

Investigation into linear regression influence measure diagnostics and bootstrap-based inferential techniques

C Booysen

 [orcid.org 0000-0002-2620-2327](https://orcid.org/0000-0002-2620-2327)

Dissertation accepted in partial of the requirements for the
degree *Master of Science in Mathematical Statistics* at the
North-West University

Supervisor: Prof L Santana

Graduation May 2020

25108557

Acknowledgements

“The Lord is my strength and my song; he has given me victory. This is my God, and I will praise him- my father’s God, and I will exalt him!”

Exodus 15:2

I give all the glory and honour to our Heavenly Father for giving me the great opportunity to complete this dissertation, and providing me with the necessary insight, strength and wisdom throughout this process. Without His help, none of this would have been possible. Additionally, I would like express my appreciation to the following people for their contribution in this dissertation:

- To my supervisor, Prof. Leonard Santana, for all your guidance, patience, enthusiasm and support, as well as your assistance in $\text{\LaTeX}$ and the R code, your insights in the statistical theory used in this dissertation, and lastly, your help in proof-reading the drafts of this dissertation in order to perfect it. It was a great honour for me to have you as my supervisor and I have learned so much from you. Sharing your extraordinary knowledge, is highly appreciated and I will always be thankful for that.
- To Prof. James Allison for your support and willingness to help with all the small (and large) details of the project!
- To Dr. Gerrit Grobler, for your mathematical insights in some of the proofs provided in this dissertation, and also for your willingness to help.
- To my parents, Pieter and Antoinette Booysen, for all your support, love, interest, and financial support. Without your sacrifices and encouragement throughout my life, I would not be where I am today.
- To my brother and sister, Pieter and Joey Booysen, as well as my grandpa and grandma, Willie and Clarice Naudé, for your unwavering interest, love, support, and motivation throughout my studies, and just for always being there for me.
- To my close friend and fellow masters student, Enrike le Roux, for your encouragement when it was the most needed, helping me finish and perfecting my tables among others, and for our helpful discussions on this dissertation.
- To the National Research Foundation (NRF) for providing me with the financial support in order for me to do my Master’s degree (NRF Grant Number 114631).

Abstract

Since extreme or influential observations drastically affect the fit of regression models, their detection plays a big role in regression model fitting. Many traditional diagnostic techniques employ single-case deletion methods for this purpose, but these have several drawbacks, such as an inability to detect masked or swamped influential cases. Recently, however, new simulation-based techniques have been developed to overcome these problems, such as the technique called ADAP proposed by Roberts et al. (2015). However, this method lacks a formal or data-driven choice for the cut-off values used in the procedure. Another recent technique, attributed to Martin and Roberts (2010), attempts to improve on the traditional single-case deletion methods by using a Jackknife-after-bootstrap approach to find data-driven cut-off values for traditional diagnostic statistics. In this dissertation, we combine these two approaches, and use the Jackknife-after-bootstrap method to find data-driven cut-off values for the ADAP method, thereby potentially improving the latter method. Additionally, a completely new method is proposed for the detection of influential observations that is based on a simple approximation to the traditional Cook's distance diagnostic measure. In this way, a new cut-off value for Cook's distance is obtained, which can be used as an alternative to the traditional rules of thumb cut-off values.

An intensive simulation study is presented that compares the newly proposed 'combined' approach, as well as the new method based on the approximation to Cook's distance, to the traditional Cook's distance diagnostic measure, the plain implementation of the Jackknife-after-bootstrap method, the plain implementation of the ADAP method, and a commonly used modern method called MC^3 . Disappointingly, our results show that while the newly proposed combined method does fare better than the traditional methods, the MC^3 method, and the plain implementation of the Jackknife-after-bootstrap method, it is an extremely time consuming process and it does not improve on either the plain implementation of the ADAP method, or the newly proposed Cook's distance based method. The new Cook's distance based method, on the other hand, performs better than expected. This method results in very good performances, outperforming the traditional Cook's distance method of obtaining cut-off values, the MC^3 method, the plain implementation of the Jackknife-after-bootstrap method, the newly proposed 'combined' method, and in some cases it performs as well as the plain implementation of the ADAP method, although the naïve application of the ADAP method is still overall the best of the discussed methods.

Keywords: Adaptive automatic multiple-case deletion (ADAP) technique; bootstrap; Cook's distance; cut-off values; influential measures; influential observations, jackknife-after-bootstrap; linear regression; masking; swamping.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Contents

1	Introduction	1
1.1	Overview	1
1.2	Objectives	2
1.3	Contributions	3
1.4	Chapter breakdown	3
2	Linear regression models	4
2.1	Introduction	4
2.2	The linear regression model	5
2.2.1	Interpretation of the regression parameters	6
2.2.2	Estimation of the regression function	6
2.3	Traditional measures of influential and outlier detection in linear regression	8
2.3.1	Leverages	8
2.3.2	<i>DFBETAS</i>	10
2.3.3	<i>DFFITs</i>	11
2.3.4	<i>CovRatio</i>	11
2.3.5	Cook's distance	14
2.3.6	General methods for finding cut-off values for measures of influence .	16
3	Resampling methods for linear regression	18
3.1	The bootstrap	18
3.1.1	Bootstrap notation	19
3.1.2	The empirical distribution function (EDF)	19
3.1.3	The plug-in principle	20
3.1.4	The basic implementation of the bootstrap	21
3.2	The bootstrap in linear regression	22
3.2.1	Residual-based resampling	22
3.2.2	Cases-based resampling	24
3.3	Jackknife	25
3.4	Jackknife-after-bootstrap (JaB)	26
4	Modern influential diagnostic methods	30
4.1	Jackknife-after-bootstrap regression influence diagnostics	31
4.2	An adaptive, automatic multiple-case deletion technique for detecting in- fluence in regression	34

4.2.1	Initial check stage of ADAP	35
4.2.2	Confirmation stage of ADAP	36
4.2.3	Algorithm of the ADAP method	38
4.3	MC^3 Method	40
5	Newly proposed influential diagnostic methods	42
5.1	Using jackknife-after-bootstrap cut-off values in the ADAP method (the “Combined” method)	42
5.2	A new method based on a simple approximation of the distribution of Cook’s distance (the “New” method)	46
6	Simulation study	50
6.1	Introduction	50
6.2	Simulation settings	53
6.3	Results	55
6.4	Discussion	71
6.4.1	Results when influential values are inserted using the <i>simple</i> method	71
6.4.2	Results when influential values are inserted using the <i>masked</i> method	75
6.4.3	Comparison between the results of the two different ways in which the influential observations were added to the data	78
6.4.4	Cut-off values for each method	78
6.4.5	Execution time for each method	79
7	Conclusion	80
	Bibliography	82
A	Derivation of the least squares estimates	88
B	Proofs of the properties of leverages	89
B.1	Proof of the bounds of leverages	89
B.2	Proof that the sum of the leverages add up to p	90
C	Proof of the variance of the regression residuals	91
D	Proof of the alternative form of $CovRatio$	92
E	Proof of the simple form of Cook’s distance	94
F	Proofs on restrictions for the New method	96

List of Figures

3.1	Efron's (2003) schematic representation of the plug-in principle.	20
4.1	An example illustrating why the ADAP algorithm needs the 'confirmation stage'.	37
5.1	Plot of h_{ii} against the function $h_{ii}/(1 - h_{ii})$ to show that the function $h_{ii}/(1 - h_{ii})$ is monotone everywhere except at $h_{ii} = 1$	47
6.1	An illustration of the error distributions given in Table 6.3.	55
F.1	Plot of n_1 against different values of D^2 to find the bounds of n_1	99
F.2	Plot of n_2 against different values of D^2 to find the bounds of n_2	99

List of Tables

6.1	General form of a confusion matrix.	53
6.2	Density functions used for error distributions.	54
6.3	Error distributions used in the simulation study.	54
6.4	Average TP, FP, SN and SP for the various methods across 100 data sets generated with an $N(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	57
6.5	Average TP, FP, SN and SP for the various methods across 100 data sets generated with an t_3 error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	58
6.6	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Log-normal $(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	59
6.7	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Laplace $(0, 3)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	60
6.8	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(-1.3, 1, 1.3, 1, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	61
6.9	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(0, \frac{1}{\sqrt{13}}, 0, \frac{5}{\sqrt{13}}, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.	62
6.10	Average TP, FP, SN and SP for the various methods across 100 data sets generated with an $N(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	63
6.11	Average TP, FP, SN and SP for the various methods across 100 data sets generated with an t_3 error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	64

6.12	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Log-normal $(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	65
6.13	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Laplace $(0, 3)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	66
6.14	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(-1.3, 1, 1.3, 1, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	67
6.15	Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(0, \frac{1}{\sqrt{13}}, 0, \frac{5}{\sqrt{13}}, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.	68
6.16	Cut-off values used in the New and Traditional methods.	69
6.17	Average of the JaB method cut-off values for the various methods across 100 generated data sets (estimated variances are stated in parentheses below each entry).	69
6.18	Average of the Combined method cut-off values for the various methods across 100 generated data sets. The initial check stage cut-off values appear in the row labelled “Init.” and the confirmation stage cut-off values appear in the row labelled “Conf.” (estimated variances are stated in parentheses below each entry).	70
6.19	Median run-time (in seconds) over the different error distribution data sets for each influential diagnostic method.	71

Table of abbreviations

A list of abbreviations that will be used throughout this dissertation, is given below.

