to screen hundreds of datasets to identify those for further investigation, for example, to screen hundreds of genes of a gene expression dataset of a set of patients to find potential genes with outliers. In any case, we recommend investigating the underlying distribution of the sample to draw the accurate inference.

We investigate the outlier problem in the samples with sample size from around 25 to 100 so far. When sample size is large, there are two issues to note. First, the probability of having very large values significantly increases in large samples. For example, the probability of a value from standard normal distribution greater than 3.29 is about 0.0005, i.e., $P(X > 3.29) \approx 0.0005$, where $X \sim N(0,1)$. The probability of having at least one value greater than 3.29 is 0.015 for a sample of size 30, but it increases to 0.221 for a sample of size 500, about a 15-fold increase. This leads to higher chance of false positive findings for those outlier detection methods which are not adjusted by sample size. Second, the probability of having multiple outliers also increases dramatically in large samples. Still considering the same example as above, the probability of having at least two values greater than 3.29 is about 0.0001 for a sample of size 30, but it increases to 0.0266 for a sample of size 500, about a 245-fold increase. The two issues are in the nature of outlier problem. The extended Bootlier procedure addresses these two issues well because it is automatically adjusted to the sample size. For the first issue, as proven in the Section 2.3, the separation of modes in the density plot of MTM is determined by distance between largest order statistic $Y_{(n)}$ and outliers ξ. When sample size increases, $Y_{(n)}$ increases so that it requires large outliers ξ to have the separation in the density plot. For the second issue, a remedy is proposed by adjusting the bootstrap sample size to $[rn]$ $(0 < r \leq 1)$, where r is suggested to be $\frac{\log(k+1)}{L}$. One can estimate the number of outliers to pre-specify r or explore different r's to get the best results. The following simulation study illustrates the performance of the extended Bootlier procedure for large sample.

In simulation study #4, a sample of size 500 is simulated from standard normal distribution with three outliers 3.8, 4 and 4.2 added to the sample. The total number of data points is then 503. The sample data is plotted in Figure 2.6 (a). The density plot, sample Bootlier index and P-value for normal distribution are presented in Figure 2.6 (b). Based on density plot in Figure 2.6 (b) and P-value = 0.011, we draw the conclusion of presence of outlier in the data.

[Insert Figure 2.6]

Similar conclusions are obtained by other outlier detection method. The interval to identify extreme outlier is $(-2.837, 2.892)$ for interquartile range method, the three largest values are considered as extreme outliers. By modified interquartile range, the interval to identify the mild outlier becomes $(-3.679, 3.733)$, again the three largest values are considered as extreme outliers. P-value for Grubbs' test is 0.017, which is a strong evidence of presence of outliers in the data.

The extended Bootlier procedure constructs a statistical test to assess the significance of outlier effect, thus it is very important to assess its testing power, i.e., the probability of rejection under the alternative hypothesis. We first show that sample Bootlier index as the test statistic increases with the increase of the magnitude of outliers. Consider two mixed normal density functions, $f_1(x) = \frac{2}{3}\phi(x) + \frac{1}{3}\phi(x-3)$ and $f_2(x) = \frac{2}{3}\phi(x) + \frac{1}{3}\phi(x-5)$. The Bootlier index of $f_2(\cdot)$ (0.262) is much larger than that of $f_1(\cdot)$ (0.0167), which is due to the fact that there is more separation of two modes in $f_2(\cdot)$ than that in $f_1(\cdot)$. As seen in Theorem 2.3, Corollary 2.4 and Theorem 2.5, the separation between modes in the mixture normal distribution function of MTM increases as the magnitude of outliers increase. Although it seems difficult to prove theoretically the monotonicity of sample Bootlier index with magnitude of outliers, the association between larger outliers and larger sample Bootlier index is clear. Therefore, we should see higher probability of rejection for larger outliers. Next we conduct a simulation study to assess the testing power of Bootlier procedure and prove the above arguments.

The simulation study #5 consists of three sub-studies. For the first simulation study (#5A), we draw a sample of size 30 from standard normal distribution, truncate the values greater than 3.29 to 3.29, then add 3.29 to the sample. The value 3.29 is chosen based on $P(X > 3.29) \approx 0.0005$, where $X \sim N(0,1)$. We then apply the extended Bootlier procedure to get P-value. After repeating the above steps 1,000 times, we assess the power of testing procedure under different significance level $\alpha = 0.01, 0.05$ and 0.1. For the second simulation study (#5B), we use 3.72 instead of 3.29 to assess the testing power of extended Bootlier procedure under the same significance levels. The value 3.72 is chosen based on $P(X > 3.72) \approx 0.0001$, where $X \sim N(0,1)$. For the third simulation study (#5C), we use 4.23 based on $P(X > 4.23) \approx 0.00001$. The results are presented in Table 2.1.

[Insert Table 2.1]

If we consider $\alpha = 0.1$ corresponding to a mild outlier and $\alpha = 0.05$ corresponding to an extreme outlier, the extended Bootlier procedure achieves sufficient power to identify 4.23 as an extreme outlier (power = 90.0%) and at least a mild outlier (power = 97.0%). But for 3.29, a more possible value than 4.23 in a normal sample, the procedure has a fair chance to identify it as at least a mild outlier (power = 69.1%) or as an extreme outlier (power = 48.5%).

2.6 REAL DATA EXAMPLE

2.6.1 Temperature Data of Space Shuttle Challenger

The real data examples consist of two studies. In the first study, we revisit a sample for the recorded temperature at which the primary O-ring of the space shuttle Challenger was sealed on 24 launches; see Dalal et al. (1989). The temperature at the time of Challenger explored, Jan. 28, 1986, was 31 degrees. This data example is used in a lot of literatures; see Singh and Xie (2003), Agresti (2002) and Robert and Casella (2004).

The temperatures in Fahrenheit at which O-ring was sealed are,

66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58, 31

The sample data is plotted in Figure 2.7 (a). The temperature of Challenger's last launch, 31 degree, is a suspect outlier in the lower side. The normal Q-Q plot of temperature data, Figure 2.7 (b), supports the normality of data except for the suspected outlier point. Therefore we consider normal distribution as reference. The extended Bootlier procedure is then applied for upper-sided and lower-sided trimming, and the density plots are presented in Figure 2.7 (c) and (d). The multimodality is clearly demonstrated in Figure 2.7 (d) with *P*-value = 0.004, but the density plot is nearly smooth in Figure (c) with *P*-value = 1. Consequently, the presence of outliers in lower side is confirmed.

[Insert Figure 2.7]

2.6.2 Natality Data

The second study illustrates an analysis of outliers in a large sample. We consider a dataset that contains natality information of the United States from 2007 to 2012, supplied by Wide-ranging Online Data for Epidemiologic Research (WONDER) of Centers for Disease Control and Prevention (CDC) (http://wonder.cdc.gov). We analyze the fertility rate for 572 counties or combined counties in the United States. The counties with populations of 100,000 or more are listed while those with fewer than 100,000 persons within a state are combined together and labeled as “Unidentified Counties”. The fertility rates are calculated as the number of births divided by the number of females with age 15 to 44 years old, and expressed as the rate per 10,000 persons.

The data is plotted in Figure 2.8 (a). Two candidate outliers (Hampshire County, MA [fertility rate = 280.36] and Onslow County, NC [fertility rate = 1099.27]) are suspected with one on the lower side and the other on the upper side. Figure 2.8 (b), the normal Q-Q plot of the fertility rate does not rule out the normal assumption. We consider the normal distribution as the reference. The density plots are presented in Figure 2.7 (c) and (d). *P*-values are 0.008 and 0.154 for upper-sided and lower-sided trimming respectively. Consequently, the presence of outliers in the upper side is confirmed, while the evidence of outliers in lower side is not strong enough. To further confirm if Onslow County, NC is the only outlier, we re-analyze the sample without that county. The density plot is nearly smooth, Figure 2.8 (e), and *P*-value = 0.681. Therefore Onslow County, NC is identified as the only outlier in this dataset.

[Insert Figure 2.8]

2.7 CONCLUDING REMARKS

In this chapter we propose a bootstrap-based statistical framework, the extended Bootlier procedure, to detect outliers in i.i.d. univariate sample. The key elements of the extended Bootlier procedure including bootstrap, density plot, Bootlier index and the testing method are thoroughly discussed. Then the features of the proposed framework are illustrated through various simulation studies and real data examples. In these studies, the extended Bootlier procedure presents good performance under various scenarios, such as multiple outliers, heavy-tailed distribution, and large samples. The extended Bootlier procedure also demonstrates good testing power. The comparisons with other outlier detecting methods show its overall advantages. The general result of the limiting distribution of MTM for multiple outliers is developed as the foundation of the extended Bootlier procedure. The choice of bootstrap size under multiple outlier cases is suggested based on the limiting results. Finally, the interesting areas for further study may include a sequential procedure similar to the generalized extreme generalized extreme studentized deviate (ESD) (see Rosner, 1983), and an automatic method searching the optimal r to increase the testing power.

APPENDIX

A.1 A review of Outlier Detection Methods

A review of several commonly used outlier detection methods is presented. The purpose is not to exhaust all the outlier detection methods but to provide a general picture of representative methods. The performance of proposed Bootlier procedure is compared with these methods to show the overall advantages.

Box-plot and interquartile range method

A box-plot is a graphic representation of the dispersion of the data. The graphic represents the lower quartile $Q1$, 25th percentile of data, and upper quartile $Q3$, 75th percentile of the data, along with the median. The interquartile range (IQR) is defined as $Q3 - Q1$. The upper whisker limit (upper fence) is $Q3 + 1.5 \times IQR$ and lower whisker limit (lower fence) is $Q1 - 1.5 \times IQR$. As a crud method, any observation outside the interval $(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$ is considered as an outlier, which is further classified as mild if it is within the interval $(Q1 - 3 \times IQR, Q3 + 3 \times IQR)$, or extreme if it outside that interval according to NIST/SEMATECH (2012).

The above approach does not take sample size into account so that when the sample size is large, the probability of extreme values appearing in the tails is high even the data is truly from a normal distribution, which will leads to higher false positive rate. A modified interquartile range is proposed; see Barbato et al. (2011). Assuming normality, the observations outside the interval $(Q1 - 1.5 \times IQR \times (1 + 0.1 \times \log\frac{n}{10}), Q3 + 1.5 \times IQR \times (1 + 0.1 \times \log\frac{n}{10})$ but within the interval $(Q1 - 3 \times IQR \times (1 + 0.1 \times \log\frac{n}{10}), Q3 + 3 \times IQR \times (1 + 0.1 \times \log\frac{n}{10})$ are considered as mild outliers, and the observation outside the later interval are considered as extreme outliers.

Because of its simplicity and quickness of implementation, the box-plot seems appealing in the examining the data especially for screening different features of a given data for potential outliers.

Grubbs' test

Grubb's test, also known as maximum normed residual test, is used to detect a single outlier in a univariate data set assumed to come from a normal distribution; see Grubb (1969) and Stefansky (1972).

Given a data set of size n, $\{Y_1,\ Y_2,\cdots,\ Y_n\}$, the hypothesis of interest is H_0: there are no outliers in the dataset vs. H_1: there is at least one outlier in the data set. Grubb's test statistic is defined as $G_l = \frac{\bar{Y}-Y_{(1)}}{s}$ and $G_u = \frac{\bar{Y}-Y_{(n)}}{s}$, where $\bar{Y}$ and s are sample mean and sample standard deviation respectively. For a two-sided test, the null hypothesis is rejected if $G > \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{n-2,1-\frac{\alpha}{2n}}}{n-2+t^2_{n-2,1-\frac{\alpha}{2n}}}}$, where G stands for G_l or G_u, and $t_{v,\alpha}$ is the 100α percentile from t-distribution with v degrees of freedom. The significance level of this test is α.

A generalized extreme studentized deviate (ESD) test

A generalized extreme studentized deviate (ESD) test is used to detect one or more outliers in a data set which follows approximately normal distribution (see Rosner, 1983), which can be considered as a sequential procedure of Grubb's test with adjustment of critical values based on the number of tests.

Given a data set of size n, $\{Y_1,\ Y_2,\cdots,\ Y_n\}$, and an upper bound k, the generalized ESD tests the hypothesis, H_0: there are no outliers in the data set vs. H_1: there are up to k outliers in the data set. The test statistics is computed in iteratively way. Compute $R_1 = \max_i \left|\frac{Y_i-\bar{Y}}{s}\right|$ where $\bar{Y}$ and s are sample mean and sample standard deviation respectively. Then remove the observation which maximizes R_1 and compute R_2 with the rest $n-1$ samples, repeat the process to get $R_{3,}\ldots,R_k$. Corresponding to the k test statistics, k critical values are computed by $\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{(n-i-1+t^2_{n-i-1,p})(n-i+1)}}$ for $i = 1,2\ldots,k$, $t_{v,\alpha}$ is the 100α percentile from t-distribution with v degrees of freedom and $p = 1 - \frac{\alpha}{2(n-i+1)}$. The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$.

The generalized ESD test makes the appropriate adjustment for critical values to avoid the early stopping of Grubb's test if without the adjustment. The generalized ESD is restricted to two-sided testing while the Grubb's test does not.

Dixon's Q test

This test is proposed by Dixon (see Dixon, 1950, 1951, and Barbato et al., 2011), and is appropriate for detecting outliers in small sample with sample size ≤ 40.

Given a data set of size n, $\{Y_1, Y_2, \cdots, Y_n\}$, the hypothesis for Dixon's Q test is, H_0: there are no outliers in the data set vs. H_1: there are at lease one outlier in the data set. The test statistics are the quotient of *gap* divided by *range*, where *gap* is the absolute difference between the outlier in question and the close number to it and *range* is the range of all data points. When sample size is ≥ 3, it has the form: $Q_l = \frac{Y_{(i+1)} - Y_{(1)}}{Y_{(n-j)} - Y_{(1)}}$ and $Q_u = \frac{Y_{(n)} - Y_{(n-i)}}{Y_{(n)} - Y_{(j+1)}}$, where the coefficients i and j are determined by: $i = 1, j = 0, if\ 3 \leq n \leq 7$; $i = 1, j = 1, if\ 8 \leq n \leq 12$ and $i = 2, j = 2, if\ 13 \leq n \leq 40$. The null hypothesis is rejected if Q_l or Q_u exceeds the tabulated critical values which are based on the normal distribution assumption.

Most of discordancy tests discussed above rely on the assumption of normal distribution, which limits their application in the broader data analysis. Also, these methods are all for the i.i.d. univariate data.

A.2 Proof of Results

PROOF OF THEOREM 2.1

Note that $m(\xi)$ is a $Binomial(n + L, \frac{L}{n+L})$. As $n \to \infty$, it is a Poisson random variable with mean L, denoted by $\mathcal{P}_L$. We then have the following for $G_n(x)$.

$$G_n(x) = \sum_{i=0}^{K} P\left(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \leq x | m(\xi) = i\right) P(m(\xi) = i) + R_1, \qquad \text{(A.1)}$$

where $R_1 < P(m(\xi) > K)$.

For any fixed $0 \le i \le K$, $P(m(\xi) = i)$ converges to $\frac{e^{-L}L^i}{i!}$ as $n \to \infty$. Then,

$$\sum_{i=0}^{K} P\big(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \le x | m(\xi) = i\big) P(m(\xi) = i)$$

$$= \sum_{i=0}^{K} \frac{e^{-L}L^i}{i!} P\big(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \le x | m(\xi) = i\big) + R_2 \ , \qquad \text{(A.2)}$$

where $R_2 = o(1)$.

Given $m(\xi) = i$, let $\{\xi_1^*, \xi_2^*, \cdots, \xi_i^*\}$ be those bootstrap samples from $\{\xi_1, \xi_2, \cdots, \xi_L\}$. Then $(\xi_1^*, \xi_2^*, \cdots, \xi_i^*)$ has L^i unique configurations, i.e., $(\xi_1^*, \xi_2^*, \cdots, \xi_i^*) = S_{i,j}(\xi), j = 1,2,..,L^i$, each with probability L^{-i}. Further conditional on $S_{i,j}(\xi)$, we have

$$P\left(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \le x | m(\xi) = i, (\xi_1^*, \xi_2^*, \cdots, \xi_i^*) = S_{i,j}(\xi)\right)$$

$$= P\left(\sqrt{n+L}\left(\frac{\Sigma_{k=1}^{n+L-i} Y_{(k)}^*}{n+L} - \bar{Y}_n\right) \le x - \frac{i\overline{S_{i,j}}(\xi)}{\sqrt{n+L}} | m(\xi) = i, (\xi_1^*, \xi_2^*, \cdots, \xi_i^*) = S_{i,j}(\xi)\right).$$

$$\text{(A.3)}$$

Suppose that $\{Z_1, Z_2, \cdots, Z_{n+L-i}\}$ be a bootstrap sample from $\{Y_1, Y_2, \cdots, Y_n\}$. By Lemma B of Singh and Xie (2003), the conditional distribution of $\Sigma_{k=1}^{n+L-i} Y_{(k)}^*$ given $m(\xi) = i$ and $(\xi_1^*, \xi_2^*, \cdots, \xi_i^*) = S_{i,j}(\xi)$, is same as of $\Sigma_{k=1}^{n+L-i} Z_{(k)}$. Using the central limit theorem for bootstrap sample mean; see Singh (1981) or Bickel and Freedman (1981), one can prove $\sqrt{n+L}\left(\frac{\Sigma_{k=1}^{n+L-i} Y_{(k)}^*}{n+L} - \bar{Y}_n\right)$ converges to $N(0,1)$, which leads to,

$$P\left(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \le x | m(\xi) = i, (\xi_1^*, \xi_2^*, \cdots, \xi_i^*) = S_{i,j}(\xi)\right) \to \Phi\left(x - \frac{i\overline{S_{i,j}}(\xi)}{\sqrt{n+L}}\right)$$

$$\text{(A.4)}$$

By (A.3) and (A.4), we have

$$P\big(\sqrt{n+L}(\bar{Y}_{n+L}^* - \bar{Y}_n) \le x | m(\xi) = i\big) = \frac{1}{L^i} \sum_{j=1}^{L^i} \Phi\Big(x - \frac{i\overline{S_{i,j}}(\xi)}{\sqrt{n+L}}\Big) + R_3 \qquad \text{(A.5)}$$

where $R_3 = o(1)$.

Finally by (A.1), (A.2) and (A.5), we have (2.6)

$$G_n(x) = \sum_{i=0}^{K} \frac{e^{-L}}{i!} \sum_{j=1}^{L^i} \Phi\left(x - \frac{i\overline{S_{i,j}}(\xi)}{\sqrt{n+L}}\right) + r_{n,K},$$

where $\limsup_n |r_{n,K}| \leq P(\mathcal{P}_L > K)$.