JaB	Jackknife-after-bootstrap
ADAP	Adaptive automatic multiple-case deletion
EDF	Empirical distribution function
TP	True positives
FP	False positives
TN	True negatives
FN	False negatives
SN	Sensitivity
SP	Specificity

Chapter 1

Introduction

1.1 Overview

The detection of influential or extreme observations in a data set is a very important aspect in regression analysis, since these cases can exert great influence on the conclusions made from fitted regression models. According to Cook (1979), influential observations can be defined as those observations in a data set that have the effect of, when removed from the data set, drastically change essential features of the regression analysis. Many traditional methods that are implemented for the detection of influential observations, include statistics that are based on single-case deletion techniques.

Even though single-case deletion techniques are very popular, they have some disadvantages. One of these disadvantages is that they might not perform well in cases when the data set consists of multiple influential observations, since then problems like *masking* and *swamping* might occur. According to Atkinson (1986), masking is defined as the situation where an observation is found to be influential only after a set of influential observations were deleted from the data set (i.e., the set of influential observations ‘mask’ other influential cases). Swamping, on the other hand, is when the opposite occurs, that is, when a normal observation are incorrectly flagged as influential as a result of the strong effect of other influential observations on the fitted model (i.e., influential observations have the effect of ‘swamping’ observations, causing them to be seen as influential even though they are normal) (Nurunnabi et al., 2014; Rahmatullah Imon, 2005).

One solution to this disadvantage of single-case deletion techniques are multiple-case deletion techniques. Multiple-case deletion was first proposed by Belsley et al. (1980) and, in later years, many other multiple-case deletion techniques were developed, see Rousseeuw and Van Zomeren (1990), Hadi and Simonoff (1993), Peña and Yohai (1999), Pan et al. (2000), Peña (2005), Rahmatullah Imon (2005), and Nurunnabi et al. (2014). As in the case of single-case deletion techniques, multiple-case deletion techniques also have a disadvantage, namely, they are very time consuming to execute. The reason for this is because these procedures require that one constructs subsets of size j from a sample of size n , and the number of combinations of doing this can be astronomically large for large sample sizes. For this reason, and also for the fact that it is very rarely known how many obser-

vations are necessary to be removed from a data set in order for effects like masking to be eliminated, research in finding relatively fast methods that eliminate effects like masking, have become very popular.

In this dissertation, we are specifically interested in methods used for the detection of influential observations in cases where effects like masking and swamping are present in the data. Many different influence diagnostic methods, including the traditional methods for influential value detection, relatively new methods based on simulations, new modifications of simulation-based methods, and a newly proposed method based on a simple idea relating to the approximate distribution of a diagnostic statistic, will be discussed and investigated by comparing their performances in detecting influential observations in simulated data sets.

1.2 Objectives

The main objectives of this dissertation are given as follows:

- to investigate the performance in detecting influential observations of the traditional Cook's distance influential diagnostic measure;
- to investigate the performance in detecting influential observations of the jackknife-after-bootstrap (JaB) method proposed by Martin and Roberts (2010);
- to investigate the performance in detecting influential observations of the adaptive, automatic multiple-case deletion technique (ADAP) proposed by Roberts et al. (2015);
- to provide a new method based on the combination of the JaB and the ADAP methods (Combined method);
- to investigate the performance in detecting influential observations of the Combined method;
- to provide a new method based on a simple approximation of the distribution of Cook's distance (New method);
- to investigate the performance in detecting influential observations of the New method;
- to investigate the performance in detecting influential observations of the MC^3 method proposed by Hoeting et al. (1996);
- to compare the performance of all these methods to one another and make recommendations.

In order to achieve the stated objectives for each of the different methods described above, two types of data sets will be simulated. The first data set type is where influential observations are included in the data sets in a simple, random way, and the second type is where the influential observations are included in the data sets such that masking occurs. Furthermore, both types of data sets were generated from linear regression models with

differing sample sizes (sample sizes include 100, 200, and 500), differing number of predictor variables included in the regression model (number of predictor variables include 5 and 15), differing number of influential observations included in the data (number of influential observations include 5 and 10), as well as by using one of five different choices for the error distribution, namely, the normal, t , log-normal, Laplace, and contaminated normal distributions.

1.3 Contributions

The contributions of this dissertation, given the objectives, are as follows:

- a modification of the ADAP method whereby the JaB method is used to determine cut-off values (Combined method); this modification has been suggested, but not yet implemented or investigated in the literature (see Roberts et al., 2015);
- the creation of a new method based on a simple approximation of the distribution of Cook's distance (New method);
- a comprehensive simulation study on the performance of these two new methods, when compared to existing traditional and modern methods for the detection of influential observations.

1.4 Chapter breakdown

The remainder of this dissertation will be arranged as follows. Chapter 2 will describe the background regarding linear regression models as well as the traditional measures of influence and outlier detection used in these models. The next chapter, Chapter 3, presents basic theory related to the use of the bootstrap and jackknife in linear regression, giving specific attention to the 'jackknife-after-bootstrap' methodology, which is used in later chapters. This is then followed by two chapters providing methods for the detection of influential observations, where the first chapter (Chapter 4) discusses modern simulation-based methods of influential diagnostics, and the second chapter (Chapter 5) provides the two newly proposed methods for detecting influential values. The dissertation concludes with a extensive simulation study presented in Chapter 6 as well as a discussion on the findings of this simulation study. In Chapter 7 we provide our conclusions drawn from the simulation study as well as some recommendations and possible future research. Appendices are also provided to supplement the theory in various chapters.

Chapter 2

Linear regression models

2.1 Introduction

The detection of influential observations in data is a very important aspect in regression analysis, since an influential observation can have a big effect on the results produced from your study. When a model is built on a data set in order to make predictions, for example, one influential observation is all that is necessary to drastically change the results of your predictions and therefore give faulty results. In this dissertation, we are interested in methods for finding these influential observations which are all based on fitting the multiple linear regression model and using its properties. In this chapter, we will therefore provide details about the classic linear regression model and all necessary theory and results required for later chapters, including some notes on the distribution of the model, the way the regression coefficients of the model are interpreted and a method for estimating the regression coefficients (see Section 2.2). Furthermore, we will also provide details about some traditional methods used for detecting influential observations in this chapter (see Section 2.3).

According to Kutner et al. (2005), regression analysis can be defined as the process of estimating the relationship between different variables included in a model, using a statistical method. In the end, the aim is to make predictions on the response variable Y (the variable that depends on the rest of the variables in the model, also called the *dependent variable*) using all, or some of the predictor variables X (the variables that are independent of the response variable, also called the *independent variables*).

There can be two types of relationships between the response variable and the predictor variables namely, a *functional* relationship or a *statistical* relationship. These two relationships are given as

$$Y = f(X) \quad \text{and} \quad Y = f(X) + \varepsilon,$$

respectively, where ε is some random ‘error’ term. The difference between these two types of relationships is that the functional relationship represents the exact relationship between X and Y , whereas the statistical relationship represents a stochastic relationship between X and Y . This means that when you have the functional relationship between

X and Y , you would be able to find the exact value of Y for a given value of X . The statistical relationship, on the other hand, would, for a given value of X , result in a corresponding value of Y that is some random distance (the error term) from the value expected when using the functional relationship. The error term in the statistical relationship can therefore be seen as the amount by which it deviates from the functional relationship. The reason for this deviation is because of the fact that, in many situations, predictor variables are excluded from the model that should have been included since they also affect the response variable. The error term is therefore the element in the model that combines all information about the predictor variables not included in the model that may actually have an effect on the response variable.

The linear regression model that will be used in the chapters that follows is based on the statistical relationship stated above and will be discussed in the next section.

2.2 The linear regression model

We will distinguish between two linear regression models namely, the simple linear regression model and the multiple linear regression model. The use of the word *simple*, in simple linear regression, is to indicate that only one predictor variable is included in the model. Furthermore, the word *linear* is used to indicate that the parameters and the predictor variables are in a linear form in the model. This indicates that the mean of the response variable is equal to some linear combination of the predictor variables and their corresponding regression coefficients. The multiple linear regression model, on the other hand, is an extension on the simple linear regression model. Instead of only one predictor variable which was included in the simple linear regression model, the multiple linear regression model consists of multiple predictor variables. The reason why this model is necessary and why multiple predictor variables are included in this model is because, in this case, a better understanding of the variation in the response variable is obtained. This is because of the fact that, when including more than one predictor variable in the model, the information of all the extra predictor variables are taken into account.

Theory and results of the simple linear regression model will be omitted here, since they can easily be obtained from the multiple linear regression model.

The multiple regression model is given by

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \\ &= \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \dots, n, \end{aligned}$$

and in matrix notation it is given by

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times p)}{\mathbf{X}} \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}} \quad (2.1)$$

or

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & \dots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where p is the number of regression parameters in the model, n is the total number of observations in the data set, Y_i is the value of the random response variable for the i^{th} case, $\beta_0, \beta_1, \dots, \beta_{p-1}$ are the unknown regression coefficient parameters, $X_{i1}, X_{i2}, \dots, X_{i,p-1}$ are the elements of the i^{th} row of the design matrix $\mathbf{X}$, which are fixed, known, constants, and ε_i is the random error term for the i^{th} case, with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$, for all $i = 1, 2, \dots, n$. Also, ε_i and ε_k are uncorrelated such that $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0$ for all $k \neq i, k, i = 1, 2, \dots, n$. Usually, an assumption of normality on $\boldsymbol{\varepsilon}$ is made as well, but in our case it is not necessary since we will only look at estimation.

Notes on the multiple linear regression model: From the model it is clear that Y_i consists of the sum of a constant term ($\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1}$) and a random term (ε_i), which will therefore also make Y_i a random variable. Furthermore, as mentioned above, $E(\varepsilon_i) = 0$, which leads to

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1},$$

and since $\text{Var}(\varepsilon_i) = \sigma^2$, we have that

$$\text{Var}(Y_i) = \sigma^2.$$

Lastly, the response terms Y_i and Y_k , $k \neq i, k, i = 1, 2, \dots, n$, are assumed to be uncorrelated, since the error terms ε_i and ε_k , $k \neq i, k, i = 1, 2, \dots, n$, are assumed to be uncorrelated.

2.2.1 Interpretation of the regression parameters

Since the linear regression model is now defined, we will explain how the regression parameters are interpreted in this section. The β_0 coefficient can be seen as the intercept term of the y -axis. This means that if the $p - 1$ predictor variables are set equal to zero then β_0 would, for this setting of the predictor variables, be equal to the mean of the probability distribution of Y_i . The coefficients of the predictor variables $\beta_1, \dots, \beta_{p-1}$, on the other hand, can be seen as the expected amount by which the response variable would change with a unit change in their corresponding predictor variables while the others are held constant. For example, β_j would represent the expected amount of change in the response variable Y_i with a unit increase in X_{ij} while the rest of the predictor variables are held constant.

2.2.2 Estimation of the regression function

Since the regression function is very seldom known, we want to be able to estimate it. According to James et al. (2015), the aim is to get a linear model that is the best representation of the data points. When fitting this estimated regression model, some of the

data points will lie above the fitted line and others below the fitted line, but the distance from the line to the data points must be as small as possible. There are quite a few methods for calculating a fit like this, but the *method of least squares* is one of the most commonly used methods and also the method what will be used in the chapters that follow (other methods for estimating the regression coefficients can be found in Zellner (1962), Sen (1968), Hoerl and Kennard (1970), Rao and Toutenburg (1995), Tibshirani (1996), Greene (2003), Efron et al. (2004), and Kutner et al. (2005)). Furthermore, the method of least squares is not dependent on knowing the distribution of the error term. Before the method of least squares will be explained, we first introduce some notation.

Notation

Denote the observed predictor variable vectors by $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ (note that the $p \times 1$ vector $\mathbf{X}_i$, denotes the i^{th} row of $\mathbf{X}$), and the observed response variable values, by $Y_1, Y_2, \dots, Y_n$.

Method of least squares

The method of least squares is a method that estimates regression coefficients by making use of the sample data. This is done in such a way that the distance between the data points and the fitted regression line is minimized. The fitted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}, \quad (2.2)$$

where $\mathbf{b}$ denotes the vector of estimates for the p regression coefficients, and is given by

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}.$$

The distance that needs to be minimized for the least squares method is called the residuals and is given by

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}. \quad (2.3)$$

The method of least squares estimates the regression parameters by minimizing the sum of the squared residuals. This equation is denoted by Q and is given by

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \quad (2.4)$$

$(1 \times n)$
 $(n \times 1)$

Given the sample data $(\mathbf{X}_i, Y_i), i = 1, 2, \dots, n$, if $\boldsymbol{\beta}$ is replaced by $\mathbf{b}$ in (2.4), Q will be a minimum. The estimates of the regression coefficients, $\mathbf{b}$, can be obtained by solving the normal equations of the multiple linear regression model. These estimates are given by

$$\mathbf{b}_{(p \times 1)} = (\mathbf{X}^\top \mathbf{X})_{(p \times p)}^{-1} \mathbf{X}^\top_{(p \times n)} \mathbf{Y}_{(n \times 1)}. \quad (2.5)$$

The proof of how (2.5) was obtained, can be found in Appendix A.

When estimating the regression coefficients in (2.5), and the regression curve in (2.2), it is very important to determine if any observations in the data set are extreme or influential, since observations like these can have a great effect on the estimates, and therefore also on the results obtained from them. For this reason, we are interested in methods for detecting influential observations, and we start by investigating some traditional methods given in the next section. In Chapter 4 newer methods for influence detection will be given, followed by Chapter 5 in which an improved method, as well as a completely new method, will also be provided for this purpose.

2.3 Traditional measures of influential and outlier detection in linear regression

In regression analysis, the process of detecting influential observations in a data set is very important, as already mentioned. A basis of measures of influence was laid out by the papers and books by Cook (1977), Cook (1979), Belsley et al. (1980), and Cook and Weisberg (1982). The measures discussed in these papers and books are still very popular methods to use in finding influential observations. One disadvantage of these methods is that their cut-off values are based on large sample theory, since their distributions are very complex. In some cases then, these measures might lead to faulty results. In later chapters we introduce other methods that do not depend on large sample theory, but since these traditional methods are still very popular today, their performance will be used as a benchmark for comparison of the performance of the new methods. Therefore, in this section we will give attention to some of the traditional methods used for detecting influential observations.

Belsley et al. (1980) define an influential observation as an observation that, when compared to other observations in the data set, has a large influence on some estimates, like the estimated regression coefficients. These observations can be influential on their own or together with other observations in a group. An easy method to find influential observations is to delete observations one-by-one from the data set and then calculate what the effect of this deletion has on the fitted values, estimated coefficients, residuals, or the estimated covariance matrix of the coefficients. Those observations corresponding to a big change in these values, will be flagged as influential. The following measures of influence that employ this method of deletion include: Leverages, *DFBETAS*, *DFFITS*, *CovRatio*, and Cook's Distance.

2.3.1 Leverages

Leverages are very useful since they can be used to identify influential observations on their own, but also because they feature as an element in many other influential measures. According to Belsley et al. (1980), the leverage h_{ii} is defined as the i^{th} diagonal element of the hat matrix given by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top, \quad (2.6)$$

that is,

$$h_{ii} = \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i, \quad i = 1, 2, \dots, n, \quad (2.7)$$

where $\mathbf{X}_i$ denotes the i^{th} row of $\mathbf{X}$. These leverage values can be interpreted as a form of distance of $\mathbf{X}_i$ from the centroid or mean vector of the $\mathbf{X}$ values. In addition, leverages have two very important properties, given as follow:

$$0 \leq h_{ii} \leq 1, \quad i = 1, 2, \dots, n,$$

and

$$\sum_{i=1}^n h_{ii} = p. \quad (2.8)$$

The proofs of these two properties can be found in Appendix B.1 and Appendix B.2, respectively.

From (2.8), it is clear that the average of the leverages is given by p/n , i.e., the average distance from the centroid of $\mathbf{X}$. In the case where one has a data set with no influential observations included (i.e., all observations exert almost the same influence), it would mean that all the observations would have leverage values near the mean, p/n . In the cases where the h_{ii} values are not close to p/n , it is necessary to find cut-off values which will give an indication of how big an observation's h_{ii} value should be (i.e., how much bigger than the average value) for it to be considered influential.

To find a rough cut-off value for the leverages then, we make the assumption that the predictor variables are independent and that they are multivariate Gaussian distributed (which is seldom the case). These assumptions are used, because in cases like this some leverage functions' exact distributions can be calculated and their results can be used to find cut-off values for the leverages. Note however, that they are only guidelines for the cut-off values of the leverages, since the exact distributions are based on assumptions that very seldom occur.

According to Belsley et al. (1980), if we assume that the predictor variables are independent and multivariate Gaussian distributed, the following is true

$$\frac{(n-p)[h_{ii} - (1/n)]}{(1-h_{ii})(p-1)} \sim F_{(p-1), (n-p)}.$$

From this distribution of the function of h_{ii} , we can therefore calculate some results to find rough cut-off values for h_{ii} . If we calculate the 95% quantile of the F -distribution in the case where $p > 10$ and $n-p > 50$, for example, a value less than 2 was obtained. From this result it can be concluded that $\frac{2p}{n}$ would be a good cut-off value for the leverages. Again, note that this can only be seen as a rough cut-off value for the leverages and will not be accurate in all possible cases. One situation in which this cut-off value would not work at all, is in the case where $p/n > 0.4$. In this case all the observations in the data set will be flagged as influential, since then the degrees of freedom for every parameter becomes very small. Also, in cases where the number of predictor variables is small, this cut-off value often flags too many observations as influential. This cut-off value can therefore be used as a starting point for detecting influential observation, since it is easy to use and to remember.