PROOF OF THEOREM 2.3 (A)

Let A_0 be the event that $Y_1^*, Y_2^*, \cdots, Y_{n+L}^*$ is free of the outliers $\xi_1, \xi_2, \cdots, \xi_L$, and A_i $(i = 1,2, \ldots)$ be the event that $Y_1^*, Y_2^*, \cdots, Y_{n+L}^*$ is free of the outliers and the top i members of $Y_1, Y_2, \cdots, Y_n$. Then A_i $(i = 0,1,2, \cdots)$ form a monotonically decreasing sequence of events, i.e., $A_0 \supset A_1 \supset A_2 \cdots$. We also note that, for any fixed $i \geq 1$, $P(A_{i-1}) \to e^{-(L+i-1)}$ and $P(A_{i-1} - A_i) \to e^{-(L+i-1)} - e^{-(L+i)}$ as $n \to \infty$. Therefore,

$$H_{n,0}(x) = P(T_n \leq x | A_0) = \frac{1}{P(A_0)} \sum_{i=1}^{K} P((T_n \leq x) \cap (A_{i-1} - A_i)) + R_1, \quad \text{(A.6)}$$

where $|R_1| \leq \frac{P(A_K)}{P(A_0)}$.

To conclude the theorem (a), we will show that, for any $1 \leq i \leq K$,

$$P(T_n \leq x | A_{i-1} - A_i) = \Phi\left(x + \sqrt{n+L}\, Y_{(n-i+1)}\right) + o(1) . \quad \text{(A.7)}$$

Adopting a general notation, we let $m(\{Y_{(i)}\})$ denote the number of times that $Y_{(i)}$ appears in the bootstrap sample $Y_1^*, Y_2^*, \cdots, Y_{n+L}^*$. Given the event set $A_{i-1} - A_i$, we have $Y_{(n+L)}^* = Y_{(n-i+1)}$. For a fixed positive integer $j \leq n + L$, consider $P\left(T_n \leq x | A_{i-1} - A_i, m(\{Y_{(n-i+1)}\}) = j\right)$. By Lemma B of Singh and Xie (2003), the conditional bootstrap distribution of $\sum_{k=1}^{n+L-j} Y_{(k)}^*$, given that $A_{i-1} - A_i$ and $m(\{Y_{(n-i+1)}\}) = j$, has the same distribution as that of bootstrap sample sum with size $(n + L - j)$, drawn from $Y_{(1)}, Y_{(2)}, \cdots, Y_{(n-i)}$. In the view of the fact that i and j are fixed and $\frac{Y_{(n)}}{\sqrt{n}} \to 0$, a.s. (see Singh and Xie, 2003, Lemma A), we have,

$$P\left(T_n \leq x | A_{i-1} - A_i, m(\{Y_{(n-i+1)}\}) = j\right) = \Phi\left(x + \sqrt{n+L}\, Y_{(n-i+1)}\right) + o(1). \quad \text{(A.8)}$$

For different value of j, $m(\{Y_{(n-i+1)}\}) = j$ defines that partition of $(A_{i-1} - A_i)$. As right-side of (A.8) is free of j, and $P\big(m(\{Y_{(n-i+1)}\}) = j | A_{i-1} - A_i\big) \to \frac{1}{(e-1)j!}$ as $n \to \infty$, we can conclude (A.7) from (A.8). Therefore (2.7) holds from (A.6) and (A.7).

PROOF OF THEOREM 2.3 (B)

By definition, $H_{n,1}(x) = P(T_n \leq x | m(\xi) \geq 1)$.

Given that $Y^*_{(n+L)} = \xi_{(j)}$, $j = 1,2, \dots, L$, the following holds for any fixed $1 \leq i \leq K$,

$$P\big(T_n \leq x | m(\xi) = i, Y^*_{(n+L)} = \xi_{(j)}\big)$$

$$= \frac{1}{L^{i-1}} \sum_{k=1}^{L^{i-1}} \Phi\Big(x + \sqrt{n+L}\,\xi_{(j)} - \frac{\xi_{(j)} + (i-1)\overline{S_{i-1,k}}\big(\{\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(j)}\}\big)}{\sqrt{n+L}}\Big) + o(1)\,. \qquad \text{(A.9)}$$

Note that $P\big(Y^*_{(n+L)} = \xi_{(j)} | m(\xi) = i\big) = \left(\frac{j}{L}\right)^i - \left(\frac{j-1}{L}\right)^i$, $1 \leq j \leq L$, we have,

$$P(T_n \leq x | m(\xi) = i) =$$

$$\frac{1}{L^{i-1}} \sum_{j=1}^{L} \left(\left(\frac{j}{L}\right)^i - \left(\frac{j-1}{L}\right)^i\right) \sum_{k=1}^{L^{i-1}} \Phi\Big(x + \sqrt{n+L}\,\xi_{(j)} - \frac{\xi_{(j)} + (i-1)\overline{S_{i-1,k}}\big(\{\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(j)}\}\big)}{\sqrt{n+L}}\Big) + o(1).$$

(A.10)

Therefore (2.8) holds from (A.10).

A.3 Figures and Tables

Figure 2.1: Bootlier index for a distribution with density function $f(x) = \frac{1}{2}\phi(x) + \frac{1}{3}\phi(x-3) + \frac{1}{6}\phi(x-9)$, where $\phi(\cdot)$ stands for the standard normal density.

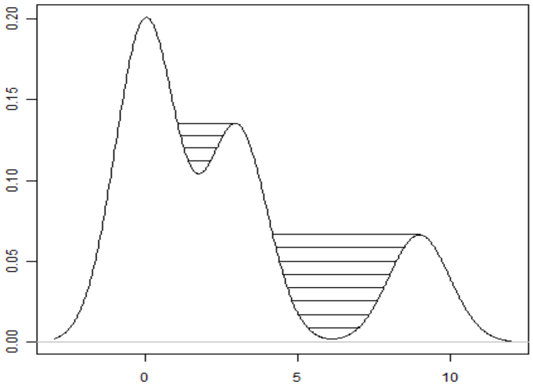
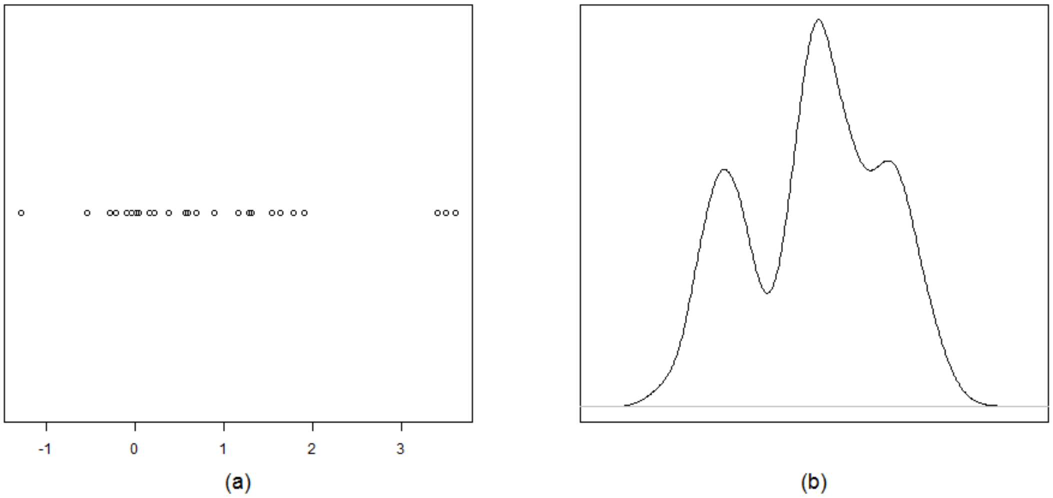


Figure 2.2 (Simulation #1) Extended Bootlier procedure for a sample of size 25 simulated from standard normal distribution with three outliers, 3.4, 3.5 and 3.6, added: (a) 28 data values; (b) the density plot (sample Bootlier index = 0.08084, and P-value = 0.012 for normal distribution)



(a)

(b)

Figure 2.3 (Simulation #1) Extended Bootlier procedure for the same 25 data values as Figure 2.2 with 3.5 added to the sample: the density plot (sample Bootlier index = 0.43435 and *P*-value = 0.008 for normal distribution)

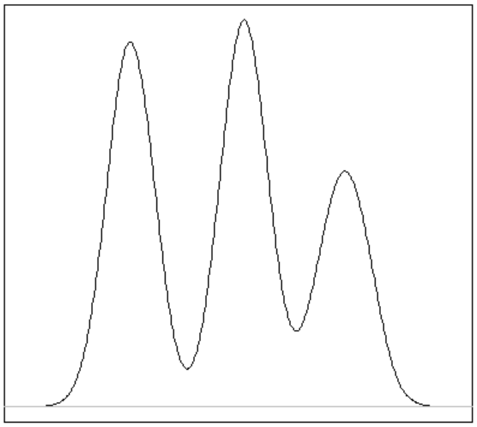


Figure 2.4 (Simulation #2) Extended Bootlier procedure for a sample of size 100 simulated from standard normal distribution with 2.8, 3.2, 3.6 and 4 added: the density plot (sample Bootlier index = 0.00015, and *P*-value = 0.586 for normal distribution)

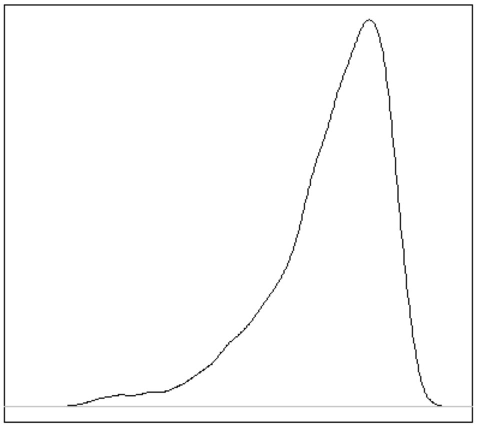
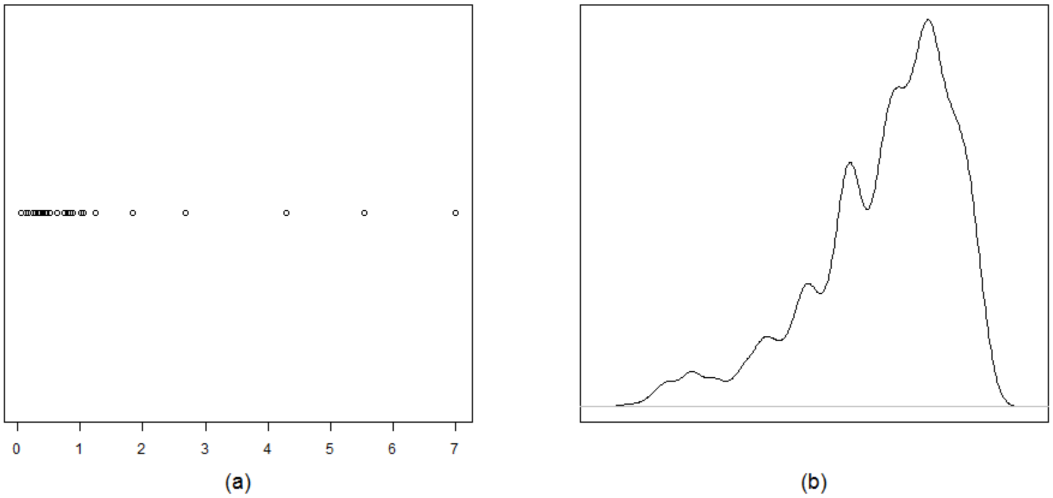


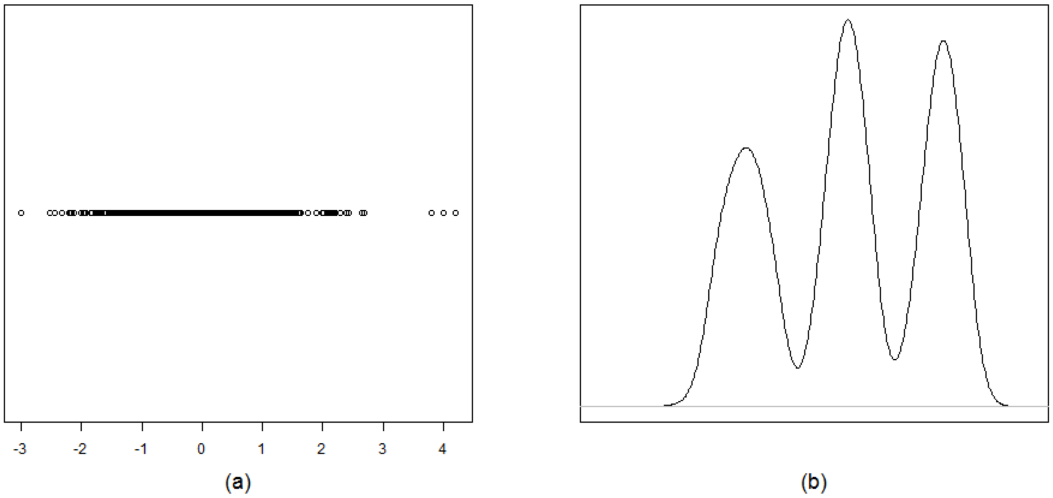
Figure 2.5 (Simulation #3) Extended Bootlier procedure for a sample of size 25 simulated from exponential distribution with 7 added to the sample: (a) 26 data values; (b) the density plot (sample Bootlier index = 0.02244 and *P*-value = 0.534 for exponential distribution)



(a)

(b)

Figure 2.6 (Simulation #4) Extended Bootlier Procedure for a sample of size 500 simulated from standard normal distribution with three outliers, 3.8, 4 and 4.2, added: (a) 503 data values (b) the density plot (sample Bootlier index = 0.44189 and *P*-value = 0.011 for normal distribution)



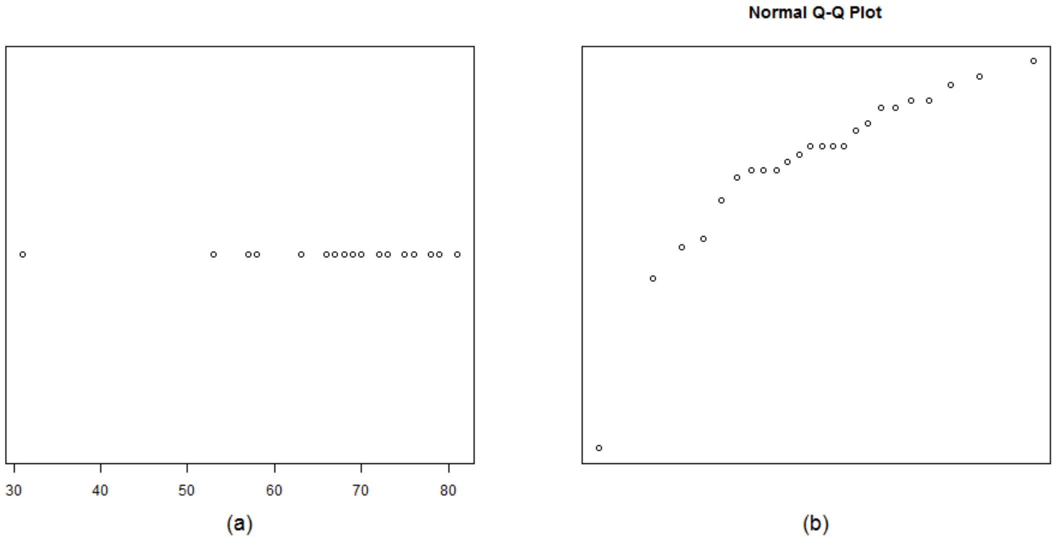
(a)

(b)

Table 2.1 (Simulation #5) Testing power of extended Bootlier procedure

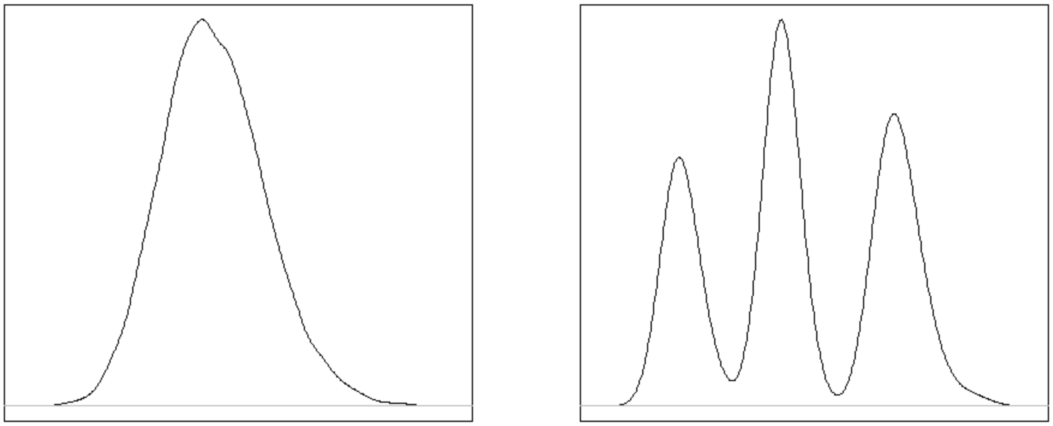
Simulation Study	Type I error α		
	0.1	0.05	0.01
#5A	69.1%	48.5%	14.5%
#5B	87.4%	71.6%	32.5%
#5C	97.0%	90.0%	56.8%

Figure 2.7 (Real Data Example #1) Extended Bootlier procedure for recorded temperature at which primary O-ring of space shuttle *Challenger* was sealed: (a) 24 original data values, (b) normal Q-Q plot of 24 data values, (c) the density plot with upper-sided trimming (sample Bootlier index = 0, and *P*-value = 1), (d) the density plot with lower-sided trimming (sample Bootlier index = 0.59032, and *P*-value = 0.004)



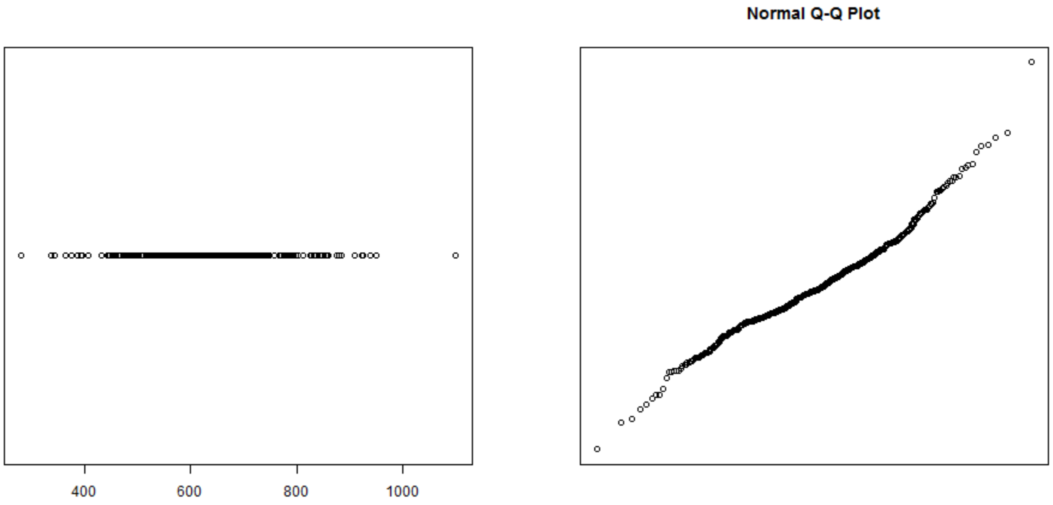
(a)