The next two measures of influence are based on the size of change occurring in the estimated regression coefficients and the fitted values of the regression model, respectively, before and after the deletion of a specific case.

2.3.2 *DFBETAS*

DFBETAS is also an important measure of influence based on the estimated regression coefficients, which often play an important role when analysing regression models. To start the discussing on *DFBETAS*, we first introduce the concept of *DFBETA* (without the ‘*S*’), which measures the change occurring in the estimated coefficients after the i^{th} observation has been deleted, and according to Belsley et al. (1980), it is calculated as follows:

$$DFBETA_i = \mathbf{b} - \mathbf{b}_{(i)} = \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i e_i}{1 - h_{ii}}, \quad i = 1, 2, \dots, n,$$

where $\mathbf{X}$ is the $n \times p$ design matrix, $\mathbf{X}_i$ is $p \times 1$ vector populated with the elements of the i^{th} row of the design matrix, h_{ii} is the leverage defined in (2.7), e_i is the i^{th} element of the vector of residuals defined in (2.3), and $\mathbf{b}$ and $\mathbf{b}_{(i)}$ are the vectors of estimated regression coefficients before and after the deletion of the i^{th} observation, respectively, with $\mathbf{b}_{(i)} = (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^\top \mathbf{Y}_{(i)}$, where $\mathbf{X}_{(i)}$ denotes the $((n-1) \times p)$ design matrix constructed from the $(n \times p)$ design matrix $\mathbf{X}$ but with the i^{th} row deleted, and where $\mathbf{Y}_{(i)}$ denotes the $((n-1) \times 1)$ vector of response values where the i^{th} value is deleted.

Furthermore, according to Belsley et al. (1980), a scaled measure for the change in b_j before and after the deletion of the i^{th} case, is defined as *DFBETAS* and is given by the following formula:

$$DFBETAS_{ij} = \frac{b_j - b_{j(i)}}{s_{(i)} \sqrt{(\mathbf{X}^\top \mathbf{X})_{jj}^{-1}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^n c_{jk}^2}} \frac{e_i}{s_{(i)}(1 - h_{ii})},$$

$i = 1, 2, \dots, n$ and $j = 0, 1, \dots, p-1$, where $s_{(i)}^2$ is the estimate of σ^2 , given by:

$$s_{(i)}^2 = \frac{1}{n - p - 1} \sum_{k \neq i} [Y_k - \mathbf{X}_k^\top \mathbf{b}_{(i)}]^2, \quad (2.9)$$

and c_{ji} is the element in the j^{th} row and i^{th} column of the matrix $\mathbf{C}$, defined by

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top.$$

For large values of $|DFBETAS_{ij}|$, the corresponding observed values will have a large influence in the calculation of b_j . Kutner et al. (2005) gave a guideline that can be used to identify observations as influential if their corresponding $|DFBETAS|$ value is large enough. In cases where we have small to moderate sample sizes, a cut-off value of 1 is suggested and in the case where one consists of a large data set, the cut-off value equal to $2/\sqrt{n}$ is suggested. This latter cut-off value will be discussed in more detail in Section 2.3.6.

2.3.3 DFFITS

Similar to the previous section, the measure of influence discussed here is also based on the change in the estimated values in a regression model. Specifically, it considers the change in the fitted values of the regression model before and after the deletion of an observation. This measure is called *DFFIT* and is performed by observing how the fit has changed after deleting an observation. The measure is given by

$$DFFIT_i = \hat{Y}_i - \hat{Y}_{i(i)} = \mathbf{X}_i[\mathbf{b} - \mathbf{b}_{(i)}] = \frac{h_{ii} e_i}{1 - h_{ii}}, \quad i = 1, 2, \dots, n,$$

where $\hat{Y}_{i(i)}$ is the fitted value obtained from the regression model where the i^{th} observation was deleted. As in the case of *DFBETA*, the measure *DFFIT* needs to be scaled. The scaling factor proposed by Belsley et al. (1980), is the standard deviation of $\hat{Y}_i$, given by $\sigma\sqrt{h_{ii}}$. Therefore, the scaled *DFFIT* is

$$DFFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \dots, n, \quad (2.10)$$

where $s_{(i)}^2$ is defined as in (2.9). We can also look at these scaled differences for when some observation other than the i^{th} one has been deleted (we could, for example, delete the k^{th} observation and compare $\hat{Y}_i$ to $\hat{Y}_{i(k)}$), however the measure *DFFITS_i* presented above is sufficient for measuring influence (Belsley et al., 1980). The other scaled differences only needs to be investigated in cases where $|DFFITS_i|$ is large (Belsley et al., 1980). Again, Kutner et al. (2005) gave some suggestions that can be used as cut-off values for the $|DFFITS|$. For small to moderate sample sizes, Kutner et al. (2005) suggested using 1 as the cut-off value, just as in the case of *DFBETAS*, and when you have a large data set, the cut-off value $2/\sqrt{p/n}$ is suggested. The latter cut-off value will also be discussed further in Section 2.3.6.

2.3.4 CovRatio

The *CovRatio* measure is once again based on the idea of deleting an observation and then measuring the change in some value. In this case, the change that is measured is in the covariance matrix of the coefficient estimates, which make use of leverages, h_{ii} , as well as the regression residuals, e_i . Two other forms of the regression residuals are also used in the *CovRatio* measure namely, the *standardized residuals* and the *studentized residuals*, defined next. The standardized residuals divides the regression residuals with their estimated standard error, which are obtained from the variance of the regression residuals given by

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1, 2, \dots, n. \quad (2.11)$$

The proof of this property can be found in Appendix C. By making use of (2.11), the standardized residuals can be defined by

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \dots, n, \quad (2.12)$$

where s is an estimate for σ , and s^2 is given by

$$s^2 = \frac{1}{n - p} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (2.13)$$

The studentized residuals on the other hand, have the same formula as the standardized residuals, but instead of estimating σ with s , we estimate it with $s_{(i)}$, as defined in (2.9). The studentized residual is therefore given by

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}, \quad i = 1, 2, \dots, n. \quad (2.14)$$

Now, for *CovRatio*, as already mentioned above, single-case deletion is implemented, but instead of measuring the change in the regression coefficients or the fitted values, the change in the covariance matrix of the coefficient estimates of the regression model are measured. The two covariance matrices before and after deletion of the i^{th} observation, are given by

$$\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} \quad \text{and} \quad \sigma^2 \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1},$$

respectively. To be able to compare these two measures, their determinants are calculated and then their ratio is taken, i.e.,

$$\frac{\det \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1}}{\det \left(\mathbf{X}^\top \mathbf{X} \right)^{-1}}.$$

This ratio consists of two matrices which only differ by the fact that one includes the i^{th} observation and the other one excludes it. For this reason, if the ratio is close to 1, the conclusion can be made that the i^{th} observation has little to no effect on the covariance matrix and that these two matrices are ‘close’ to each other. However, this ratio is not useful by itself, since it only includes the two versions of the $\mathbf{X}$ matrix and excludes the estimate of σ , which would also change when the i^{th} observation is removed from the data. A better ratio, called the *CovRatio*, makes use of $s^2(\mathbf{X}^\top \mathbf{X})^{-1}$ and $s_{(i)}^2 \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1}$, and is given by

$$\begin{aligned} \text{CovRatio}_i &= \frac{\det \left(s_{(i)}^2 \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1} \right)}{\det \left(s^2 \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right)} \\ &= \frac{s_{(i)}^{2p}}{s^{2p}} \left(\frac{\det \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1}}{\det \left(\mathbf{X}^\top \mathbf{X} \right)^{-1}} \right), \quad i = 1, 2, \dots, n. \end{aligned}$$

In the process of finding influential observations using *CovRatio*, we are interested in observations that have *CovRatio*_{*i*} values that (greatly) differ from 1. To determine how much this ratio should deviate from 1 such that the value is identified as a possible influential value, another form of the *CovRatio* measure will be used, defined as

$$\text{CovRatio}_i = \frac{1}{\left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p} \right)^p (1-h_{ii})}. \quad (2.15)$$

The proof of how this form of *CovRatio* was obtained can be found in Appendix D. Using the new form of *CovRatio* given in (2.15), together with two extreme cases, the cut-off values will be obtained for *CovRatio* (i.e., we obtain the magnitude of the distance of

$CovRatio$ from 1 that will indicate a possible influential observation).

The first extreme case that will be used is where the studentized residual is ‘extreme’, specifically, we consider the case where $|t_i| \geq 2$. In this case, note that h_{ii} has a minimum equal to $1/n$, which will go to 0 as n goes to infinity, producing the following approximate form for $CovRatio$:

$$CovRatio_i \approx \frac{1}{\left(1 + \frac{t_i^2 - 1}{n-p}\right)^p}.$$

However, in the extreme case where $|t_i| \geq 2$, we have the following bound for this value:

$$CovRatio_i \approx \frac{1}{\left(1 + \frac{t_i^2 - 1}{n-p}\right)^p} \leq \frac{1}{\left(1 + \frac{3}{n-p}\right)^p}.$$

Now, replacing $n - p$ by n in the denominator of the second part of the above expression (for simplicity), and multiplying out this denominator, we note that

$$\left(1 + \frac{3}{n}\right)^p = 1 + \frac{3p}{n} + O\left(\frac{1}{n^2}\right).$$

Further, recall that a geometric series is given by:

$$\sum_{n=0}^{\infty} a^n = 1 + a + a^2 + \dots = \frac{1}{1-a}, \quad |a| < 1.$$

Finally, using the above two results, we can say

$$\frac{1}{\left(1 + \frac{3}{n-p}\right)^p} \approx \frac{1}{\left(1 + \frac{3p}{n}\right)} \approx 1 - \frac{3p}{n}. \quad (2.16)$$

Note that, in our case, $a = -\frac{3p}{n}$ and that we only have the first part of the geometric series, since we can conclude that, as $n \rightarrow \infty$, the terms greater or equal to a^2 will become very small. From (2.16), it can be concluded that an observation with corresponding $CovRatio_i < 1 - \frac{3p}{n}$, is possibly influential and further investigation would be necessary. Note however, that this cut-off value can only be used when $n > 3p$.

The second extreme case that will be used is where the leverage is ‘extreme’, i.e., $h_{ii} \geq 2p/n$, and where the studentized residual is at a minimum, i.e., $t_i = 0$. In this case, the bound for the $CovRatio$ becomes

$$CovRatio_i \geq \frac{1}{\left(1 - \frac{1}{n-p}\right)^p \left(1 - \frac{2p}{n}\right)},$$

and, following an approach similar to the one used to obtain (2.16), the above expression can be further approximated and reduced as follows:

$$\begin{aligned} \frac{1}{\left(1 - \frac{1}{n-p}\right)^p \left(1 - \frac{2p}{n}\right)} &\approx \frac{1}{\left(1 - \frac{p}{n-p}\right) \left(1 - \frac{2p}{n}\right)} \\ &\approx \frac{1}{\left(1 - \frac{3p}{n}\right)} \\ &\approx 1 + \frac{3p}{n}. \end{aligned}$$

From this result, it can be concluded that an observation with $CovRatio > 1 + \frac{3p}{n}$ might be influential and needs further investigation. In summary, it can be concluded that all observations with $|CovRatio - 1| > 3p/n$ are flagged as potential influential observations and need further investigation.

The last measure of influence for this chapter, which will be discussed next, is called Cook's distance. This measure will be used as influential measure in some modern influential diagnostic methods and two newly proposed methods, that will be discussed in later chapters. Furthermore, Cook's distance will form the focus of this dissertation and will be used in the simulation study as the traditional influence diagnostic method to compare the results of the modern influential diagnostic methods to, as well as the results of the newly proposed methods. This measure is again based on investigating the change in the fitted values of the regression model when deleting an observation from the data set, however, instead of considering the change occurring in only the 'one' fitted value, $\hat{Y}_i$, Cook's distance considers the change in all n fitted values when the i^{th} observation is deleted.

2.3.5 Cook's distance

Cook's distance is a measure of influence that indicates the effect on all the n fitted values by deleting the i^{th} observation from the data set, as mentioned above and is defined by Kutner et al. (2005) as follows:

$$D_i = \frac{\sum_{j=1}^n \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{ps^2},$$

where s^2 is defined as in (2.13), and $\hat{Y}_j$ and $\hat{Y}_{j(i)}$ are the j^{th} fitted values before and after the deletion of the i^{th} observation, respectively. Cook's distance can also be written in matrix notation as

$$\begin{aligned} D_i &= \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} \right)^{\top} \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} \right)}{ps^2} \\ &= \frac{\left(\mathbf{b} - \mathbf{b}_{(i)} \right)^{\top} \mathbf{X}^{\top} \mathbf{X} \left(\mathbf{b} - \mathbf{b}_{(i)} \right)}{ps^2}, \end{aligned} \quad (2.17)$$

where $\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\mathbf{b}_{(i)}$.

From (2.17), it is easy to note that Cook's distance has the same form as the $(1 - \alpha) \times 100\%$ confidence ellipsoid of $\boldsymbol{\beta}$, defined by Cook (1977) as the set of all vectors, say $\tilde{\mathbf{b}}$, satisfying the following equation:

$$\frac{\left(\mathbf{b} - \tilde{\mathbf{b}} \right)^{\top} \mathbf{X}^{\top} \mathbf{X} \left(\mathbf{b} - \tilde{\mathbf{b}} \right)}{ps^2} \leq F_{p, n-p}(1 - \alpha),$$

where $F_{p, n-p}(1 - \alpha)$ is the $1 - \alpha$ value of the central F -distribution with degrees of freedom p and $n - p$. Even though this measure is designed to be used for the difference between $\boldsymbol{\beta}$ and $\mathbf{b}$, the above expression also provides a rough measure of the distance between $\mathbf{b}_{(i)}$

and $\mathbf{b}$ in terms of a known probability distribution. Suppose, for example, that Cook's distance is approximately equal to $F_{p,n-p}(0.5)$. This would mean that $\mathbf{b}_{(i)}$ is on the 50% confidence ellipsoid of $\boldsymbol{\beta}$ based on $\mathbf{b}$. Therefore, in this case, when the i^{th} observation was deleted, the least squares estimates of the regression coefficients were greatly altered. From this example it is clear that, when an observation has a large Cook's distance value, it indicates that that observation has a great effect on the estimate of $\boldsymbol{\beta}$ and is probably an influential observation. The way in which the effect can be measured then, is by comparing Cook's distance to the central F -distribution with p and $n - p$ degrees of freedom, and obtaining the corresponding percentile value of the F -distribution corresponding to the calculated value of Cook's distance. This is the same as saying that we are searching for the level of the confidence ellipsoid with center value equal to $\mathbf{b}$, that passes through $\mathbf{b}_{(i)}$. For non-influential observations, one would expect that the Cook's distance would stay within the 10%, or at least the 20%, confidence ellipsoid of $\boldsymbol{\beta}$. According to Kutner et al. (2005), a good cut-off value for Cook's distance is the 50% probability value of the F -distribution with p and $n - p$ degrees of freedom, i.e., the median of this distribution, denoted by $F_{p,n-p}(0.5)$. Cook and Weisberg (1982), on the other hand, suggested a much simpler cut-off value for Cook's distance, namely, 1. For a Cook's distance value greater than 1, Cook and Weisberg (1982) state that this value corresponds to a distance between $\mathbf{b}$ and $\mathbf{b}_{(i)}$, which is greater than the 50% confidence region.

From the formula of Cook's distance given in (2.17), it seems that it would be necessary to refit a regression model for each time a different observation is deleted, which would result in an exhaustive number of calculations needed for finding influential observations in the data set by making use of Cook's distance. This is not the case though, since Cook's distance can be written in the following simplified form:

$$D_i = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right), \quad i = 1, 2, \dots, n, \quad (2.18)$$

where r_i are the standardized residuals (as defined in (2.12)), p is the number of coefficient parameters in the linear regression model, and h_{ii} are the leverages. The proof of how this formula is obtained can be found in Appendix E. Note now, that Cook's distance depends only on three elements, namely, r_i , p , and h_{ii} . Therefore, Cook's distance will have a large value for an observation i , when either its corresponding standardized residual value is large or its leverage value is large.

In Sections 2.3.1 to 2.3.5, we have discussed various measures of influence, but we have only discussed cut-off values in detail for the leverages (see Section 2.3.1), *CovRatio* (see Section 2.3.4), and for Cook's distance, as can be seen above. In the section that follows, we will give attention to the cut-off values for those measures of influence that were not discussed yet, or that were only mentioned, as well as explanations why they are chosen the way they are.

2.3.6 General methods for finding cut-off values for measures of influence

In this subsection, two methods proposed by Belsley et al. (1980) that are used to decide on suitable cut-off values for influence diagnostic measures will be discussed. The first of these methods will be used to explain how the cut-off values of *DFBETAS* and *DFFITS* are commonly and traditionally obtained (given in Sections 2.3.2 and 2.3.3, respectively), whereas the second method is given as a possible alternative method for finding these cut-off values. These two methods are: External scaling and internal scaling.

External scaling According to Belsley et al. (1980), external scaling is defined as the method that calculates cut-off values using properties known in statistical theory. Measures like *DFBETAS* and *DFFITS* for example, all use appropriate estimated standard errors as a scaling factor. Furthermore, when the Gaussian assumption holds, these estimated standard errors are stochastically independent of their corresponding influential measures. In these cases a good starting point is to compare the magnitude of the influential measures' values with two (i.e., if any observation consists of an influential measure with absolute value larger than two, that observation is seen as a possible influential observation). Cut-off values chosen in this way, are defined as 'absolute cut-offs'. Even though *DFBETAS* and *DFFITS* are directly based on the sample size, absolute cut-offs can be used for them as well. This is because of the fact that it would be a strange event if the deletion of an observation from a data set with sample size 100 or more, results in a change in some estimated statistic that is greater or equal to two standard errors. For leverages and *CovRatio*, on the other hand, absolute cut-off values cannot be used, since they have not been scaled by some standard error.