(b)

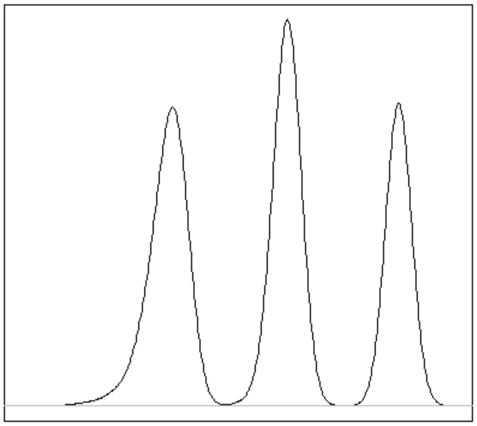


(c) (d)

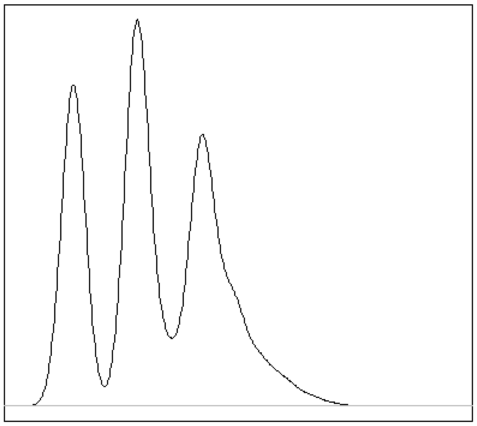
Figure 2.8 (Real Data Example #2) Extended Bootlier procedure for county-level fertility rates of the United States supplied by US CDC WONDER database: (a) 572 original data values, (b) normal Q-Q plot of data, (c) the density plot with upper-sided trimming (sample Bootlier index =1.20468, and *P*-value = 0.008), (d) the density plot with lower-sided trimming (sample Bootlier index = 0.46990, and *P*-value = 0.154), (e) the density plot with upper-sided trimming (sample Bootlier index = 0.00270, and *P*-value = 0.681) without Onslow County, NC



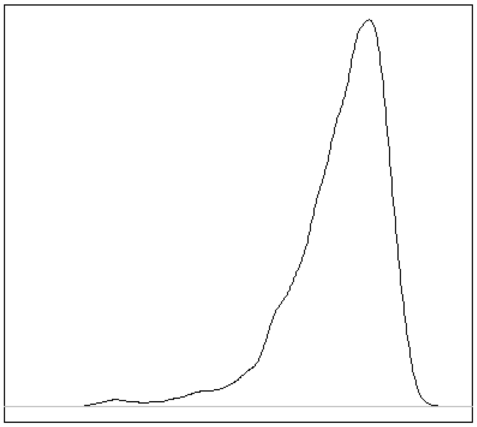
(a) (b)



(c)



(d)



(e)

REFERENCES

Agresti, A. (2002), Categorical Data Analysis 2nd Edition, John Wiley & Sons, Inc.

Barbato, G., Barini, E.M., Genta, G. and Levi, R. (2011), Features and Performance of Some Outlier Detection Methods, *Journal of Applied Statistics*, Vol. 38, No. 10, 2133-2149

Barnett, V. and Levis, T. (1994), Outliers in Statistical Data (3rd edition), Wiley

Bickel, P. J. and Freedman, D. (1981), Some Asymptotic Theory for the Bootstrap, *The Annals of Statistics*, Vol. 9, No. 6, 1196-1217

Dalal, S.R., Fowlkes, E.B. and Hoadley B. (1989), Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure, *Journal of American Statistical Association*, Vol. 84, No. 408, 945-957

Dixon, W.J. (1950), Analysis of Extreme Values, *Annals of Mathematical Statistics*, Vol. 21, No. 4, 488-506

Dixon, W.J. (1951), Ratios Involving Extreme Values, *Annals of Mathematical Statistics*, Vol. 22, No. 1, 68-78

Grubbs, F. (1969), Procedure for Detecting Outlying Observations in Samples, *Technometrics*, Vol. 11, No.1, 1-21

Jones, M.C., Marron, J.S., and Sheather, S.J. (1996), A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of American Statistical Association*, Vol. 91, No. 433, 401-407

Natality data, United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics, Natality public-use data 2007-2012, on CDC WONDER Online Database, November 2013. Accessed at *http://wonder.cdc.gov/natality-current.html*

NIST/SEMATECH, e-Handbook of Statistical Methods, 2012

Parzen, E. (1962), On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics*, Vol. 33, No. 3, 1065-1076

Robert, C. P. and Casella, G. (2004), Monte Carlo Statistical Methods 2nd Edition, Springer

Rosenvblatt, M. (1956), Remarks on some non-parametric estimates of a density function, *Annals of Mathematical Statistics*, Vol. 27, No. 3, 832-837

Rosner, B. (1983), Percentage Points for a Generalized ESD Many-Outlier Procedure, *Technometrics*, Vol. 25, No.2, 165-172

Sheather, S.J. (2004), Density Estimation, *Statistical Science*, Vol. 19, No. 4, 588-597

Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis, Chapman and Hall

Singh, K. (1981), On the Asymptotic Accuracy of Efron's Bootstrap, *The Annals of Statistics*, Vol. 9, No. 6, 1187-1195

Singh, K. and Xie, M. (2003), Bootlier-Plot – Bootstrap Based Outlier Detection Plot, *Sankhya: The Indian Journal of Statistics*, Vol. 65, Part 3, 532-559

Stefansky, W. (1972), Rejecting Outliers in Factorial Designs, *Technometrics*, Vol. 14, No. 2, 469-479

CHAPTER 3

IDENTIFYING OUTLIERS AND INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION ANALYSIS BY EXTENDED BOOTLIER PROCEDURE

ABSTRACT

The identifying outliers and influential observations is a fundamental question in linear regression model diagnostics. In Chapter 2 of this dissertation a distribution-free testing procedure is proposed for detection of outliers in the independent and identically distributed (i.i.d.) univariate sample. As an extension of that testing procedure, in this chapter, a statistical framework is proposed to detect outliers and influential observations in linear regression analysis by analyzing the outliers in the residuals and influential measures from regression model fitting. In the analysis of outliers, the ordinary residuals, studentized residuals, and studentized deletion results are suggested, while a square-root version of Cook’s distance is proposed to analyze influential observations. In contrast to the case with i.i.d. samples, the residuals and Cook’s distance are dependent. Therefore we develop large sample theory that explains the association between the test statistic and the outliers (or the influential observations). The proposed framework is then illustrated through simulation studies and a real data example of sperm motility data to show its usefulness.

Keywords: linear regression; outlier; influential observation; residual; Cook’s distance; bootstrap; large sample theory; Bootlier index

3.1 INTRODUCTION

This chapter develops a bootstrap-based statistical framework for detecting outliers and influential observations in linear regression analysis. The framework extends the statistical method proposed in Chapter 2 for detection of outliers in the independent and identically distributed (i.i.d.) univariate sample to linear regression analysis, a more complicated setting. The outliers and influential observations are detected by analyzing the residuals and the influential measures from regression model fitting using the same techniques as the procedure for univariate sample. Although, in regression, it is generally assumed that error terms are i.i.d., the residuals and influential measures no longer are. Therefore some adjustments are made to the procedure and the large sample theory is developed for regression setting, which are two foci of this chapter.

There is certainly a vast literature on the detection of outliers and influential observations in linear regression analysis, as the study on these is a major topic in linear regression model diagnostics; see Beckman and Cook (1983) and Barnett and Levis (1994). The topic is present in almost every textbook of linear regression model. An outlier usually means an outlying point which departs surprisingly far from the regression line. For an influential observation, a definition given by Belsley et al. (1980) seems most appropriate: “an influential observation is which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates … than is the case for the most of the other observations”. This definition is also adopted by Chatterjee and Hadi (1986) and they classify the common measure of influence into the residuals, the prediction matrix, the volume of confidence ellipsoids, the influence functions, and the partial influence. The difference between influential observations and outliers, to some extent, is vague. While an outlier is often an influential observation, an influential observation need not be an outlier. The methods for the detection of outliers and influential observations are generally classified into two groups, namely the graphical and the analytical methods. The graphic methods usually display different statistics measuring the degree of departure of outliers or influential observations from other data points. These graphic methods include scatter plot, residual plot, box-plot and normality plot to name a few. The analytical methods include various discordancy tests based on different residuals and influence measures. Chatterjee and Hadi (1986) have a complete review of the measures and discordancy tests based on those measures. Some discordancy tests are discussed in Barnett and Levis (1994).

The method proposed in this chapter falls into the group of analytical approaches. It also includes a graphical plot to facilitate the interpretation of analysis results. The extended Bootlier procedure developed in Chapter 2 for univariate sample is motivated by a non-parametric and graphic outlier detection method proposed by Singh and Xie (2003). They prove that, when the data has outliers and we draw a large number of bootstrap samples, a sample statistic “mean – trimmed mean” (referred afterwards as MTM) will have multiple modes in its density plot (a bootstrap histogram). Therefore detecting outliers is equivalent to checking multimodality in the density plot of MTM. A quantitative index is proposed to measure the degree of multimodality of density plot. By utilizing that index as the test statistic, the extended Bootlier procedure developed in Chapter 2 provides a statistical inference approach for testing the outliers in the univariate sample. In this chapter, we extend the procedure with some modifications to regression setting.

There are several points to note. First, the outliers and influential observations have distinct meanings in regression and there are different quantities to measure them; see Cook (1979) and Chatterjee and Hadi (1986). Thus the inference drawn from the procedure depends on what measures we use. We explore the ordinary residuals, studentized residual and studentized deletion residual as the measures for outliers and a square-root version of Cook’s distance (referred afterward as SRCD) as the measure for influential observations, but the procedure is not limited to those. Second, the residuals and Cook’s distance are dependent, thus the large sample theory developed in Chapter 2 do not apply directly. Noticing certain properties of the residuals and SRCD, we develop large sample results similar to those in Chapter 2. Third, the extended Bootlier procedure for i.i.d. univariate data demonstrates lower false positive rate when the sample is from heavy-tailed distribution. This good property is still possessed in regression setting – when the distribution of error terms deviates from normal with high probability in the tail, the extended Bootlier procedure has lower false positive rate to claim large values as outliers. We illustrate it in a simulation study.

The rest of the chapter is organized as follows. Section 3.2 introduces the basic definitions of linear regression model including residuals and SRCD, and describes the proposed procedure with details. Section 3.3 develops the large sample theory. In section 3.4, two simulation studies are conducted to illustrate the performance of the proposed procedure. Section 3.5 presents a real data example. Finally, concluding remarks are present in Section 3.6.

3.2 EXTENDED BOOTLIER PROCEDURE FOR LINEAR REGRESSION ANALYSIS

3.2.1 Notations

We consider a linear regression model,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}\,, \tag{3.1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of values of the response (dependent) variable, $\boldsymbol{X}$ is a $n \times p$ matrix of independent variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_i)$ is an unobservable $n \times 1$ vector of error terms that are independent and identically distributed with mean 0 and unknown variance σ^2, i.e., $\varepsilon_i \sim F(0, \sigma^2)$ for $i = 1,2, \dots, n$.

Let $\widehat{\boldsymbol{\beta}}$ be the least square estimator of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{Y}}$ be fitted value of $\boldsymbol{Y}$. The vector of ordinary residuals $\boldsymbol{e} = (e_i)$ is

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{3.2}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}' = (h_{ij})$ is a $n \times n$ projection matrix (often called hat matrix), with i, jth element h_{ij}, $i, j = 1,2, \dots, n$. Its diagonal elements h_{ii} are often termed as leverages to reflect the influence of an observation may have. Detailed discussion related to the hat matrix and leverages can be found in Hoaglin and Welsch (1978), and Cook and Weisberg (1982).

Noticing that each e_i has different variance $\sigma^2(1-h_{ii})$, a scaled version of residual is proposed as studentized residual,

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}\,, \tag{3.3}$$

where $\hat{\sigma}^2$ is the estimate of σ^2.

A different refinement to make residuals more sensitive to outliers is given by Beckman and Trussell (1974). It is often called studentized deletion residual,

$$s_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_{ii})}}\,. \tag{3.4}$$

where $\hat{\sigma}^2_{(-i)}$ is the estimate of variance by fitting the regression model deleting the ith data point.

Cook's distance is generally considered as a good measure for evaluation the influence of a data point to the regression. In this chapter, we propose a square-root version of Cook's distance (SRCD), which is linear in the residual and has large sample theory proved in Section 3.3,

$$d_i = \frac{\sqrt{h_{ii}}}{\sqrt{p\hat{\sigma}^2(1-h_{ii})}}e_i\ . \tag{3.5}$$

There, of course, exist other residuals and influential measures. As the above are well-studied and representative measures, we use them in this chapter for the extended Bootlier procedure. We believe other measures can also be used after careful examination.

Typical graphical methods usually plot these measures and visually identify the data points that yield outliers in these measures as outliers or influential observations, or identify those that yield extreme values in these measures for further investigation. Most analytical methods detect outliers and influential observations by testing the outliers in those measures. For example, a discordancy test by Barnett and Levis (1994) uses $\max(r_i)$ as the test statistic to detect outliers, and the critical values for $\max(r_i)$ are presented for different p (pp. 522-3). Montgomery et al. (2003) suggests that an observation with $r_i > 3$ indicates an outlier. Because each studentized deletion residual s_i has student t distribution with $n - p - 1$ degrees of freedom, a Bonferroni type of test that is adjusting for n tests therefore has the critical value as $t_{n-p-1,1-\alpha/2n}$ to detect outliers; see Neter et al. (1996). In interpreting Cook's distance, Bollen and Jackman (1990) suggests to investigate those with Cook's distance greater than $4/n$. While these discordancy tests are developed specifically for certain residuals or influential measures, the extended Bootlier procedure provides a universal way to analyze the residuals and influential measures. A question is often raised in the real data analysis – which measure is the best? There is no definitive answer. The choice sometimes relies on specific problem setting, and sometimes may be just a personal choice. For example, Belsley et al. (1980) prefer studentized deletion residual over studentized residual (also see Chatterjee and Hadi, 1986). In what follows we elaborate our framework.

3.2.2 Extended Bootlier Procedure for Linear Regression Analysis

Consider the linear regression model (3.1) and let $\{t_1, t_2, \ldots, t_n\}$ be the general form of residuals or influential measures after fitting the model, which could be (3.2), (3.3), (3.4) or (3.5). We apply the procedure proposed in Chapter 2 to $\{t_1, t_2, \ldots, t_n\}$ to obtain a set of MTM, the density plot of MTM, and the quantitative index measuring multimodality of the density as follows.

Let $\{t_1^*, t_2^*, \cdots, t_n^*\}$ be a bootstrap sample from $\{t_1, t_2, \ldots, t_n\}$, $\left\{t_{(1)}^*, t_{(2)}^*, \cdots, t_{(n)}^*\right\}$ be the order statistics of $\{t_1^*, t_2^*, \cdots, t_n^*\}$, $\bar{t}_n = \frac{1}{n}\sum_{i=1}^{n} t_i$, and $\bar{t}_n^* = \frac{1}{n}\sum_{i=1}^{n} t_i^*$. MTM of the bootstrap sample with trimming size k_0 is obtained as,

$$\text{Upper-sided trimming: MTM} = \bar{t}_n^* - \frac{1}{n-k_0}\sum_{i=1}^{n-k_0} t_{(i)}^* \,, \tag{3.6}$$

$$\text{Lower-sided trimming: MTM} = \bar{t}_n^* - \frac{1}{n-k_0}\sum_{i=k_0+1}^{n} t_{(i)}^* \quad . \tag{3.7}$$

Suppose m bootstrap samples are drawn from $\{t_1, t_2, \dots, t_n\}$. The set of MTM is obtained as $\{MTM_1, MTM_2, \cdots, MTM_m\}$. The density plot of MTM is expressed as $\hat{f}_{MTM}(x)$ using the kernel density estimation method,

$$\hat{f}_{MTM}(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x - MTM_i}{h}\right), \tag{3.8}$$

where $K(.)$ is a kernel function and h is the bandwidth.

In the i.i.d. univariate sample case discussed in Chapter 2 (i.e., suppose $\{t_1, t_2, \dots, t_n\}$ were an i.i.d. sample), we prove the association between multimodality of $\hat{f}_{MTM}(x)$ and presence of outliers in the sample, and construct a statistical test based on a quantitative index, Bootlier index (Singh and Xie 2003), for detection the outliers in the sample. Adopting the same definition, we obtain the Bootlier index of $\hat{f}_{MTM}(x)$ (referred as sample Bootlier index) as $BI(\hat{f}_{MTM})$.

As the large sample theory in Section 3.3 shows, the similar association exists between the multimodality of $\hat{f}_{MTM}(x)$ and the presence of the outliers (or influential observations) in linear regression analysis. A formal test is then constructed for detection of outliers (or influential observations), which is formulated as the following hypothesis,

H_0: The sample contains outliers (or influential observations)

H_1: The sample contains no outliers (or influential observations)

The sample Bootlier index $BI(\hat{f}_{MTM})$ is used as the test statistic to test this hypothesis. We denote its distribution function under H_0 (null distribution) by F_0. The P-value for this test is expressed as $Prob = 1 - F_0(BI(\hat{f}_{MTM}))$. Since it is difficult to obtain the closed form of F_0, we propose a simulation method to estimate it instead. Suppose N independent samples with each sample having n independent values are drawn from the distribution of $\{t_1, t_2, \dots, t_n\}$ under H_0. The sample Bootlier index is obtained for each independent sample using the same step as $BI(\hat{f}_{MTM})$, and is denoted by BI_i for $i = 1,2, \dots, N$. The P-value is then computed as $\widehat{Prob} = \frac{1}{N}\sum_{i=1}^{N} 1_{(BI(\hat{f}_{MTM}) \le BI_i)}$.

Combining the above steps we obtain the extended Bootlier procedure for regression setting, which can be summarized as follows.