For some influential measures it is sufficient to have cut-off values that do not directly depend on the sample size. However, in other cases, a method might be needed which obtains cut-off values that reveal, regardless of the size of your sample, roughly the same proportion of possible influential observations. These type of cut-off values are called 'size-adjusted cut-offs'. A suitable size-adjusted cut-off value for *DFBETAS* can be obtained by making use of a special form of *DFBETAS*, obtained in the case where the linear regression model only includes the intercept term. For this special case, $h_{ii} = 1/n$, $c_{ji} = 1/n$, and therefore, *DFBETAS* becomes:

$$DFBETAS_i = \frac{\sqrt{n}e_i}{(n-1)s_{(i)}}. \quad (2.19)$$

From (2.19), it is clear that $DFBETAS_i$ would decrease as the sample size increases. The size of this decrease *DFBETAS* is proportional to $\sqrt{n}$ and therefore, a suitable size-adjusted cut-off for *DFBETAS* would be given by $2/\sqrt{n}$. The same can be done for *DFFITS*. First note that, since the leverages add up to p , in the case where you have a perfectly balanced design matrix, we have that $h_{ii} = p/n$ for all $i = 1, 2, \dots, n$. Replacing h_{ii} with p/n in (2.10), the *DFFITS* measure is given by

$$DFFITS_i = t_i \sqrt{\frac{p}{n-p}}, \quad (2.20)$$

where t_i is the studentized residuals given in (2.14). Therefore, from (2.20) it is clear that after replacing $n - p$ with n (for the sake of simplicity), the size-adjusted cut-off value for *DFFITs* is $2\sqrt{p/n}$. Note that these two cut-offs are based on a number of model assumptions that would not necessarily hold for all models in practice.

Internal scaling The second method proposed by Belsley et al. (1980), is internal scaling. When the influence diagnostic measures are calculated, a series of values are obtained. In the case of *DFFITs*, for example, a series of size n is obtained and for *DFBETAS*, a series of size $p - 1$ is obtained. Now, according to Belsley et al. (1980), internal scaling is defined as the method that flags observations as influential by looking at the weights of their corresponding diagnostic measure's value relative to the other values in the series of the diagnostic measure. Internal scaling is therefore a method used for obtaining cut-off values by calculating its series of values and using them to determine the extent in which an observation shows itself as influential relative to the other observations. The way in which this is done is by calculating the sample inter-quartile range, q_R , of the series of values obtained from the diagnostic measure. A cut-off value that can be used for the diagnostic measures, as proposed by Tukey (1977), is $7q_R/2$.

Since the measures of influence discussed in this section have cut-off values based on large sample theory and a number of other restrictive model assumptions, it might happen that these measures give faulty results. In cases where the sample size is small and the data are skewed for example, the large sample conditions are not met. In similar cases, a better choice for cut-off values would be to calculate them by making use of the true distribution of the influential diagnostic measure or even an approximation for this distribution. In Chapter 4, modern methods based on this idea will be provided, and in Chapter 5, new and improved methods based on this idea, will be investigated. These methods mostly make use of resampling methods and therefore, in the next chapter, a brief discussion will follow on resampling methods.

Chapter 3

Resampling methods for linear regression

Recall that the focus of this dissertation is to determine appropriate cut-off values for the measures of influence discussed in Chapter 2.3, and so a major focus will be on approximating the sampling distribution of these measures (specifically we will primarily consider Cook's distances). Since approximating distributions of complex statistics is one of the main goals, the bootstrap will play a very important role in this discussion. In this chapter we will provide a basic outline of the bootstrap methods as given by Efron and Tibshirani (1993) and Davison and Hinkley (1997) (among others), that will be required for later sections. We first start with some important bootstrap notation that will be used throughout this dissertation after which a brief discussion of the EDF will be provided. Next, the plug-in principle will be discussed (an important aspect used in the bootstrap), followed by the methodology used to practically implement the bootstrap to perform simulations. In this latter section, we will provide an example in the form of an algorithm, where the quantiles of the distribution of a statistic are estimated using the bootstrap. The last three sections are the most relevant to this dissertation and include information about the application of the bootstrap to the linear regression model, the jackknife, as well as the jackknife-after-bootstrap (JaB) method that will be used in Chapters 4 and 5. For more detailed theoretical information about the bootstrap, see Hall (1992) and Shao and Tu (1995).

3.1 The bootstrap

According to Efron and Tibshirani (1993) and (Varian, 2005), the bootstrap is a computer based method that uses quick and easy resampling methods to assign measures of accuracy to statistical estimates, and it obtains approximations of the distribution of many statistics. Not only is the bootstrap method quick and easy to implement, Hall (1992) stated that bootstrap estimators are often desirable as they typically outperform traditional normal approximations in many cases, for example where the sample size is small or where model assumptions cannot be met.

The main idea behind the bootstrap is that it tries to reproduce the process of sampling

from the true distribution function, F , that produced the given sample data. However, instead of sampling from F , the bootstrap method makes use of the plug-in-principle (see Section 3.1.3) to samples from $\hat{F}$ instead, which is an estimate of the true distribution function, F . There are many different choices for $\hat{F}$, but the most common one, and also the one that will be used in this dissertation, is the empirical distribution function (EDF) of the sample data, denoted by F_n (this distribution function will be discussed in Section 3.1.2). Note that the resulting sample obtained by sampling from $\hat{F}$ will be called a *bootstrap resample*. If a statistic is then calculated from data obtained from $\hat{F}$, the core idea underpinning the bootstrap method is then that this statistic's distribution would be close to the distribution of the statistic obtained when the data were generated from F . By repeating the process of sampling from $\hat{F}$, calculating the statistic of interest, and noting the values of each of these statistics calculated in this way, one can obtain an approximation of the distribution of the statistic.

3.1.1 Bootstrap notation

We will now formally define some of the notation that will be used throughout this chapter. Let $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, with $\mathbf{X}_i \in \mathbb{R}^k$, $i = 1, 2, \dots, n$ and $k \geq 1$, denote a random sample that was drawn independently from an unknown distribution function, F . When a bootstrap resample is generated from the data set $\mathbf{X}$, it will be denoted by $\mathbf{X}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_n^*\}$. This bootstrap resample can also be defined as a sample drawn independently from an estimator of the true distribution function F . The estimator for the function F is then denoted by $\hat{F}$ and, in many cases, is chosen as the EDF, denoted F_n (this distribution function will be discussed in the next section). Throughout this chapter, population parameters will typically be defined using Greek letters, for example θ , and their corresponding estimates will often be represented using the same symbol, but with an added hat to the parameter, i.e., $\hat{\theta}$. Lastly, if the bootstrap resample $\mathbf{X}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_n^*\}$ is used in the calculation of a bootstrap statistic, this statistic will then be denoted by $\hat{\theta}^*$.

3.1.2 The empirical distribution function (EDF)

As mentioned earlier, the EDF serves as an estimator of the true distribution function F and can be used to generate bootstrap resamples. Furthermore, the EDF depends only on the sample data, which makes it a non-parametric estimator. Suppose, for example, that you have an observed univariate random sample, $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, that was drawn from some unknown distribution function F . The EDF, F_n , is then defined as a discrete function that assigns equal probability to each observation in the sample. According to Davison and Hinkley (1997), the EDF is defined as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x),$$

where I indicates the indicator function defined as

$$I(A) = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } A^c \text{ occurs.} \end{cases}$$

Note that sampling independently from F_n , is equivalent to sampling with replacement from the sample data $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$.

Next, we will discuss the plug-in principle, which is a very important tool used in the bootstrap, since the plug-in principle is directly applied when the bootstrap estimates properties of some statistic, $\hat{\theta}$.

3.1.3 The plug-in principle

The plug-in principle is a simple method that can be used to estimate population parameters. Suppose we want to estimate some parameter θ defined as a functional of an unknown distribution function F , that is, define $\theta = \varphi(F)$. The plug-in principle is then performed by applying the same functional, φ , to an empirical estimator for F , $\hat{F}$, where we typically choose $\hat{F} = F_n$. Formally, the plug-in estimator for θ is defined as

$$\hat{\theta} = \varphi(\hat{F}).$$

Efron (2003) provide the following schematic representation of the bootstrap and its implementation of the plug-in principle to explain these concepts:

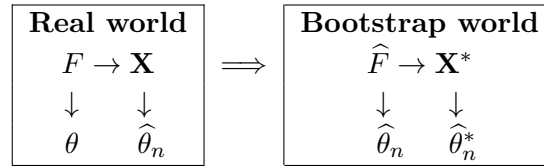


Figure 3.1: Efron's (2003) schematic representation of the plug-in principle.

On the left hand side of Figure 3.1, F is some unknown probability structure from which the sample $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ was sampled. From $\mathbf{X}$, we can then calculate the statistic $\hat{\theta}$, which is an estimate of some true population parameter θ . We define the left hand side of the representation above as the 'Real world' and the right hand side as the 'Bootstrap world'. In the Bootstrap world, $\hat{F}$ is an estimate of the true distribution function F from which the bootstrap samples, $\mathbf{X}^* = \{X_1^*, X_2^*, \dots, X_n^*\}$, can be drawn. Lastly, using this bootstrap sample, the statistic $\hat{\theta}^*$ can be calculated. If the probability structure was known in the Real world, we would have been able to calculate the unknown population parameter, θ , and obtain the distribution of the statistic $\hat{\theta}$. This is all possible in the Bootstrap world though, since here, the probability structure is known. The bootstrap therefore shifts from a world where the probability structure is unknown (the Real world), to a world where it is known (the Bootstrap world) and then uses the information obtained in the Bootstrap world to estimate similar elements in the Real world. The way in which the bootstrap shifts between worlds is by applying the plug-in principle. Once the shift has occurred, one is able to determine the distributional properties of the statistic $\hat{\theta}^*$, and the results obtained in this way can be used as estimates for corresponding elements in the Real world.

The plug-in principle can therefore be implemented in the following way: Suppose F

is the unknown distribution function and $\hat{F}$ is chosen as F_n , the EDF of the data set $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, sampled from F . Suppose also that the unknown population parameter θ , is some functional of F , that is,

$$\theta = \varphi(F).$$

To estimate this parameter, we can use the plug-in estimator, which is obtained by directly applying the plug-in principle as follows:

$$\hat{\theta} = \varphi(F_n).$$

Applying the plug-in principle again, we shift from the real world to the Bootstrap world to estimate the statistic $\hat{\theta}$. This will be done as follows:

$$\hat{\theta}^* = \varphi(F_n^*),$$

where F_n^* is the EDF of the resampled bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, \dots, X_n^*\}$.

Now that we know how the bootstrap works, we will discuss the basic implementation of the bootstrap in the next section, followed by a discussion on how the bootstrap can be use in linear regression in Section 3.2.

3.1.4 The basic implementation of the bootstrap

The way in which the bootstrap estimates the properties of some statistic $\hat{\theta}$ is by making use of the idea that, in the Bootstrap world, it is accepted that the observed sample data is the whole population. A sample, $\mathbf{X}^* = \{X_1^*, X_2^*, \dots, X_n^*\}$, is then drawn independently from F_n (the true population distribution function in the Bootstrap world), and will be referred to as a *bootstrap resample*. From this bootstrap resample, the bootstrap statistic can be calculated, denoted by $\hat{\theta}^*$. The process of generating a bootstrap sample and calculating a statistic from this samples is then repeated B times to give B *bootstrap replicates* of the statistic, $\hat{\theta}^*$, that is, we have the B replicates $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. These replicates can be used as an approximation to the distribution of the bootstrap statistic $\hat{\theta}^*$ as $B \rightarrow \infty$. In turn, this approximation can then be used as an estimate for the true distribution of the statistic $\hat{\theta}$.

As an illustration of the implementation of the bootstrap, the following example is provided: Suppose we are interested in using the bootstrap to estimate the α -quantile value, q_α , of the distribution of the statistic $\hat{\theta}$, defined as $P(\hat{\theta} < q_\alpha) = \alpha$. Applying the plug-in principle, we obtain the bootstrap estimator, $\hat{q}_\alpha$, defined as $P^*(\hat{\theta}^* < \hat{q}_\alpha) = \alpha$, where $P^*(\cdot)$ denotes the ‘bootstrap’ probability operator and is simply defined as the conditional probability $P^*(\cdot) = P(\cdot | X_1, X_2, \dots, X_n)$. According to Johnson (2001), the estimator $\hat{q}_\alpha$ can then be approximated using the following algorithm:

Algorithm 1 *Approximating the bootstrap estimate of q_α*

1. Suppose we have the observed data set $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$.
2. Sample n observations with replacement from $\mathbf{X}$ to obtain the bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, \dots, X_n^*\}$.
3. Using the bootstrap sample $\mathbf{X}^*$, calculate the statistic $\hat{\theta}^*$.
4. Independently repeat steps 2 and 3 B times to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$.
5. Sort these bootstrap replicates to obtain the order statistics $\hat{\theta}_{1:B}^* \leq \hat{\theta}_{2:B}^* \leq \dots \leq \hat{\theta}_{B:B}^*$.
6. The desired quantile is now given by:

$$\hat{q}_{\alpha,B}^* = \theta_{a:B}^*,$$

where

$$a = \lfloor B \times \alpha \rfloor.$$

3.2 The bootstrap in linear regression

In later chapters, we consider a bootstrap-based method proposed by Martin and Roberts (2010) for finding an approximation to the distribution of Cook's distance in order for us to obtain more appropriate cut-off values than those proposed in the previous chapter. In cases where data are non-normal or when data sets are very small for example, the new cut-off values based on the approximated Cook's distance will potentially perform better than the traditional cases in the process of identifying influential observations. In order for us to obtain such an approximation, an important step is to apply the bootstrap to regression models. In this section, two commonly used methods will be discussed for this purpose namely, 'residual-based resampling' and 'case-based resampling'.

3.2.1 Residual-based resampling

The first method that we will consider is a model-based method that makes an assumption about the form of the relationship between the expected value of the response variable and a number of covariates. This method, called the 'residual-based' method of bootstrap resampling, also makes an additional assumption that the covariates are fixed from one sample to the next. To introduce the concepts, consider the multiple linear regression setup as defined in (2.1) where we have a fixed $(n \times p)$ design matrix $\mathbf{X}$ and a $(n \times 1)$ random vector $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^\top$, with the values $Y_1, \dots, Y_n$ defined by the regression function

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

and where $\varepsilon_1, \dots, \varepsilon_n$ is a sample from some unknown distribution function G . The first step towards generating bootstrap samples of the response variable would be to estimate the distribution function G , and good estimate for it would be to use the EDF of the regression sample residuals, which are defined as in (2.3). This can easily be done, since we are able to obtain the least squares estimates of the regression coefficients for this

model. According to Davison and Hinkley (1997), the following modified form of the residuals can be used as well:

$$r_i = \frac{e_i}{\sqrt{1 - h_{ii}}},$$

where h_{ii} is defined as in (2.7). Davison and Hinkley (1997) state that these modified residuals are a better option in practical use, since their variances agree with the variances of ε_i . Lastly, since we assume that the error terms, ε_i , have expected value equal to 0 in this model, the EDF of the residuals must also have a zero mean so that we mimic the property of the errors when drawing the bootstrap samples. This can be accomplished by centering the residuals as follows:

$$\hat{\mathbf{e}} = \mathbf{e} - \mathbf{1}\bar{e}, \quad (3.1)$$

where

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i \quad \text{and} \quad \mathbf{1} = [1, 1, \dots, 1]^\top.$$

Note that for the case where ordinary least squares is used with the multiple linear regression model in (2.1), we have that $\bar{e} = 0$, and so, for this case we may use either the centered or un-centered residuals for this procedure. The EDF of these residuals is now used as an estimate for G , and is denoted by G_n . From this estimated distribution, bootstrap resamples can be obtained by sampling independently from G_n , which is the same as sampling with replacement from $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$, and the resulting bootstrap sample residuals are denoted by $\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*$. The bootstrap sample of response values can now be obtained by making use of the following expression

$$Y_i^* = b_0 + b_1 X_{i1} + \dots + b_{p-1} X_{i,p-1} + \hat{e}_i^*, \quad i = 1, 2, \dots, n,$$

where $\mathbf{b} = [b_0, b_1, \dots, b_{p-1}]^\top$ are the ordinary least squares estimates of $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_{p-1}]^\top$ as given by (2.5).

Now that we have these bootstrap samples it is possible to estimate the properties of various statistics related to regression. For example, it is possible to estimate the distributional properties of the least squares estimators of the regression parameters. By applying the method of least squares on the observed covariates $\mathbf{X}$ and the bootstrap sample, $\mathbf{Y}^* = \{Y_1^*, Y_2^*, \dots, Y_n^*\}$, one can obtain a vector of bootstrap regression coefficients, denoted by

$$\underset{(p \times 1)}{\mathbf{b}^*} = \underset{(p \times p)}{(\mathbf{X}^\top \mathbf{X})^{-1}} \underset{(p \times n)}{\mathbf{X}^\top} \underset{(n \times 1)}{\mathbf{Y}^*}. \quad (3.2)$$

Using these bootstrap statistics we can estimate the distributional properties of $\mathbf{b}$. In the following algorithm, we illustrate how to approximate the bootstrap estimate of the sampling distribution of the elements of $\mathbf{b}$, using the residual-based method:

Algorithm 2 *Approximating the bootstrap estimate of the sampling distribution of the elements of $\mathbf{b}$ using the residual-based method*

1. Suppose we have the following observed data set: $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^\top$ and fixed $(n \times p)$ design matrix, $\mathbf{X}$.
2. Calculate the estimate of the regression coefficient vector, β , by applying the method of least squares. We now have, $\mathbf{b}$.
3. Calculate the residuals using the following expression:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}.$$

Therefore, we have $\mathbf{e} = [e_1, e_2, \dots, e_n]^\top$.