1. Draw m bootstrap samples from data $\{t_1, t_2, \dots, t_n\}$ with each bootstrap sample of size n, and compute the sample statistic MTM as $\{MTM_1,\ MTM_2, \cdots,\ MTM_m\}$.
2. Obtain $\hat{f}_{MTM}$ and $BI(\hat{f}_{MTM})$.
3. Estimate the null distribution for $BI(\hat{f}_{MTM})$ and obtain P-value.

   (a) Draw N independent samples with each sample having n independent values drawn from the distribution $F(0, \hat{\sigma}^2)$ as $Z_i = (z_{i1}, z_{i2}, \dots, z_{in})'$ for $i = 1,2, \dots, N$.

   (b) Compute $(t_{i1}, t_{i2}, \dots, t_{in})' = (I - H)Z_i$ for $i = 1,2, \dots, N$ if $\{t_1, t_2, \dots, t_n\}$ is the ordinary residual, and compute $(t_{i1}, t_{i2}, \dots, t_{in})$ by (3.3), (3.4) or (3.5) if $\{t_1, t_2, \dots, t_n\}$ is the studentized residual, the studentized deletion residual, or SRCD.

   (c) Obtain BI_i for $\{t_{i1}, t_{i2}, \dots, t_{in}\}$, $i = 1,2, \dots, N$ by repeating 1&2.

   (d) Obtain P-value as $\frac{1}{N}\sum_{i=1}^{N} 1_{(BI(\hat{f}_{MTM}) \le BI_i)}$.

While the general results for limiting distribution of MTM for i.i.d. univariate sample are developed in Chapter 2, they do not apply directly because the residuals and Cook's distance are dependent. In the following section, we develop similar results for regression case, which explains why multimodality in the density plot of MTM is sensitive to the outliers (or influential observations) to provide the support for proposed procedure.

3.3 LARGE SAMPLE THEORY

In this section, we develop several theorems that explain the association between the multimodality in the density plot of MTM and the outliers (or the influential observations) in the sample. Throughout this section, we consider F be the normal distribution $N(0, \sigma^2)$. Lemma 3.1 introduce a basic result, the limiting distribution of the bootstrap sample mean for an increasing sequence converges to normal distribution given the sequence having certain properties. Those properties are proved to be possessed for residuals and SRCD in the later theorems so that Lemma 3.1 forms the basis for later theorems. Lemma 3.2 gives the result of limiting distribution of bootstrap sample mean for the residuals (or SCRD) when no outliers (or influential observations) are present in the sample, which is the immediate conclusion of Lemma 3.1. Theorem 3.3 and Theorem 3.5 present the limiting distribution of MTM with and without outliers (or influential observations) respectively, which are key results to support the proposed procedure. While the limiting distribution of MTM tends to normal distribution (Corollary 3.4) without outliers (or influential observations) present, it presents multimodality (Theorem 3.5) when there are outliers (or influential observation) in the sample.

Lemma 3.1

Let $X_{n1}, X_{n2}, \cdots, X_{nm}$ be i.i.d. random variables uniformly distribution on a sequence $\{c_{n1}, c_{n2}, \cdots, c_{nn}\}$, where $\frac{m}{n} \geq r$ for some constant $r > 0$. Let $\mu_n = \frac{1}{n}\sum_{i=1}^{n} c_{ni}$, $v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(c_{ni} - \mu_n)^2$ and $\bar{X}_{n+} = \frac{1}{m}\sum_{i=1}^{m} X_{ni}$. Suppose (a) $c_1 \leq v_n^2 \leq c_2$ for $0 < c_1 < c_2$, and (b) $\max_{1 \leq i \leq n} \frac{|c_{ni}|}{\sqrt{n}} \to 0$ as $n \to \infty$. Then $\sqrt{m}\frac{\bar{X}_{n+} - \mu_n}{v_n} \xrightarrow{D} N(0,1)$ as $n \to \infty$.

Lemma 3.1 can be considered as a special case of Linderberg-Feller central limit theorem. The proof is provided in Appendix A.1. If we consider $\{c_{n1}, c_{n2}, \cdots, c_{nn}\}$ as a sample from certain distribution, Lemma 3.1 provides the limiting distribution of its bootstrap sample mean. We notice the sequence $\{c_{n1}, c_{n2}, \cdots, c_{nn}\}$ do not necessarily come out from same distribution, and are not necessarily independent, but the bootstrap sample mean converges to normal distribution as long as the sample exhibits the asymptotic properties defined in (a) and (b). The motivation of this lemma is the residuals and influential measures that have the same properties.

The next theorem gives the limiting distribution of bootstrap sample mean for residuals and SRCD in (3.2) to (3.5). In this theorem, we require the elements of hat matrix are bounded by $\frac{c}{n}$ for some constant c, i.e., $max_{1\leq i,j\leq n}\,|h_{ij}| < \frac{c}{n}$ as $n \to \infty$. This condition is related to the independent variables only. The purpose is to limit the impact of outliers in the independent variable space by excluding arbitrary outliers in the independent variables. The assumptions generally hold in linear regression analysis. For example, in simple linear regression, $h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{\sum_{k=1}^{n}(x_k-\bar{x})^2}$ for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,n$. Suppose $\bar{x} \to \mu_X$ and $\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 \to \sigma_X^2 < \infty$ as $n \to \infty$. The assumption $max_{1\leq i,j\leq n}\,|h_{ij}| < \frac{c}{n}$ is equivalent to that X is bounded, which is a general case in the real data analysis.

Lemma 3.2

Assume the above condition holds. Let $\{t_1, t_2, \ldots, t_n\}$ be any residuals or SRCD in (3.2) to (3.5), $\mu_n = \frac{1}{n}\sum_{i=1}^{n} t_i$ and $v_n^2 = \frac{1}{n}\sum_{i=1}^{n} t_i^2$. Let $\{t_1^*, t_2^*, \ldots, t_m^*\}$ be a bootstrap sample from $\{t_1, t_2, \ldots, t_n\}$, where $\frac{m}{n} \geq r$ for some constant $r > 0$, $\bar{t}_m^* = \frac{1}{m}\sum_{i=1}^{m} t_i^*$ and $G_n(x) = P\left(\sqrt{m}\frac{\bar{t}_m^*-\mu_n}{v_n} \leq x\right)$. Then, $G_n(x) \to \Phi(x)$ as $n \to \infty$.

Lemma 3.2 shows that the bootstrap sample mean of the residuals and SRCD without outliers has the asymptotic normal distribution. Proof of Lemma 3.2 is provided in Appendix A.1.

The next theorem studies the limiting distribution of bootstrap sample mean of residuals and SRCD with an outlier present in the data. Let us augment the data by an outlier from upper side, say $Y_{n+1} = X_{n+1}\beta + \xi$, where ξ is sufficiently large to make t_{n+1} the largest residual or SRCD after fitting regression model.

Note we also let $m = n$ and $\sigma = 1$ for simplicity.

Theorem 3.3

Assume the conditions for independent variables in Lemma 3.2 hold. Let $\{t_1, t_2, \ldots, t_{n+1}\}$ be any residuals or SRCD in (3.2) to (3.5), $\mu_n = \frac{1}{n}\sum_{i=1}^{n} t_i$, $v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i - \mu_n)^2$, and

$\sigma_{n+1}^2 = \frac{1}{n+1}\sum_{i=1}^{n+1} {e_i}^2$. Let $\{t_1^*, t_2^*, \dots, t_n^*, t_{n+1}^*\}$ be a bootstrap sample from $\{t_1, t_2, \dots, t_n, t_{n+1}\}$, $\bar{t}_{n+1}^* = \frac{1}{n+1}\sum_{i=1}^{n+1} t_i^*$, and $G_{n+1}(x) = P\left(\sqrt{n+1}\,\frac{\bar{t}_{n+1}^* - \mu_n}{v_n} \le x\right)$. Then, $G_{n+1}(x)$ can be expressed as a mixture of normal distribution as

$$G_{n+1}(x) = \sum_{i=0}^{K} \frac{e^{-1}}{i!}\Phi\left(x - \frac{i\xi'}{\sqrt{n+1}}\right) + r_{n,K}\ , \tag{3.9}$$

where $\text{limsup}_{n\to\infty} |r_{n,K}| \le P(\mathcal{P}_1 > K)$ and $\mathcal{P}_1$ is a Poisson random variable with mean 1, K is any positive integer, $\xi' = \frac{\xi}{v_n}$ for ordinary residuals and studentized deletion residuals, $\xi' = \frac{\xi}{v_n \sigma_{n+1}}$ for studentized residual, and $\xi' = \frac{\sqrt{h_{(n+1)(n+1)}}}{\sqrt{p}\, v_n \sigma_{n+1}}\xi$ for SRCD.

Proof of Theorem 3.3 is provided in Appendix A.1.

COROLLARY 3.4

If $\xi/\sqrt{n} \to 0$ as $n \to \infty$, we have the limiting result $G_{n+1}(x) = \Phi(x) + o(1)$ a.s.. That is, unless ξ increases at a rate of $\sqrt{n}$ or faster as $n \to \infty$, its effect on the distribution of the normalized bootstrap mean vanishes in limit.

Theorem 3.3 and Corollary 3.4 show that the limiting distribution of bootstrap sample mean for residuals and SRCD in (3.2) to (3.5) with outliers converges to normal distribution, similar to the result we have in Lemma 3.2. It implies that the bootstrap sample mean is not sensitive to outliers.

Now we turn to the limiting distribution of MTM of bootstrap sample. With trimming size 1 from upper side of $\{t_1, t_2, \dots, t_n, t_{n+1}\}$,

$$\frac{1}{n}\sum_{k=1}^{n} t_{(k)}^* - \frac{1}{n+1}\sum_{k=1}^{n+1} t_{(k)}^* = \frac{1}{n}\left(\bar{t}_{n+1}^* - t_{(n+1)}^*\right).$$

It is sufficient to study the distribution of normalized $\bar{t}_{n+1}^* - t_{(n+1)}^*$.

Let $T_n = \sqrt{n+1}\,\frac{\bar{t}_{n+1}^* - t_{(n+1)}^* - \mu_n}{v_n}$, and let $H_n(x) = P(T_n \le x)$ that is further expressed as the mixture

$$H_n(x) = \lambda_{n,0}H_{n,0}(x) + \lambda_{n,1}H_{n,1}(x),$$

where $\lambda_{n,0} = 1 - \lambda_{n,1} = e^{-1} + o(1)$, $H_{n,0}(x) = P(T_n \le x | m(\xi) = 0)$, and $H_{n,1}(x) = P(T_n \le x | m(\xi) \ge 1)$.

The rescaled T_n for MTM of $\{t_1, t_2, \dots, t_n, t_{n+1}\}$ has a mixture of distributions $H_{n,0}$ and $H_{n,1}$, where $H_{n,0}$ is free of outliers and $H_{n,1}$ involves of outliers. Under the same assumption as Theorem 3.3, the following theorem reveals the representation of $H_{n,0}$ and $H_{n,1}$ as the sample size increases to infinity.

Theorem 3.5

Under the above setting and same assumptions as Theorem 3.3, $H_{n,0}(x)$ and $H_{n,1}(x)$ can be expressed as a mixture of normal distribution as

$$\text{(a) } H_{n,0}(x) = \sum_{i=1}^{K}(e^{-i+1} - e^{-i})\Phi\left(x + \frac{\sqrt{n+1}}{v_n}t_{(n-i+1)}\right) + s_{n,K}\ , \tag{3.10}$$

where $\limsup_{n\to\infty} |s_{n,K}| \le e^{-K}$, K is any positive integer,

$$\text{(b) } H_{n,1}(x) = \sum_{i=1}^{K}\frac{1}{(e-1)i!}\Phi\left(x + \frac{\sqrt{n-i+1}}{v_n}t_{n+1}\right) + t_{n,K}\ , \tag{3.11}$$

where $\limsup_{n\to\infty} |t_{n,K}| \le \frac{1}{1-e^{-1}}P(\mathcal{P}_1 > K)$ and $\mathcal{P}_1$ is a Poisson random variable with mean 1, K is any positive integer.

For (b), if $\xi/\sqrt{n} \to 0$, it can be further expressed as

$$\text{(c) } H_{n,1}(x) \to \Phi\left(x + \sqrt{n+1}\xi'\right), \tag{3.12}$$

where $\xi' = \xi$ for residuals (3.2) to (3.4) and $\xi' = \frac{\sqrt{h_{(n+1)(n+1)}}}{\sqrt{p}v_n}\xi$ for SRCD.

Proof of Theorem 3.5 is provided in Appendix A.1.

Without proof, for certain special cases where the error terms are from some short-tailed distribution such that we could further assume that $\sqrt{n}(t_{(n)} - t_{(n-i)}) \to 0$ a.s. for any fixed i, $H_n(x)$ of Theorem 3.5 can be further simplified as a mixture of two normal distributions,

$$H_n(x) \to e^{-1}\Phi\left(x + \sqrt{n+1}\,t_{(n)}\right) + (1 - e^{-1})\Phi\left(x + \sqrt{n+1}\,\xi'\right), \tag{3.13}$$

where $\xi' = \xi$ for residuals (3.2) to (3.4) and $\xi' = \frac{\sqrt{h_{(n+1)(n+1)}}}{\sqrt{p}v_n}\xi$ for SRCD.

The representations in (3.10) to (3.13) reveal that multimodality of limiting distribution of MTM for residuals and SRCD is caused by the outlier. Lemma 3.2, Theorem 3.3 and 3.5 are developed for one outlier in the linear regression problem. When there are multiple outliers present in the sample, these results are similar to those discussed in Chapter 2. We omit them here.

3.4 SIMULATION STUDY

In this section, we conduct two simulation studies to illustrate the performance and features of the extended Bootlier procedure. The first simulation study illustrates a regression analysis with a data point which is both an outlier and an influential observation. The second simulation study demonstrates that the extended Bootlier procedure has lower false positive rate when the error terms distributed with high probability in the tail. In particular, the results from those analytic methods discussed in Section 3.2.1 are provided in both simulation studies for comparison.

In the simulation study #1, a sample size of 30 $\{(y_i, x_i); i = 1,2,\ldots,30\}$ is simulated to fit a linear regression model $\boldsymbol{Y} = \boldsymbol{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$. The independent variables $\{x_i; i = 1,2,\ldots,30\}$ are simulated from Uniform [-5, 5], error terms $\{\varepsilon_i; i = 1,2,\ldots,30\}$ are simulated from standard normal $N(0,1)$, and $\boldsymbol{Y}$ is calculated by $y_i = 1 + x_i + \varepsilon_i$. An additional data point $(0,4.72)$ is added to the sample as an outlier (Note: $P(X > 3.72) = 0.0001$ where $X \sim N(0,1)$). The sample data with the fitted regression line is plotted in Figure 3.1 (a). Figure 3.1 (b) to (e) are plots of residuals and SRCD. The corresponding density plot of MTM and sample Bootlier index are provided in Figure 3.1 (f) to (i).

For Figure 3.1 (b) to (i), the residuals and SRCD of the added data point are well separated from others and clear multimodality of the density plot is presented for all three residuals and SRCD. P-values are 0.001, 0.001, 0.001 and 0.024 for ordinary residuals, studentized residuals, studentized deletion residuals and SRCD respectively. These present the strong evidence that the added point is both an outlier and an influential observation. We notice that although the four measures all provide significant P-values, P-value for SRCD does not have same strength the other three for this example. A close look of the residuals and SRCD explains the reason. As seen in (3.2) to (3.4), the three residuals are usually at same scale when there are no obviously high leverage points in the sample or when the high leverage points are close to the regression line (i.e., with small error). When there are high leverage points with large error, the studentized deletion residual is most sensitive to the outliers among the three, and studentized residual is more sensitive to the outliers than ordinary residual. However, the difference among residuals is mainly determined by the weights $\frac{1}{\sqrt{1-h_{ii}}}$ for $i = 1,2,\dots,n$, which usually does not present large variability. In this simulation study, $\frac{1}{\sqrt{1-h_{ii}}}$ only ranges from 1.02 to 1.07. Therefore the results based on residuals generally are consistent. But SRCD is very different from residuals because it depends on leverage through the weight of $\frac{\sqrt{h_{ii}}}{1-h_{ii}}$. Even without the high leverage points, its variability is much higher. The data points with relative higher leverage are more likely to have large SRCD values than those with small leverage. In this simulation study, the range of $\frac{\sqrt{h_{ii}}}{1-h_{ii}}$ is $[0.19, 0.41]$ and the added point has the value (0.19) at the lower end of range, which mathematically explains why P-value for SRCD is on a less severe scale than the residuals. Essentially SRCD is a measure of influential observations.

The studentized residual (3.502) for the added point is the only value greater than 3. It is considered as an outlier by Montgomery et al. (2003). This value is also beyond the 1% critical value for the discordancy test proposed in Barnett and Levis (1994). For Bonferroni type test by Neter et al. (1996), the studentized deletion residual (4.529) for the added point is greater than the 1% critical value. These results are consistent with the extended Bootlier procedure to confirm the added point is an outlier. For Cook's distance, the added point has value 0.205 that is beyond the cutoff (i.e., $\frac{4}{31}$) suggested by Bollen and Jackman (1990), which agrees the conclusion from the extended Bootlier procedure.

[Insert Figure 3.1]

The simulation study #2 investigates a linear regression analysis with the error terms distributed as the standard double exponential distribution $DB(0,1)$ and no outliers of influential observations added. A sample size of 100, $\{(y_i, x_i); i = 1,2, \dots, 100\}$, is simulated to fit a linear regression model $\boldsymbol{Y} = \boldsymbol{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, with the $x_i{}'s$ drawn from Uniform [-5, 5] and $\varepsilon_i's$ drawn from $DB(0,1)$ and $\boldsymbol{Y}$ calculated by $y_i = 1 + x_i + \varepsilon_i$. The sample data with the fitted regression line is plotted in Figure 3.2.

[Insert Figure 3.2]

The observation #30 has the studentized residual -3.507 thus it is considered as an outlier by Montgomery et al. (2003). This value is also beyond the 5% critical value for the discordancy test proposed by Barnett and Levis (1994). The studentized deletion residual is -3.731 for the observation #30 that is beyond the 5% critical value for Bonferroni type test by Neter et al. (1996). Therefore, observation #30 is considered as an outlier. For Cook's distance, there are several observations beyond the cutoff (i.e., $\frac{4}{100}$) suggested by Bollen and Jackman (1990). While all those analytic approaches suggest the outliers or influential observations in the data, the extended Bootlier procedure does not have the significant findings. P-values for the lower-side and upper sided trimming are 0.298 and 0.212 for the ordinary residual, 0.304 and 0.219 for the studentized residual, 0.284 and 0.217 the for studentized deletion residual, 0.539 and 0.503 for SRCD. None of these P-values are significant. In other words, the extended Bootlier procedure has lower false positive rate for such cases. This is contributed to the fact the procedure inherently takes all data points into account, which Singh and Xie (2003) describe as "seems to capture the essence of what statisticians call outliers". When there are data points filling in the gaps between those extreme values and the rest of data, it is more reasonable to consider them from some heavy-tailed distribution rather than outliers.

3.5 REAL DATA EXAMPLE

In the real data example, we analyze the sperm data studied in Clarke (2008). The dataset includes sixty observations of bivariate data (Table 3.1), which are the Sperm Motility Index (SMI) and sperm motility (motility) as measured using the SQA-I1 B machine. Their objective of the analysis is to illustrate the usefulness of diagnostic approaches and the adaptive trimmed likelihood algorithm (ATLA) that are aiming at the detection of outliers in exploring the relationship between those two variables.

[Insert Table 3.1]

The regression model is first fitted with all data points with SMI as the response variable and motility as the independent variable, then the extended Bootlier procedure is applied to the residuals and SRCD. The sample data with the fitted regression line is plotted in Figure 3.3 (a). Figure 3.3 (b) to (e) are plots of residuals and SRCD and Figure 3.2 (f) to (i) are the corresponding density plots of MTM.

P-values for sample Bootlier index are 0.004, 0.002, 0.003 and 0.038 for ordinary residuals, studentized residuals, studentized deletion residuals and SRCD respectively. The presence of outliers and influential observations in data is apparent.

[Insert Figure 3.3]

Similar to the diagnostic approach in Clarke (2008), we first exclude two largest outliers, observation #1 and #2, fit the linear regression model with the rest of data, and then apply the extended Bootlier procedure to residuals and SRCD again. Figure 3.4 (a) to (d) are the density plots with sample Bootlier index. *P*-values are 0.020, 0.021, 0.024 and 0.003 for ordinary residuals, studentized residuals, studentized deletion residuals and SRCD respectively, which still suggest the presence of outliers and influential observations. This is consistent with the results in Clarke (2008).

[Insert Figure 3.4]

Next, four outliers, observation #1, #2, #3 and #15, are excluded from the model and the extended Bootlier procedure is applied to residuals and SRCD afterwards; see Figure 3.5 (a) to (d) for the density plots with sample Bootlier index. *P*-values are >0.999, >0.999, >0.999 and 0.045 for ordinary residuals, studentized residuals, studentized deletion residuals and SRCD respectively. The results for residuals suggest no outliers, but SRCD suggests borderline influential observations. The investigation of SRCD reveals that observation #4 has a large residual (studentized residual 2.55) and a high leverage in X, which are the largest values among studentized residuals and leverages respectively. This is a good example that an observation is not an outlier but an influential observation.

[Insert Figure 3.5]

In Clarke (2008), the diagnostic approach does not go further to remove more outliers or influential observations, and the regression model is fitted with 56 observations. In our study we further exclude observation #4 and apply the extended Bootlier procedure to the rest of data. Figure 3.6 (a) to (d) shows the density plots for the residuals and SRCD. *P*-values are >0.999, >0.999, >0.999 and 0.451. There is no evidence of outliers or influential observations. This is consistent with the results by the ATLA method in Clarke (2008) given the number of potential outliers is up to five. While the ATLA does not continue checking more outliers or influential observations due to the explosion of computation, not because of no outlier or influential observations, the extended Bootlier procedure suggests no further outliers or influential observations in the data.

[Insert Figure 3.6]

3.6 CONCLUDING REMARKS

This chapter has proposed a statistical framework, the extended Bootlier procedure, as a diagnostic method for the outliers and the influential observations in linear regression model. The framework is an extension of the similar method that is developed for i.i.d. univariate sample. Three residuals as the measures of outliers and a square-version of Cook’s distance as the measure of influential observations are studied using the extended

Bootlier procedure. As suggested, other measures could be applied after careful examinations. The features of the proposed framework are illustrated through two simulation studies and a real data example, and the procedure demonstrates advantages when the error terms have higher tail probability than normal distribution. Because the residuals and influential measures are dependent, the large sample results are developed to provide the theoretical support. Finally, the interesting area for further study may include a sequential procedure similar to the generalized extreme generalized extreme studentized deviate (ESD); see Rosner (1983).

APPENDIX

A.1 Proof of Results

PROOF OF LEMMA 3.1

Let $Y_{ni} = \frac{X_{ni}-\mu_n}{v_n}$, $i = 1,2 \dots, m$, then $E(Y_{ni}) = 0$ and $E(Y_{ni}^2) = 1$. Let $\varphi_{ni}(t) = E\left(e^{it\frac{Y_{ni}}{\sqrt{m}}}\right)$, then any $\varepsilon > 0$,

$$\left|\varphi_{ni}(t) - (1 - \tfrac{t^2}{2m})\right| \le E\, min\left(\left|t\,\tfrac{Y_{ni}}{\sqrt{m}}\right|^3, 2\left|t\,\tfrac{Y_{ni}}{\sqrt{m}}\right|^2\right)$$

$$\le E\left(\left|t\,\tfrac{Y_{ni}}{\sqrt{m}}\right|^3 ; \left|\tfrac{Y_{ni}}{\sqrt{m}}\right| < \varepsilon\right) + E\left(2\left|t\,\tfrac{Y_{ni}}{\sqrt{m}}\right|^2 ; \left|\tfrac{Y_{ni}}{\sqrt{m}}\right| \ge \varepsilon\right)$$

$$\le \frac{\varepsilon t^3}{m} + 2t^2 E\left(\left|\tfrac{Y_{ni}}{\sqrt{m}}\right|^2 ; \left|\tfrac{Y_{ni}}{\sqrt{n}}\right| \ge \sqrt{\tfrac{m}{n}}\,\varepsilon\right)$$

$$= \frac{\varepsilon t^3}{m} + \frac{2t^2}{m} E\left({Y_{ni}}^2 ; |X_{ni} - \mu_n| \ge v_n\sqrt{n}r\varepsilon\right).$$

Then,

$$\sum_{i=1}^{m}\left|\varphi_{ni}(t) - \left(1 - \frac{t^2}{2m}\right)\right| \le \varepsilon t^3 + 2t^2 E\left({Y_{n1}}^2 ; |X_{n1} - \mu_n| \ge v_n\sqrt{n}r\varepsilon\right). \qquad \text{(A.1)}$$

By Lemma 3.1 (b),

$$E\left(Y_{n1}^2 ; |X_{n1} - \mu_n| \ge v_n\sqrt{n}r\varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Then,

$$\text{limsup}_{n\to\infty} \sum_{i=1}^{m}\left|\varphi_{ni}(t) - (1 - \tfrac{t^2}{2m})\right| \le \frac{\varepsilon t^3}{6}. \qquad \text{(A.2)}$$

By Lemma (3.4.3) of Durett (2010),

$$\left|\prod_{i=1}^{m}\varphi_{ni}(t) - \prod_{i=1}^{m}(1 - \tfrac{t^2}{2m})\right| \to 0 \text{ as} \to \infty. \qquad \text{(A.3)}$$

Then we have

$$\prod_{i=1}^{m} \varphi_{ni}(t) \to e^{-\frac{x^2}{2}} \text{ as } n \to \infty, \text{ which implies } \sqrt{m}\frac{\bar{X}_{n+}-\mu_n}{v_n} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$

PROOF OF LEMMA 3.2

Given that $\{t_1, t_2, \dots, t_n\}$ is the residuals or SRCD defined in (3.2) to (3.5), we will exam conditions (a) and (b) of Lemma 3.1 for each residuals and SRCD.

Ordinary residuals

(a) $\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{n-p}{n}\left(\frac{1}{n-p}\sum_{i=1}^{n} e_i^2\right) \to \sigma^2$ as $n \to \infty$

(b) $\max_{1\le i\le n}\frac{|e_i|}{\sqrt{n}} \to 0$ as $n \to \infty$, which is obtained by the same arguments of Lemma A of Singh and Xie (2003).

Studentized residuals

(a) $\frac{1}{n}\sum_{i=1}^{n} t_i^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}$

$\frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})} \le \frac{1}{\hat{\sigma}^2}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right)\max_{1\le i\le n}\frac{1}{1-h_{ii}} \to 1$ as $n \to \infty$

$\frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})} \ge \frac{1}{\hat{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n} e_i^2 \to 1$ as $n \to \infty$

Then $\frac{1}{n}\sum_{i=1}^{n} t_i^2 \to 1$ as $n \to \infty$.

(b) $\max_{1\le i\le n}\frac{|t_i|}{\sqrt{n}} \le \frac{1}{\sqrt{\hat{\sigma}^2}}\max_{1\le i\le n}\frac{1}{\sqrt{1-h_{ii}}}\max_{1\le i\le n}\frac{|e_i|}{\sqrt{n}} \to 0$ as $n \to \infty$

Studentized deletion residuals

Note that $\hat{\sigma}_{(-i)}^2 = \frac{(n-p)\hat{\sigma}^2}{n-p-1} + \frac{e_i^2}{(n-p-1)(1-h_{ii})}$ (see Beckman and Tryssell, 1974), the studentized deletion residuals (3.4) can also be expressed as $s_i = e_i\sqrt{\frac{n-p-1}{(1-h_{ii})(n-p)\hat{\sigma}^2-e_i^2}}$.

Let $g(e_i, h_{ii}) = \sqrt{\frac{n-p-1}{(1-h_{ii})(n-p)\hat{\sigma}^2-e_i^2}}$, then $t_i = g(e_i, h_{ii})e_i$.

(a) $\frac{1}{n}\sum_{i=1}^{n} t_i^2 = \frac{1}{n}\sum_{i=1}^{n} g(e_i, h_{ii})^2 e_i^2 \to 1$ as $n \to \infty$

(b) $\max_{1\le i\le n} \frac{|t_i|}{\sqrt{n}} = \max_{1\le i\le n} \frac{|g(e_i,h_{ii})e_i|}{\sqrt{n}} \to 0$ as $n \to \infty$

SRCD

The coefficient for the SRCD $\frac{\sqrt{h_{ii}}}{\sqrt{p\hat{\sigma}^2}(1-h_{ii})}$ goes to 0 as $n \to \infty$. So, we rescale SRCD by a common constant to fit the proof into our frame. Let $t_i' = \sqrt{n} t_i = \frac{\sqrt{n}}{\sqrt{p\hat{\sigma}^2}} \frac{\sqrt{h_{ii}}}{1-h_{ii}} e_i$. Next, we exam the conditions (a) to (b) of Lemma 3.1 for $\{t_1', t_2', \dots, t_n'\}$.

(a) $\frac{1}{n}\sum_{i=1}^{n} {t_i'}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{nh_{ii}}{p\hat{\sigma}^2} \frac{e_i^2}{(1-h_{ii})^2}$

$\frac{1}{n}\sum_{i=1}^{n} \frac{nh_{ii}}{p\hat{\sigma}^2} \frac{e_i^2}{(1-h_{ii})^2} \le \frac{c}{p\hat{\sigma}^2} \max_{1\le i\le n} \frac{1}{(1-h_{ii})^2} \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) \to \frac{c}{p}$ as $n \to \infty$, and

$\frac{1}{n}\sum_{i=1}^{n} \frac{nh_{ii}}{p\hat{\sigma}^2} \frac{e_i^2}{(1-h_{ii})^2} \ge \frac{1}{p\hat{\sigma}^2} \frac{1}{n}\sum_{i=1}^{n} e_i^2 \to \frac{1}{p}$ as $n \to \infty$

Then $\frac{1}{p} \le \frac{1}{n}\sum_{i=1}^{n} {t_i'}^2 \le \frac{c}{p}$

(b) $\max_{1\le i\le n} \frac{|t_i'|}{\sqrt{n}} \le \frac{\sqrt{c}}{\sqrt{p\hat{\sigma}^2}} \max_{1\le i\le n} \frac{1}{1-h_{ii}} \max_{1\le i\le n} \frac{|e_i|}{\sqrt{n}} \to 0$ as $n \to \infty$

By Lemma 3.1, we conclude that $G_n(x) = P\left(\sqrt{m}\frac{\bar{t}_m^* - \mu_n}{v_n} \le x\right) \to \Phi(x)$ as $n \to \infty$.

PROOF OF THEOREM 3.3

Let $m(\xi)$ be the number of times t_{n+1} appearing in the bootstrap sample $\{t_1^*, t_2^*, \dots, t_n^*, t_{n+1}^*\}$. Then $m(\xi)$ has $Binomial(n+1, \frac{1}{n+1})$. When $n \to \infty$, $m(\xi)$ is asymptotically distributed as Poisson random variable with mean 1, and is denoted by $\mathcal{P}_1$.

The distribution of $\sqrt{n+1}\frac{\bar{t}_{n+1}^* - \mu_n}{v_n}$ is expanded as

$$G_{n+1}(x) = \sum_{i=0}^{K} P\left(\sqrt{n+1}\frac{\bar{t}_{n+1}^* - \mu_n}{v_n} \le x, m(\xi) = i\right) + r_{n,K}, \tag{A.4}$$

where $r_{n,K} < P(m(\xi) > K)$.

Let $\{t^*_{(1)}, t^*_{(2)}, \dots, t^*_{(n)}, t^*_{(n+1)}\}$ be the order statistics of $\{t^*_1, t^*_2, \dots, t^*_n, t^*_{n+1}\}$. By assumption, $t^*_{(n+1)} = t_{n+1}$ if t_{n+1} is in $\{t^*_1, t^*_2, \dots, t^*_n, t^*_{n+1}\}$.

$$P\left(\sqrt{n+1}\,\frac{\bar{t}^*_{n+1}-\mu_n}{v_n} \le x, m(\xi) = i\right)$$

$$= P(m(\xi) = i)P\left(\sum_{j=1}^{n-i+1} t^*_{(j)} \le \sqrt{n+1}v_n x + (n+1)\mu_n - i t_{n+1} | m(\xi) = i\right)$$

$$= P(m(\xi) = i)P\left(\frac{\sqrt{n-i+1}}{v_n}\left(\frac{\sum_{j=1}^{n-i+1} t^*_{(j)}}{n-i+1} - \mu_n\right) \le \frac{\sqrt{n+1}}{\sqrt{n-i+1}}x + \frac{i}{v_n\sqrt{n-i+1}}\mu_n - \frac{i}{v_n\sqrt{n-i+1}}t_{n+1} | m(\xi) = i\right).$$

(A.5)

For $t^*_{(j)}$'s, $j = 1,2,\dots,n-i+1$, they are a bootstrap sample from $\{t_1, t_2, \dots, t_n\}$ given that $m(\xi) = i$. Let's rewrite $t^*_{(j)}$'s as $t^*_{j,[n]}$, a form without order. Next, we show that $\frac{\sqrt{n-i+1}}{v_n}\left(\frac{\sum_{j=1}^{n-i+1} t^*_{j,[n]}}{n-i+1} - \mu_n\right) \xrightarrow{D} N(0,1)$ given $m(\xi) = i$ for each residual and SRCD.

We first consider ordinary residuals. That is, $t_j = e_j$ for $j = 1,2,\dots,n,n+1$. Correspondingly we denote $t^*_{j,[n]}$ by $e^*_{j,[n]}$.

Let $w_j = e_j + h_{j(n+1)}\xi$ for $j = 1,2,\dots,n$, then we have $e_j = w_j - h_{j(n+1)}\xi$ for $j = 1,2,\dots,n$, while $e_{n+1} = (1 - h_{(n+1)(n+1)})\xi$.

If we denote $\bar{w}_n = \frac{1}{n}\sum_{j=1}^n w_j$ and $\bar{h}_n = \frac{1}{n}\sum_{j=1}^n h_{j(n+1)}$, the following hold for μ_n and v_n^2.

$$\mu_n = \bar{w}_n - \bar{h}_n\,\xi$$

$$v_n^2 = \frac{1}{n}\sum_{j=1}^n (e_j - \mu_n)^2 = \frac{1}{n}\sum_{j=1}^n \left(w_j - \bar{w}_n\right)^2 + \left(\frac{\xi}{n}\right)^2 \frac{1}{n}\sum_{j=1}^n \left(n h_{j(n+1)} - n\bar{h}_n\right)^2$$

Adopting the same notations, let $e^*_{j,[n]} = w^*_{j,[n]} + h^*_{j(n+1),[n]}\xi$ for $j = 1,2,\dots,n-i+1$, where $w^*_{j,[n]}$ and $h^*_{j(n+1),[n]}$ are the two components of $e^*_{j,[n]}$ corresponding to w_j's and $h_{j(n+1)}$'s.

Then $w_{j,[n]}^{*}$ and $h_{j(n+1),[n]}^{*}\xi$ can be considered as bootstrap samples from $\{w_1, w_2, \dots, w_n\}$ and $\{h_{1(n+1)}\xi, h_{2(n+1)}\xi, \dots, h_{n(n+1)}\xi\}$ respectively.

We have

$$\frac{\sqrt{n-i+1}}{v_n}\left(\frac{\sum_{j=1}^{n-i+1} t_{j,[n]}^{*}}{n-i+1}-\mu_n\right)=\sqrt{n-i+1}\,\frac{\frac{\sum_{j=1}^{n-i+1} w_{j,[n]}^{*}}{n-i+1}-\bar{w}_n+\left(\frac{\sum_{j=1}^{n-i+1} nh_{j(n+1),[n]}^{*}}{n-i+1}-n\bar{h}_n\right)\frac{\xi}{n}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(w_j-\bar{w}_n\right)^2+\left(\frac{\xi}{n}\right)^2\frac{1}{n}\sum_{j=1}^{n}\left(nh_{j(n+1)}-n\bar{h}_n\right)^2}}. \text{ (A.6)}$$

When $\frac{1}{n}\sum_{j=1}^{n}\left(nh_{jn}-n\bar{h}_n\right)^2=0$, which represents a special case that X_{n+1} resides at the center of X cloud, (A.6) is simplified as $\sqrt{n-i+1}\,\frac{\frac{\sum_{j=1}^{n-i+1} w_{j,[n]}^{*}}{n-i+1}-\bar{w}_n}{\frac{1}{n}\sum_{j=1}^{n}\left(w_j-\bar{w}_n\right)^2}$. By the same arguments for Lemma 3.2, we can conclude that, given $m(\xi)=i$,

$$P\left(\sqrt{n-i+1}\,\frac{\frac{\sum_{j=1}^{n-i+1} w_{j,[n]}^{*}}{n-i+1}-\bar{w}_n}{\frac{1}{n}\sum_{i=1}^{n}(w_i-\bar{w}_n)^2}\le x\right)\to\Phi(x) \text{ as } n\to\infty.$$

When $n\sum_{i=1}^{n}\left(h_{in}-\bar{h}_n\right)^2>0$, which represents that $h_{j(n+1)}$ for $j=1,2,\dots,n$ have a fixed variation. We consider the following three cases assuming $\lim_{n\to\infty}\frac{\xi}{n}$ exists for simplicity.

If $\lim_{n\to\infty}\frac{\xi}{n}=0$, (A.6) is dominated by the first component $\frac{\sum_{j=1}^{n-i+1} w_{j,[n]}^{*}}{n-i+1}-\bar{w}_n$. If $\lim_{n\to\infty}\frac{\xi}{n}=\infty$, (A.6) is dominated by the second component $\left(\frac{\sum_{j=1}^{n-i+1} nh_{j(n+1),[n]}^{*}}{n-i+1}-n\bar{h}_n\right)\frac{\xi}{n}$. If $0<\lim_{n\to\infty}\frac{\xi}{n}<\infty$, both components contribute to the distribution in (A.6). In all three cases, we can apply the same arguments for Lemma 3.2 to get the following results, given $m(\xi)=i$,

$$P\left(\sqrt{n-i+1}\,\frac{\frac{\sum_{j=1}^{n-i+1} w_{j,[n]}^{*}}{n-i+1}-\bar{w}_n+\left(\frac{\sum_{j=1}^{n-i+1} nh_{j(n+1),[n]}^{*}}{n-i+1}-n\bar{h}_n\right)\frac{\xi}{n}}{\frac{1}{n}\sum_{j=1}^{n}\left(w_j-\bar{w}_n\right)^2+\left(\frac{\xi}{n}\right)^2 n\sum_{j=1}^{n}\left(nh_{j(n+1)}-n\bar{h}_n\right)^2}\le x\right)\to\Phi(x) \text{ as } n\to\infty.$$

Therefore, (A.5) becomes

$$P\left(\sqrt{n+1}\,\frac{\bar{t}^*_{n+1}-\mu_n}{v_n} \le x, m(\xi)=i\right) \to \frac{e^{-1}}{i!}\Phi\left(x-\frac{i}{v_n\sqrt{n+1}}\xi\right). \tag{A.7}$$

From (A.4) and (A.7), we conclude (3.6),

$$G_{n+1}(x) = \sum_{i=0}^{K}\frac{e^{-1}}{i!}\Phi\left(x-\frac{i}{v_n\sqrt{n+1}}\xi\right)+r_{n,K} \text{ as } n\to\infty.$$

The results for the other residuals and SRCD can be obtained by similar arguments as above. We do not repeat the arguments here.

PROOF OF THEOREM 3.5

The proof of (a) will follow same arguments of the proof of Theorem 2.3 (a); see Appendix A.1 of Chapter 2. We do not repeat the arguments here.

For (b), $H_{n,1}(x)$ can be expressed as

$$H_{n,1}(x) = \frac{1}{P(m(\xi)>0)}\sum_{i=1}^{K}P(T_{n+1}\le x, m(\xi)=i)+t_{n,K},$$

where $\limsup_{n\to\infty}|t_{n,K}| \le \frac{1}{1-e^{-1}}P(\mathcal{P}_1 > K)$.

For $i \ge 1$,

$$P(T_n \le x | m(\xi)=i)$$

$$= P\left(\sqrt{n+1}\,\frac{\bar{t}^*_{n+1}-t^*_{(n+1)}-\mu_n}{v_n} \le x | m(\xi)=i\right)$$

$$= P\left(\sum_{j=1}^{n+1} t^*_{(j)} \le \sqrt{n+1}v_n x + (n+1)\mu_n + (n+1)t_{n+1} | m(\xi)=i\right)$$

$$= P\left(\sum_{j=1}^{n-i+1} t^*_{(j)} \le \sqrt{n+1}v_n x + (n+1)\mu_n + (n-i+1)t_{n+1} | m(\xi)=i\right)$$

$$= P\left(\frac{\sqrt{n-i+1}}{v_n}\left(\frac{\sum_{j=1}^{n-i+1}t^*_{(j)}}{n-i+1}-\mu_n\right) \le \frac{\sqrt{n+1}}{\sqrt{n-i+1}}x + \frac{i}{v_n\sqrt{n-i+1}}\mu_n + \frac{\sqrt{n-i+1}}{v_n}t_{n+1} | m(\xi)=i\right).$$

Applying the same arguments as proof in Lemma 3.2, we have

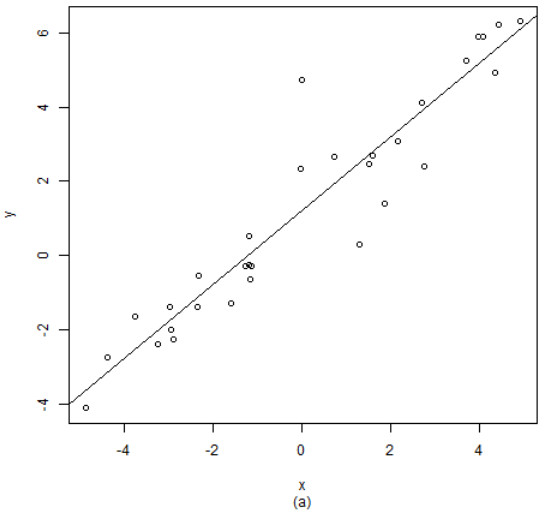
$$P(T_n \le x | m(\xi) = i) \to \Phi\left(x + \frac{\sqrt{n-i+1}}{v_n} t_{n+1}\right).$$

Then $H_{n,1}(x) = \sum_{i=1}^{K} \frac{1}{(e-1)i!} \Phi\left(x + \frac{\sqrt{n-i+1}}{v_n} t_{n+1}\right) + t_{n,K}$.

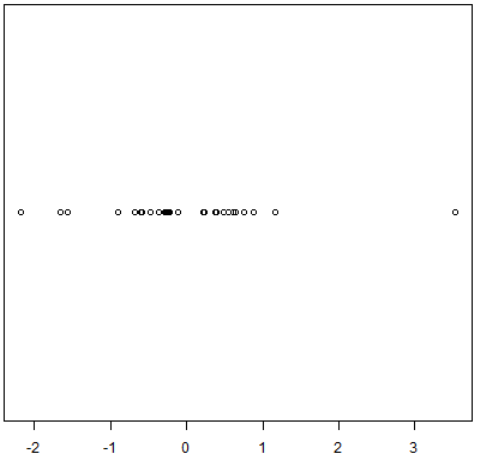
Proof of (c) is trivial.

A.2 Figures and Tables

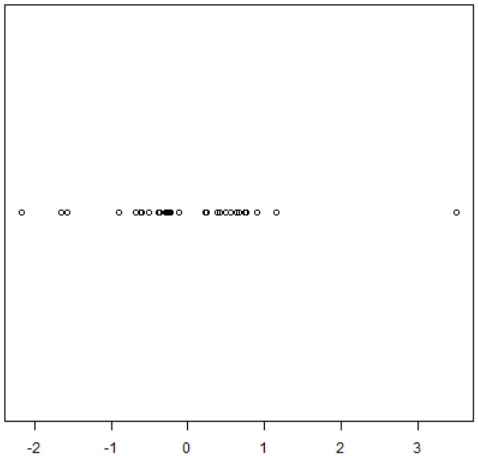
Figure 3.1 (Simulation #1): Extended Bootlier procedure for a sample of size 30 simulated for a univariate regression model with an outlier added: (a) 31 data points in (Y, X) plane with fitted regression line, (b) ordinary residuals, (c) studentized residuals, (d) studentized deletion residuals, (e) SRCD, (f) the density plot for ordinary residuals, sample Bootlier index = 0.80838, (g) the density for studentized residuals, sample Bootlier index = 0.81510, (h) the density plot for studentized deletion residuals, sample Bootlier index = 1.11021, (i) the density plot for SRCD, sample Bootlier index = 0.45858



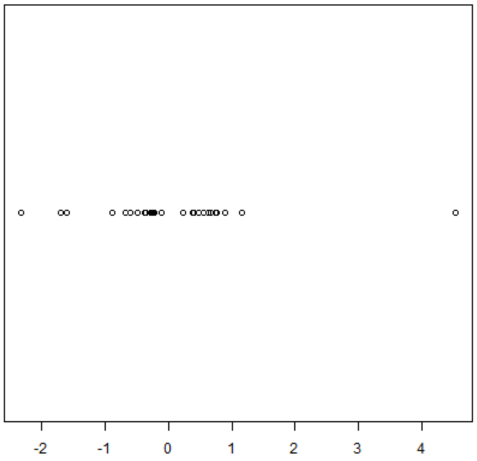
(a)



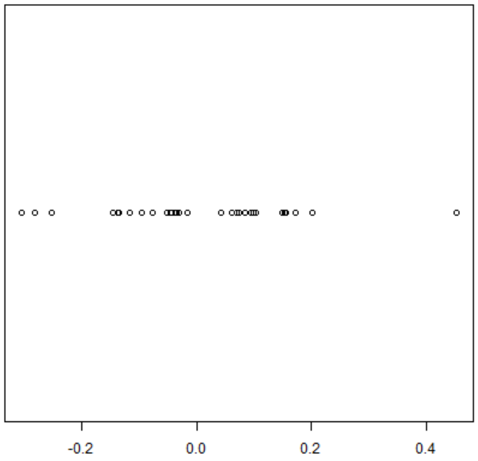
(b)



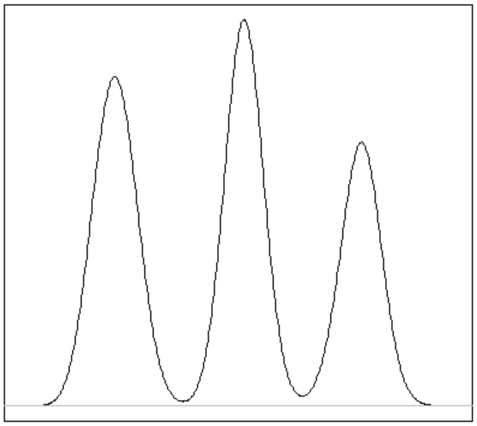
(c)



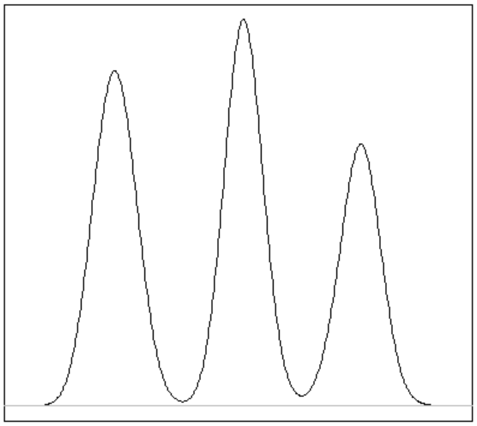
(d)



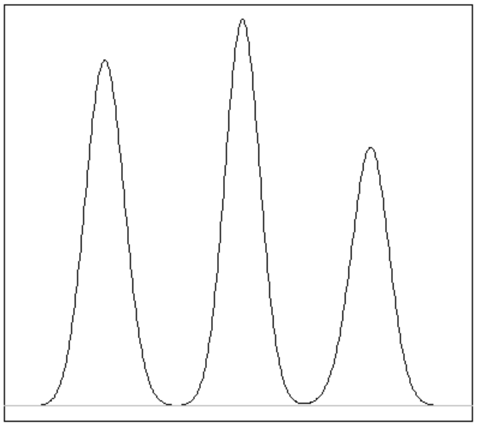
(e)



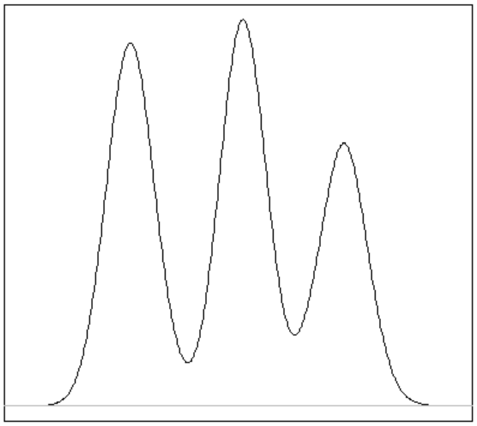
(f)



(g)



(h)



(i)

Figure 3.2 (Simulation #2): Extended Bootlier procedure for a sample of size 100 simulated for a univariate regression model.

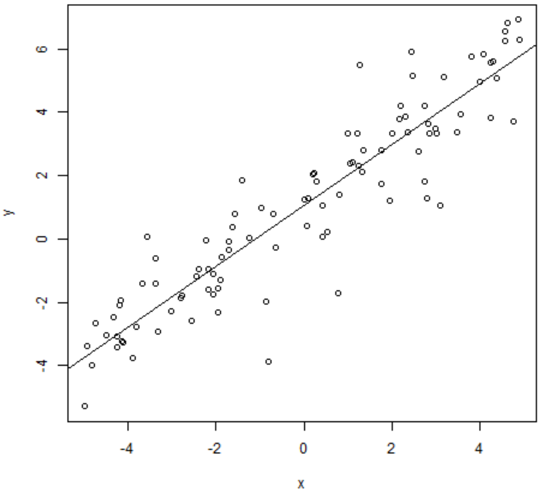
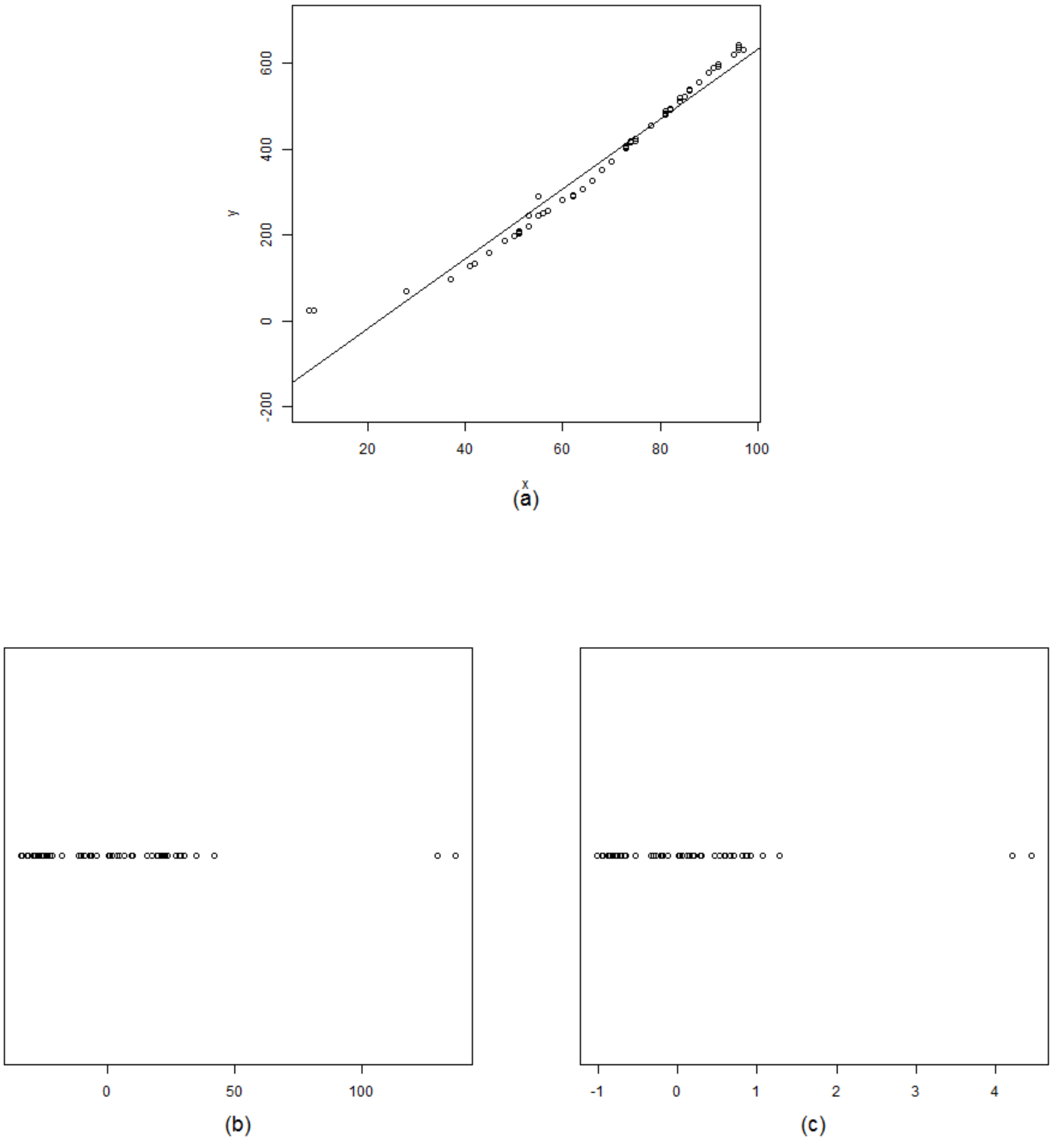


Table 3.1 (Real Data Example): Sperm data in Clarke (2008)

# Obs.	Motility	SMI	# Obs.	Motility	SMI	# Obs.	Motility	SMI
1	8	23	21	62	291	41	81	481
2	9	24	22	62	293	42	82	492
3	28	70	23	64	307	43	82	494
4	37	98	24	66	326	44	82	491
5	41	128	25	66	326	45	84	510
6	42	134	26	66	326	46	84	519
7	45	158	27	68	351	47	85	521
8	48	186	28	70	372	48	86	537
9	50	198	29	70	372	49	86	539
10	51	208	30	73	407	50	88	556
11	51	206	31	73	403	51	88	556
12	51	204	32	73	405	52	90	579
13	53	245	33	74	416	53	91	589
14	53	220	34	74	418	54	92	597
15	55	290	35	75	424	55	92	592
16	55	246	36	75	420	56	95	621
17	56	250	37	78	455	57	96	643
18	56	252	38	81	484	58	96	636
19	57	256	39	81	480	59	96	631
20	60	281	40	81	489	60	97	632

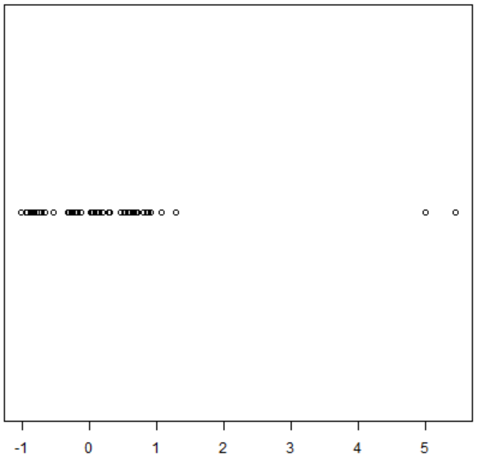
Figure 3.3 (Real Data Example): Extended Bootlier procedure for Sperm data: (a) 60 data points in (Y, X) plane with fitted regression line, (b) ordinary residuals, (c) studentized residuals, (d) studentized deletion residuals, (e) SRCD, (f) the density plot for ordinary residuals, sample Bootlier index = 0.77918, (g) the density plot for studentized residuals, sample Bootlier index = 0.82472, (h) the density plot for studentized deletion residuals, sample Bootlier index = 0.86479, (i) the density plot for SRCD, sample Bootlier index = 1.08405



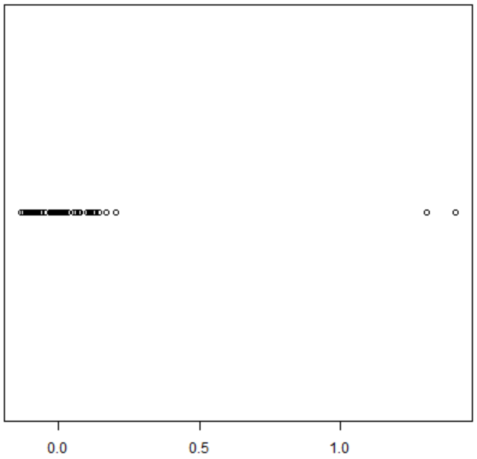
(a)

(b)

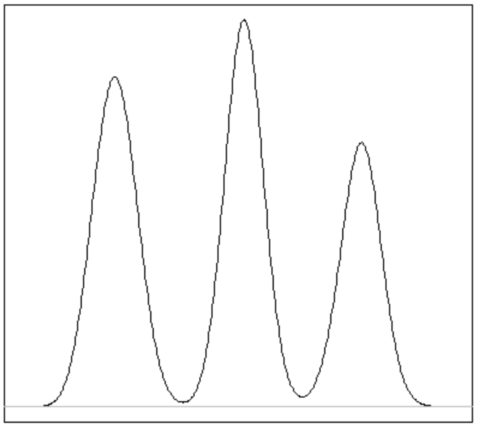
(c)



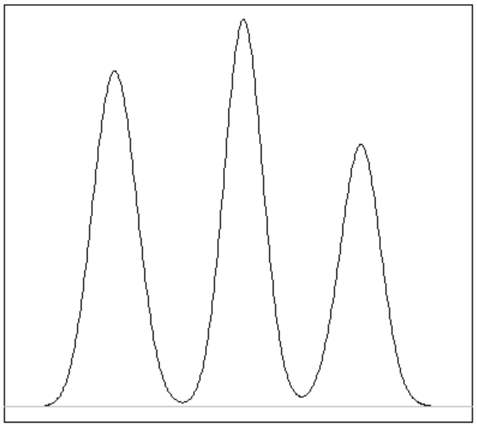
(d)



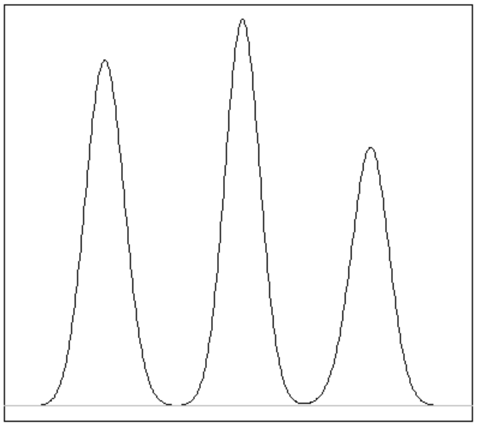
(e)



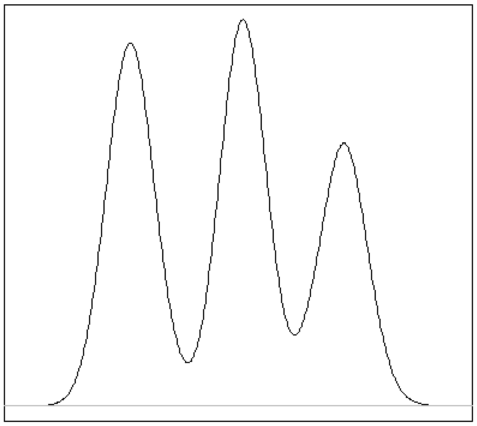
(f)



(g)

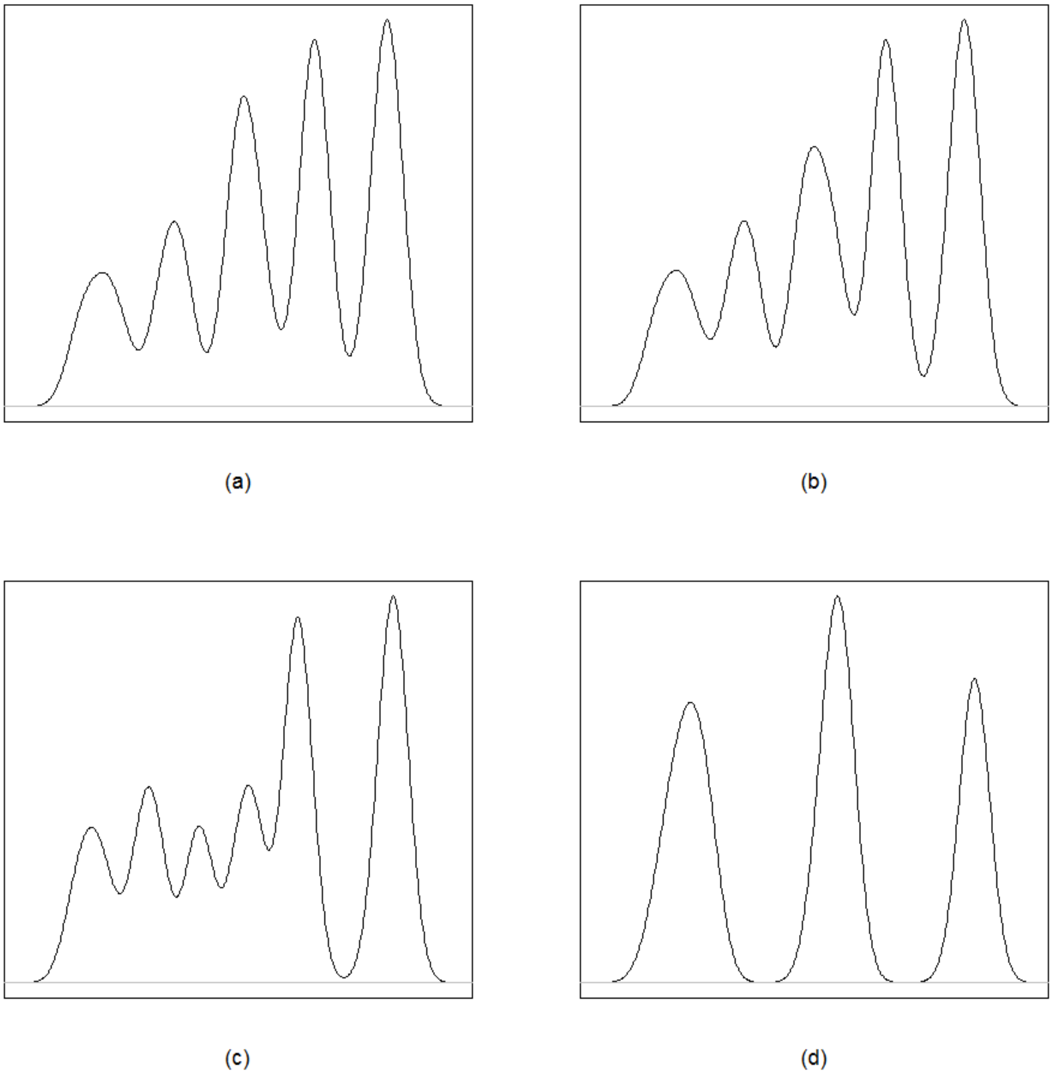


(h)



(i)

Figure 3.4 (Real Data Example): Extended Bootlier procedure for SMI data with observations #1 and #2 excluded from the regression model: (a) the density plot for ordinary residuals, sample Bootlier index = 0.43834, (b) the density plot for studentized residuals, sample Bootlier index = 0.42557, (c) the density plot for studentized deletion residuals, sample Bootlier index = 0.52484, (d) the density plot for SRCD, sample Bootlier index = 1.22618



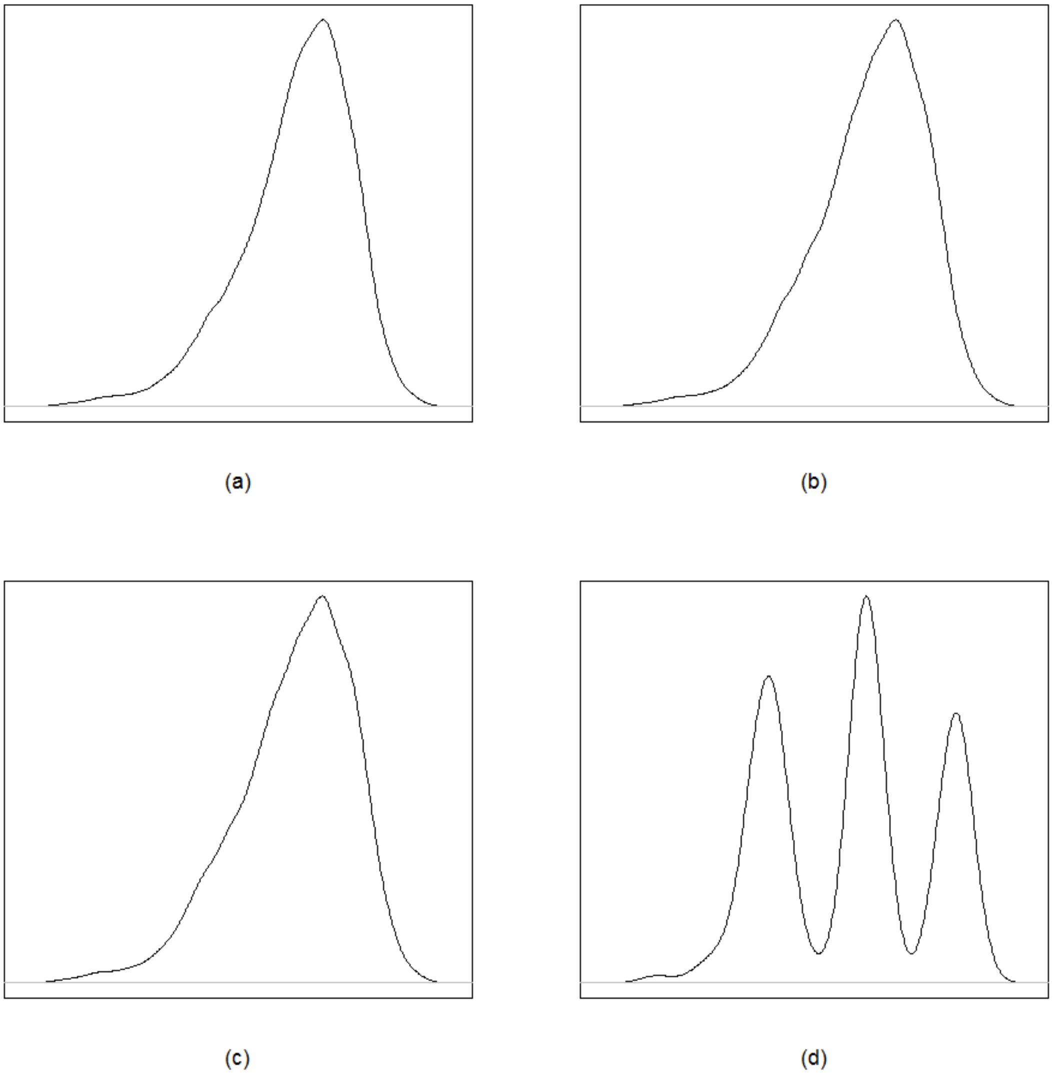
(a)

(b)

(c)

(d)

Figure 3.5 (Real Data Example): Extended Bootlier procedure for Sperm data with observations #1, #2, #3 and #15 excluded from the regression model: (a) the density plot for ordinary residual, sample Bootlier index = 0, (b) the density plot for studentized residuals, sample Bootlier index = 0, (c) the density plot for studentized deletion residuals, sample Bootlier index = 0, (d) the density plot for SRCD, sample Bootlier index = 0.53587



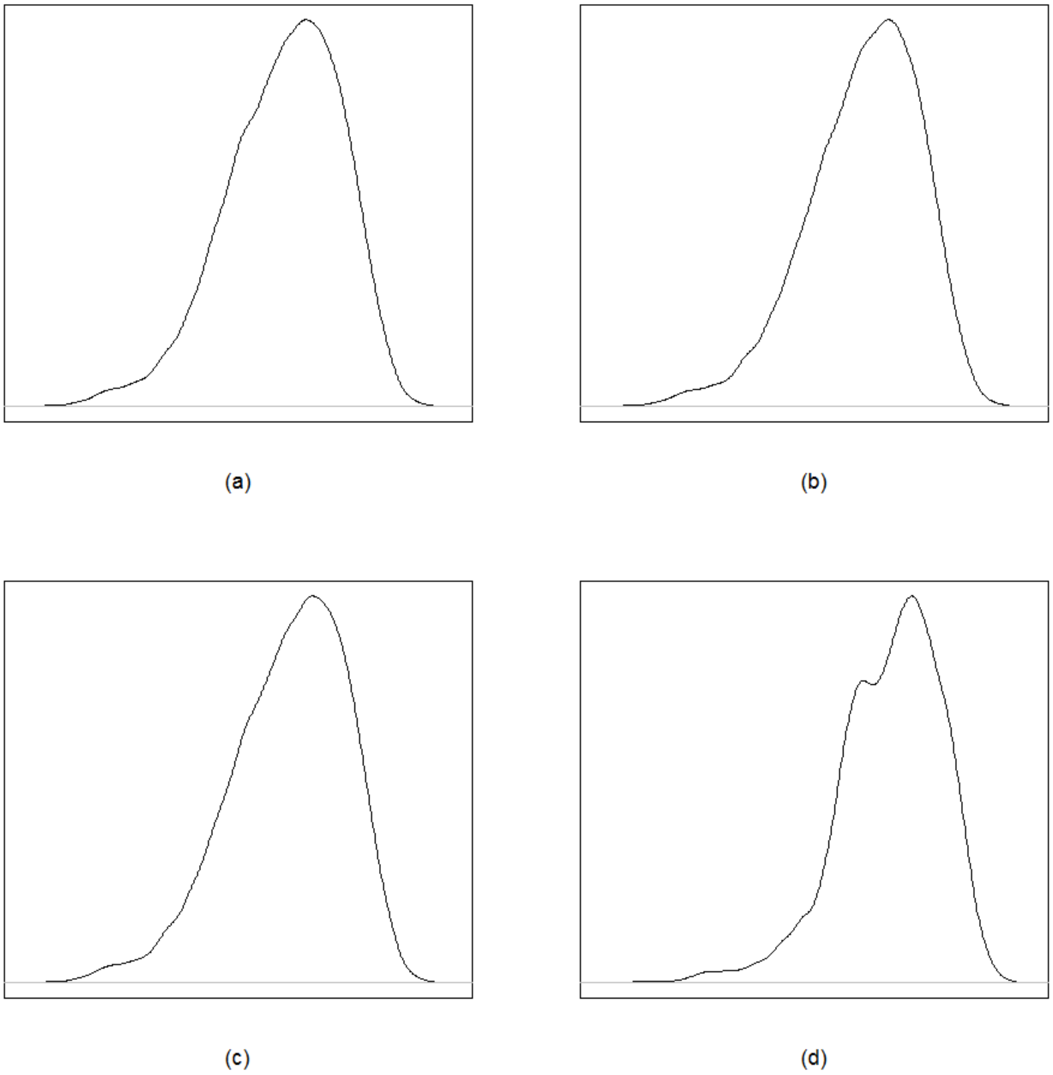
(a)

(b)

(c)

(d)

Figure 3.6 (Real Data Example): Extended Bootlier procedure for Sperm data with observations #1, #2, #3, #14 and #15 excluded from the regression model. (a) the density plot for ordinary residuals, sample Bootlier index = 0, (b) the density plot for studentized residuals., sample Bootlier index = 0, (c) the density plot for studentized deletion residuals, sample Bootlier index = 0, (d) the density plot for SRCD, sample Bootlier index = 0.00078



(a)

(b)

(c)

(d)

REFERENCES

Barnett, V. and Levis, T. (1994), Outliers in Statistical Data (3rd edition), Wiley

Beckman, R.J. and Cook, R.D. (1983), Outlier………s (with discussion), *Technometrics*, Vol. 25, No.2, 119-163

Beckman, R.J., and Trussell, H.J. (1974), The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regression, *Journal of the American Statistical Association*, Vol. 69, No. 345, 199-201

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, Inc.

Bollen, K.A., and Jackman, R.W. (1990), Regression Diagnostics: An Exploratory Treatment of Outliers and Influential Case, Modern Methods of Data Analysis (Fox, J., and Long, J.S.), Sage

Chatterjee, S. and Hadi, A. (1986), Influential Observations, High Leverage Points, and Outliers in Linear Regression, *Statistical Science,* Vol. 1, No. 3, 379-393

Clarke, B. R. (2008), Linear Models: The Theory and Application of Analysis of Variance, John Wiley & Sons, Inc.

Cook, R.D. (1979), Influential Observations in Linear Regression, *Journal of the American Statistical Association*, Vol. 74, No. 365, 169- 174

Cook, R.D. and Weisberg, S. (1982), Residuals and Influence in Regression, Chapman and Hall

Durett, R. (2010), Probability: Theory and Examples, Cambridge University Press

Hoaglin, D., and Welsch, R. (1978), The Hat Matrix in Regression and ANOVA, *The American Statistician*, Vol. 32, No. 1, 17-22

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2003), Introduction to Linear Regression Analysis, 3rd Edition, John Wiley & Sons, Inc.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996), Applied Linear Regression Modes, 3rd Edition, IRWIN

Rosner. B. (1983), Percentage Points for a Generalized ESD Many-Outlier Procedure, *Technometrics*, Vol. 25, No.2, 165-172

Singh, K. and Xie, M. (2003), Bootlier-Plot – Bootstrap Based Outlier Detection Plot, *Sankhya: The Indian Journal of Statistics*, Vol. 65, Part 3, 532-559

CHATPER 4

SOFTWARE DEVELOPMENT

4.1 INTRODUCTION

Software development is presented in this chapter. The software includes a numerical algorithm for calculation of the Bootlier index and R/C functions for the extended Bootlier procedure developed in Chapter 2 and Chapter 3.

4.2 COMPUTATION ALGORITHMS

The Bootlier index is expressed in a closed form in (2.4) of Chapter 2, and the concept is easy to understand, but it is not trivial to use (2.4) directly to compute Bootlier index because it involves integral from $-\infty$ to $-\infty$. The analytical solution seems not feasible. However, a numerical solution is proposed noticing that the density plot is expressed as the kernel density function.

Let $f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$ be the kernel density function for a given sample $\{X_1, X_2, \cdots, X_n\}$. An interesting finding is that $f(x)$ monotonically increases on $\{x; x < \min_i\{X_i\}\}$ and monotonically decreases on $\{x; x > \max_i\{X_i\}\}$. One immediate result is that the valley area as illustrated in Figure 2.1 only exists in the interval $[\min_i\{X_i\}, \max_i\{X_i\}]$, which leads to a numerical algorithm to calculate Bootlier index as follows.

1. Let $\Delta = \frac{\max_i\{X_i\} - \min_i\{X_i\}}{K}$ and divide the interval $[\min_i\{X_i\}, \max_i\{X_i\}]$ equally into K partitions. Denote $t_0 = \min_i\{X_i\}$, $t_K = \max_i\{X_i\}$, and $t_k = t_0 + k\Delta$, $k = 1, .., K$.

2. Find the global maximum $\{f(t_0), f(t_1), \cdots, f(t_K)\}$ and denote it by $f(t_M)$.

3. Calculate the valley area between t_0 and t_M iteratively by the steps below.

   (a) Let $S_{LEFT} = 0$, $t_{START} = t_0$ and $t_{NEXT} = t_1$.

   (b) If $f(t_{NEXT}) \geq f(t_{START})$, let $t_{START} = t_{NEXT}$ and $t_{NEXT} = t_{NEXT+1}$.

   If $f(t_{NEXT}) < f(t_{START})$, let $S_{LEFT} = S_{LEFT} + \Delta(f(t_{NEXT}) - f(t_{START}))$ and $t_{NEXT} = t_{NEXT+1}$.

   (c) Repeat 3(a) and 3(b) until $t_{NEXT} = t_M$.

4. Calculate the valley area between t_M and t_K similar to Step 3 and denote it by S_{RIGHT}.

5. Bootlier index is $S = S_{LEFT} + S_{RIGHT}$

The procedure above is similar to the standard numerical methods for calculating the integrals. Without repeating the general proof, we claim that S converges to Bootlier index for $f(x)$ as n goes to infinity.

4.3 R/C FUNCTIONS

4.3.1 Functions

Function for Bootlier index calculation

To improve computational efficiency, this function is first written in C language based on the algorithm described in Section 4.2, and compiled into dynamic link library (DLL). A R wrapper that calls the DLL by .Call function is developed to provide usual R interface.

```
################################################################################
#                                                                              #
# # C function - Calculate Bootlier index                                      #
#                                                                              #
################################################################################

#include <R.h>
#include <Rmath.h>
void getBIndexC(double *kfpoint, int *kflength, double *h, int *n, double
*BIndex)
{
  int i,j, pointmax;
  double kfmin, kfmax, x[*n+1], y[*n+1], ymax, movingmax;
  kfmin=kfpoint[0];
  kfmax=kfpoint[0];
  for(i=1; i<*kflength; i++) {
     if (kfmin>kfpoint[i]) kfmin=kfpoint[i];
     if (kfmax<kfpoint[i]) kfmax=kfpoint[i];
  }
  for (i=0; i<=*n; i++) {
    x[i]=kfmin+i*(kfmax-kfmin)/(*n);
    y[i]=0;
    for (j=0; j<*kflength; j++) y[i]=y[i]+dnorm((x[i]-kfpoint[j])/(*h),0,1,0);
    y[i]=y[i]/((*kflength)*(*h));
  }

  ymax=y[0];
  pointmax=0;
  for (i=0; i<=*n; i++) {
    if (ymax<y[i]) {
      ymax=y[i];
      pointmax=i;
    }
  }
  *BIndex=0;
  movingmax=y[*n];
  for (i=*n;i>=pointmax;i--) {
    if (y[i]<movingmax) *BIndex+=movingmax-y[i];
    else                 movingmax=y[i];
  }
  movingmax=y[0];
  for (i=0;i<=pointmax;i++) {
    if (y[i]<movingmax) *BIndex+=movingmax-y[i];
    else                 movingmax=y[i];
  }
  *BIndex=*BIndex*(kfmax-kfmin)/(*n);
}

################################################################################
#                                                                              #
# R wrapper function - Calculate Bootlier index                                #
#                                                                              #
# Function: getBIndexByC                                                       #
# Parameter Description                                                        #
#     kfpoint: sample for kernel density estimation                            #
#     h:       bandwidth of kernel density function                            #
#     n:       number of partitions                                            #
#                                                                              #
################################################################################

setwd("C:/…/Clib")
getBIndexByC<-function(kfpoint,h,n=2000){
 return(.C("getBIndexC",as.double(kfpoint),as.integer(length(kfpoint)),as.doubl
e(h),as.integer(n),as.double(vector("double",1)))[[5]])
}
```

Function for the extended Bootlier procedure for i.i.d. univariate sample

This R function calculates the sample Bootlier index, generates density plot, and produces *P*-value for a pre-specified distribution.

```
###############################################################################
#                                                                             #
# R function - Calculate Bootlier index, generate estimated density plot, and #
#              produce P-value based on pre-specified distribution            #
#                                                                             #
# Function: BProc                                                             #
# Parameter Description                                                       #
#     sample:    univariate sample                                            #
#     numTrim:   trimming number [positive integer] [default=2]               #
#     BPSide:    trimming side [1: two-sided, 2: upper-sided, 3: lower-sided] #
#                [default=2]                                                  #
#     numMTM:    number of bootstrap samples [positive integer] [default =    #
#                20,000]                                                      #
#     numPar:    number of partitions for Bootlier index calculation          #
#                [positive integer] [default=2,000]                           #
#     BPFrac:    fraction of total sample size for bootstrap samples          #
#                [positive number between 0 and 1] [default=1]                #
#     BPPlot:    plot estimated density function? [TRUE, FALSE] [default=TRUE]#
#     BPTest:    produce P-value? [TRUE, FALSE]  [default=TRUE]               #
#     BPDist:    pre-specified distribution [1: normal, 2: T6, 3: Exponential #
#                4: Uniform, 5: Cauchy] [default=1 normal]                    #
#     numDist:   number of observations for empirical distribution [positive  #
#                integer] [default=1,000]                                     #
#     BPSeed:    seed for random number generation [default=1]                #
#     StatusBar: display the progress bar? [TRUE, FALSE] [default=TRUE]       #
#                                                                             #
###############################################################################

BProc<-function(sample,numTrim=2,BPSide=2,numMTM=20000,numPar=2000,BPFrac=1,
BPPlot=TRUE,BPTest=TRUE,BPDist=1,numDist=1000,BPSeed=1,StatusBar=TRUE){
  # Initialization
  set.seed(BPSeed)
  numObs=length(sample)
  dyn.load("getBIndexC.dll")
  # Get the MTM sample

MTM=getMTM(sample=sample,numTrim=numTrim,BPSide=BPSide,numMTM=numMTM,BPFrac=BPF
rac)
  # Plot the kernel density estimator
  if (BPPlot==TRUE) plot(density(MTM),xlab="",ylab="",main="")
  # Calculate Bootlier index
  hSRT=0.9*min(IQR(MTM)/1.34,sd(MTM))*(numMTM^(-0.2))
  BIndex=getBIndexByC(MTM,hSRT,n=numPar)
  # Simulate distribution of Bootlier index and produce P-value
  if(BPTest==TRUE){
     # Simulate distribution
     set.seed(BPSeed)
     BIndexEmp=rep(0,numDist)
     if(StatusBar==TRUE) BPStatus=txtProgressBar(min=0,max=numDist,style=3)
```

```
      for (i in 1:numDist){
        if(BPDist==1) sampleEmp=rnorm(numObs)
        if(BPDist==2) sampleEmp=rt(numObs,6)
        if(BPDist==3) sampleEmp=rexp(numObs)
        if(BPDist==4) sampleEmp=runif(numObs)
        if(BPDist==5) sampleEmp=rcauchy(numObs)

MTMEmp=getMTM(sample=sampleEmp,numTrim=numTrim,BPSide=BPSide,numMTM=numMTM,BPFr
ac=BPFrac)
        hSRTEmp=0.9*min(IQR(MTMEmp)/1.34,sd(MTMEmp))*(numMTM^(-0.2))
        BIndexEmp[i]=getBIndexByC(MTMEmp,hSRTEmp,n=numPar)
        if(StatusBar==TRUE) setTxtProgressBar(BPStatus,i)
      }
      if(StatusBar==TRUE) close(BPStatus)

      # Calculate P-value
      BIndexEmpQ=quantile(BIndexEmp,probs=(seq(0,1,0.05)))
      BPpval=sum(BIndexEmp>=BIndex)/numDist
   }
   # Produce the final result
   if(BPTest==FALSE){
     BIndexEmp=NULL
     BIndexEmpQ=NULL
     BPpval=NULL
   }
   BPList=list(MTM,BIndex,BIndexEmp,BIndexEmpQ,BPpval)
   names(BPList)=c("MTM","BIndex","BIndexEmp","BIndexEmpQuantile","BPpvalue")
   return(BPList)
}

################################################################################
#                                                                              #
# R function - Generate MTM samples                                            #
#               This is a sub-function called by BProc to generate MTM samples.#
#                                                                              #
################################################################################

getMTM<-function(sample,numTrim,BPSide,numMTM,BPFrac){
   numSample=length(sample)
   n=floor(numSample*BPFrac)
   # Get the MTM sample
   MTM=rep(0,numMTM)
   for(i in 1:numMTM){
     select=ceiling(numSample*runif(n))
     y=sort(sample[select])
     if(BPSide==1) MTM[i]=(-2*numTrim)*sum(y)/(n*(n-2*numTrim))+sum(y[1:numTrim],y[(n-
numTrim+1):n])/(n-2*numTrim)
     if(BPSide==2) MTM[i]=(-numTrim)*sum(y)/(n*(n-numTrim))+sum(y[(n-numTrim+1):n])/(n-
numTrim)
     if(BPSide==3) MTM[i]=(-numTrim)*sum(y)/(n*(n-numTrim))+sum(y[1:numTrim])/(n-numTrim)
   }
   return(MTM)
}
```

Function to produce the null distribution of sample Bootlier index

This R function produces the null distribution of the sample Bootlier index. In addition to the main function BProc, this function provides flexibility for those who only need the null distribution of sample Bootlier index.

```
################################################################################
#                                                                              #
# R function - Produce the simulated distribution                              #
#                                                                              #
# Function: BProcEmp                                                           #
# Parameter Description                                                        #
#     numObs:    sample size of original sample [positive number]              #
#     numTrim:   trimming number [positive integer] [default=2]                #
#     BPSide:    trimming side [1: two-sided, 2: upper-sided, 3: lower-sided]  #
#                [default=2]                                                   #
#     numMTM:    number of bootstrap samples [positive integer]                #
#                [default = 20,000]                                            #
#     numPar:    number of partitions for Bootlier index calculation           #
#                [positive integer] [default=2,000]                            #
#     BPFrac:    fraction of total sample size for bootstrap samples           #
#                [positive number between 0 and 1] [default=1]                 #
#     BPDist:    pre-specified distribution [1: normal, 2: $T_6$, 3: Exponential #
#                4: Uniform, 5: Cauchy] [default=1 normal]                     #
#     numDist:   number of observations for empirical distribution [positive   #
#                integer] [default=1,000]                                      #
#     BPSeed:    seed for random number generation [default=1]                 #
#     StatusBar: display the progress bar? [TRUE, FALSE] [default=TRUE]        #
#                                                                              #
################################################################################

BProcEmp<-function(numObs,numTrim=2,BPSide=2,numMTM=20000,numPar=2000,BPFrac=1,
BPDist=1,numDist=1000,BPSeed=1,StatusBar=TRUE){
  # Initialization
  dyn.load("getBIndexC.dll")
  set.seed(BPSeed)
  # Generate the simulated distribution and P-value
  BIndexEmp=rep(0,numDist)
  if(StatusBar==TRUE) BPStatus=txtProgressBar(min=0,max=numDist,style=3)
  for (i in 1:numDist){
     if(BPDist==1) sampleEmp=rnorm(numObs)
     if(BPDist==2) sampleEmp=rt(numObs,6)
     if(BPDist==3) sampleEmp=rexp(numObs)
     if(BPDist==4) sampleEmp=runif(numObs)
     if(BPDist==5) sampleEmp=rcauchy(numObs)

MTMEmp=getMTM(sample=sampleEmp,numTrim=numTrim,BPSide=BPSide,numMTM=numMTM,BPFr
ac=BPFrac)
     hSRTEmp=0.9*min(IQR(MTMEmp)/1.34,sd(MTMEmp))*(numMTM^(-0.2))
     BIndexEmp[i]=getBIndexByC(MTMEmp,hSRTEmp,n=numPar)
     if(StatusBar==TRUE) setTxtProgressBar(BPStatus,i)
  }
  if(StatusBar==TRUE) close(BPStatus)
  BIndexEmpQ=quantile(BIndexEmp,probs=(seq(0,1,0.05)))
  # Produce the final result
  BPList=list(BIndexEmp,BIndexEmpQ)
  names(BPList)=c("BIndexEmp","BIndexEmpQuantile")
  return(BPList)
}
```

Function for the extended Bootlier procedure for linear regression model

This R function provides the sample Bootlier index, density plot, and *P*-value for the residuals and the square-root of Cook's distance for linear regression analysis.

```
###############################################################################
#                                                                             #
# R functions - Calculate Bootlier index, generate estimated density plot, and#
#              produce P-value for linear regression analysis                 #
#                                                                             #
# Function: BProcLinMod                                                       #
# Parameter Description                                                       #
#     linFit:    lm object from linear regression model                       #
#     resType:   residual type [1: ordinary residual, 2: studentized residual,#
#                3: studentized deletion residual, 4: SRCD]                   #
#     numTrim:   trimming number [positive integer] [default=2]               #
#     BPSide:    trimming side [1: two-sided, 2: upper-sided, 3: lower-sided] #
#                [default=2]                                                  #
#     numMTM:    number of bootstrap samples [positive integer] [default =    #
#                20,000]                                                      #
#     numPar:    number of partitions for Bootlier index calculation          #
#                [positive integer] [default=2,000]                           #
#     BPFrac:    fraction of total sample size for bootstrap samples          #
#                [positive number between 0 and 1] [default=1]                #
#     BPPlot:    plot estimated density function? [TRUE, FALSE] [default=TRUE]#
#     BPTest:    produce P-value? [TRUE, FALSE]  [default=TRUE]               #
#     numDist:   number of observations for empirical distribution [positive  #
#                integer] [default=1,000]                                     #
#     BPSeed:    seed for random number generation [default=1]                #
#     StatusBar: display the progress bar? [TRUE, FALSE] [default=TRUE]       #
#                                                                             #
###############################################################################

BProcLinMod<-
function(linFit,resType,numTrim=2,BPSide=2,numMTM=20000,numPar=2000,
BPFrac=1,BPPlot=TRUE,BPTest=TRUE,numDist=1000,BPSeed=1,StatusBar=TRUE){
  # Initialization
  set.seed(BPSeed)
  dyn.load("getBIndexC.dll")
  y=linFit$model[,1]
  numObs=length(y)
  x=cbind(rep(1,numObs),linFit$model[,-1])
  numDim=length(x[1,])
  variance=summary(linFit)$sigma
  hatMatrix=x%*%solve(t(x)%*%x)%*%t(x)
  hat=diag(hatMatrix) # hat=lm.influence(linFit)$hat
  residual=resid(linFit)
  if(resType==1) resValue=residual
  if(resType==2) resValue=rstandard(linFit)
  if(resType==3) resValue=rstudent(linFit)
  if(resType==4) resValue=sqrt(cooks.distance(linFit))*sign(residual)
  # Get the MTM sample

MTM=getMTM(sample=resValue,numTrim=numTrim,BPSide=BPSide,numMTM=numMTM,BPFrac=B
PFrac)
  # Plot the kernel density estimator
  if (BPPlot==TRUE) plot(density(MTM),xlab="",ylab="",main="")
  # Calculate Bootlier index
  hSRT=0.9*min(IQR(MTM)/1.34,sd(MTM))*(numMTM^(-0.2))
  BIndex=getBIndexByC(MTM,hSRT,n=numPar)
  # Generate the simulated distribution and P-value
  if(BPTest==TRUE){
     # Simulated distribution
     BIndexEmp=rep(0,numDist)
     if(StatusBar==TRUE) BPStatus=txtProgressBar(min=0,max=numDist,style=3)
     for (i in 1:numDist){
       sampleEmp=variance*(diag(numObs)-hatMatrix)%*%rnorm(numObs)
       varEst=sqrt(sum(sampleEmp**2)/(numObs-numDim))
       if(resType==2) sampleEmp=sampleEmp/sqrt(1-hat)/varEst
       if(resType==3) sampleEmp=sampleEmp*sqrt((numObs-numDim-1)/(1-
hat)/((numObs-numDim)*varEst**2-sampleEmp**2/(1-hat)))
       if(resType==4) sampleEmp=sampleEmp*sqrt(hat/numDim)/varEst/(1-hat)

MTMEmp=getMTM(sample=sampleEmp,numTrim=numTrim,BPSide=BPSide,numMTM=numMTM,BPFr
ac=BPFrac)
       hSRTEmp=0.9*min(IQR(MTMEmp)/1.34,sd(MTMEmp))*(numMTM^(-0.2))
       BIndexEmp[i]=getBIndexByC(MTMEmp,hSRTEmp,n=numPar)
       if(StatusBar==TRUE) setTxtProgressBar(BPStatus,i)
     }
     if(StatusBar==TRUE) close(BPStatus)
     # Cacluate P-value
     BIndexEmpQ=quantile(BIndexEmp,probs=(seq(0,1,0.05)))
     BPpval=sum(BIndexEmp>=BIndex)/numDist
  }
  # Produce the final result
   if(BPTest==FALSE){
   BIndexEmp=NULL
   BIndexEmpQ=NULL
   BPpval=NULL
  }
  BPList=list(MTM,BIndex,BIndexEmp,BIndexEmpQ,BPpval)
  names(BPList)=c("MTM","BIndex","BIndexEmp","BIndexEmpQuantile","BPpvalue")
  return(BPList)
}
```

4.3.2 Examples

Example 1 Codes for Real Data Example #1 in Chapter 2

```
  # Codes for real data example #1 in Chapter 2: 24 recorded temperatures at
which the primary O-ring of the space shuttle Challenger was sealed
  sample=c(66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70,
81, 76, 79, 75, 76, 58,)
  BProc(sample=sample,BPSide=3)
```

Example 2 Codes for Simulation Study #1 in Chapter 3

```
  # Codes for Simulation Study #1 in Chapter 3
  set.seed(1)
  x=10*(runif(30)-0.5)
  y=x+1+rnorm(30)
  x=c(x,0)
  y=c(y,4.72)
  linFit=lm(y~x)
  BProcLinMod(linFit=linFit,resType=1,BPSide=2)
  BProcLinMod(linFit=linFit,resType=2,BPSide=2)
  BProcLinMod(linFit=linFit,resType=3,BPSide=2)
  BProcLinMod(linFit=linFit,resType=4,BPSide=2)
```