4. Obtain the centered residuals $\hat{\mathbf{e}} = [\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n]^\top$, as shown in (3.1). We now have G_n , the EDF of these centered residuals.
5. Sample independently from G_n to obtain the bootstrap residuals, denoted by $\hat{\mathbf{e}}^* = [\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*]^\top$.
6. Determine the bootstrap sample $\mathbf{Y}^* = \{Y_1^*, Y_2^*, \dots, Y_n^*\}$, using the following:

$$\mathbf{Y}^* = \mathbf{X}\mathbf{b} + \hat{\mathbf{e}}^*.$$

7. Calculate the bootstrap estimate of the vector $\mathbf{b}$, denoted by $\mathbf{b}^*$, as defined in (3.2).
 8. Independently repeat steps 5 to 7 B times to get bootstrap replicates of $\mathbf{b}^*$, denoted by $\mathbf{b}_1^*, \mathbf{b}_2^*, \dots, \mathbf{b}_B^*$, which is then an approximation to the sampling distribution of $\mathbf{b}$.
-

3.2.2 Cases-based resampling

The cases-based resampling method differs from the residual-based resampling method in that, when using the residual-based method, interest lies in the estimation of the error distribution, whereas now interest lies in estimating the joint distribution of $(\mathbf{X}_i, Y_i)$, where $\mathbf{X}_i \in \mathbb{R}^p$, $i = 1, 2, \dots, n$, are $(p \times 1)$ vectors denoting the i^{th} row of the $(n \times p)$ design matrix $\mathbf{X}$. Therefore, this approach makes no model based assumptions about the form of the relationship between the expected value of the response and the covariates, but assumes that both the response and covariates are random variables. The cases-based method therefore generates bootstrap resamples, $(\mathbf{X}_1^*, Y_1^*), (\mathbf{X}_2^*, Y_2^*), \dots, (\mathbf{X}_n^*, Y_n^*)$, by sampling with replacement from the pairs $(\mathbf{X}_i, Y_i)$, $i = 1, 2, \dots, n$ (Freedman, 1981).

As before, when the bootstrap resamples are obtained, one can calculate the distributional properties of the various statistics related to regression. For example, applying the method of least squares on these bootstrap resamples, one can approximate the estimated distribution of $\mathbf{b}$. This procedure can be summarised in the following algorithm:

Algorithm 3 *Approximating the bootstrap estimate of the sampling distribution of the elements of $\mathbf{b}$ using the cases-based method*

1. Suppose we have the observed data set $T_1 = (\mathbf{X}_1, Y_1), T_2 = (\mathbf{X}_2, Y_2), \dots, T_n = (\mathbf{X}_n, Y_n)$.
2. Sample with replacement from $T_1, T_2, \dots, T_n$ to obtain the bootstrap resamples $T_1^*, T_2^*, \dots, T_n^*$.
3. Applying the method of least squares to the elements of $T_1^*, T_2^*, \dots, T_n^*$ we can obtain the following bootstrap statistic:

$$\underset{(p \times 1)}{\mathbf{b}^*} = (\underset{(p \times p)}{\mathbf{X}^{*\top} \mathbf{X}^*})^{-1} \underset{(p \times n)}{\mathbf{X}^{*\top}} \underset{(n \times 1)}{\mathbf{Y}^*}.$$

4. Independently repeat steps 2 and 3 B times to obtained the bootstrap replicates $\mathbf{b}_1^*, \mathbf{b}_2^*, \dots, \mathbf{b}_B^*$, that can be seen as an approximation to the distribution of $\mathbf{b}$.
-

In the next section, an important method will be discussed which will be implemented in some of the influence diagnostics methods used in this dissertation. This is the jackknife-after-bootstrap method, but before it can be explained, we will need some information about the jackknife method, since the jackknife-after-bootstrap method implements the idea of the jackknife method on a bootstrap approximation to an estimator. The jackknife will therefore be discussed in the next section first, then followed by a discussion on the jackknife-after-bootstrap.

3.3 Jackknife

The jackknife is presented here as an alternative method to the bootstrap when one is interested in estimating a statistic's bias and standard error, however the main concept that needs to be introduced in this section is the manner in which new samples are formed by deleting single observations from a data set.

The basic procedure for estimating the bias and standard error of a statistic using the jackknife, starts by obtaining *jackknife samples* which are constructed by deleting a single observation from the data set; one for each separate jackknife sample. From a data set $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, the i^{th} jackknife sample will have the following form:

$$\mathbf{X}_{(i)} = \{X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}, \quad i = 1, 2, \dots, n,$$

i.e., this new jackknife sample is formed by simply deleting X_i from the original sample $X_1, X_2, \dots, X_n$. Note that a total of n jackknife samples will therefore be obtained in this way.

For each of the jackknife samples $\mathbf{X}_{(i)}, i = 1, 2, \dots, n$, the statistic of interest is calculated to further obtain the n *jackknife replicates*. Suppose for example the statistic that is investigated is $\hat{\theta}$, then the jackknife replicates of this statistic are given by

$$\hat{\theta}_{(i)} = \hat{\theta}(X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n), \quad i = 1, 2, \dots, n.$$

From these jackknife replicates, the jackknife estimates of bias and variance of $\widehat{\theta}$ can be obtained using the following two formulas, respectively, as given by Efron and Tibshirani (1993):

$$\widehat{bias}_{jack} = (n - 1) \left(\widehat{\theta}_{(\bullet)} - \widehat{\theta}_n \right),$$

where

$$\widehat{\theta}_{(\bullet)} = \frac{1}{n} \sum_{i=1}^n \widehat{\theta}_{(i)},$$

and

$$\widehat{\sigma}_{jack}^2 = \frac{n-1}{n} \sum_{i=1}^n \left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(\bullet)} \right)^2.$$

Note also that the jackknife can be extended to a delete- d jackknife where, instead of deleting a single observation at a time from the original data set to obtaining the jackknife samples, d observations are deleted. More on multiple-deletion jackknife methods can be found in Hinkley (1977), Shao and Wu (1989), Wu (1990), and Politis and Romano (1994).

Since our focus for this dissertation is not on estimating the bias or the variance of a statistic, but rather on finding quantiles for the statistic using the jackknife-after-bootstrap method, we will not go into further detail on the jackknife method, however the ideas discussed here (specifically the concepts related to the deletion of observations) are useful for these further discussions. For more information on the jackknife, see also Quenouille (1949) (the paper that first proposed the concept of the jackknife), Tukey (1958) (the paper that extended the concept to include standard error estimation), Gray and Schucany (1972), and Miller (1974).

3.4 Jackknife-after-bootstrap (JaB)

In this section, we are interested in finding a way in which the jackknife can be applied to bootstrap estimators, for example, $\text{Var}^*(\widehat{\theta}_n^*)$. This is not an easy task, since the jackknife idea of deleting the observation X_i from $X_1, X_2, \dots, X_n$ to form jackknife samples, is much more complex when one has to make use of bootstrap resamples. Therefore, in this section we will primarily focus on this concept of deleting a single observation from the bootstrap resamples. This concept is naturally very useful in the context of linear regression diagnostic methods since many of the influential diagnostic statistics involve case-deletion methods in their calculation. These concept will be explained using an example where we apply the jackknife to estimate the variance of the approximate bootstrap estimate of the quantile $\widehat{q}_{\alpha,B}^*$ (as calculated in Section 3.1.4).

Suppose we have the data set, $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, and we applied the bootstrap to this data set in order to approximate the estimate of the α -quantile, $q_{(\alpha)}$, i.e., $\widehat{q}_{(\alpha)}$, of the distribution of the statistic $\widehat{\theta}$. In Section 3.1.4, we found that this bootstrap approximation was given by

$$\widehat{q}_{\alpha,B}^* = \theta_{a:B}^*,$$

where $\widehat{\theta}_{1:B}^* \leq \widehat{\theta}_{2:B}^* \leq \dots \leq \widehat{\theta}_{B:B}^*$, and

$$a = \lfloor B \times \alpha \rfloor.$$

In some cases, one might then further be interested in determining the uncertainty in the bootstrap approximation, $\hat{q}_{\alpha,B}^*$. A way in which this can be done is by making use of a double bootstrap procedure to determine the variance of $\hat{q}_{\alpha,B}^*$. Alternatively, one can apply the basic idea of the jackknife to estimate this variance based on these bootstrap quantities, which is done by making use of the jackknife-after-bootstrap method.

If the jackknife is to be used for this purpose, one would need to calculate the bootstrap estimator $\hat{q}_{\alpha,B}^*$ based on a set of bootstrap resamples where the i^{th} value was deleted. Let b_i denote the number of bootstrap samples that do not contain X_i generated in this way. Note, however, that the approach followed here will *not* involve first deleting the i^{th} value from $X_1, X_2, \dots, X_n$ and then resampling. Instead, the jackknife-after-bootstrap method that will be used begins by first generating a complete set of B bootstrap resamples and then employs the jackknife idea of case-deletion by separating these bootstrap samples into n sets such that the i^{th} set of bootstrap resamples does not contain the i^{th} observation from the original sample. In the end, the bootstrap statistic replicates obtained from the bootstrap resamples that exclude the i^{th} observation, denoted $\hat{\theta}_{(i),j}^*$, $j = 1, 2, \dots, b_i$, are used to approximate the bootstrap estimator

$$\hat{q}_{\alpha,b_i}^* = \hat{\theta}_{(i),a:b_i}^*,$$

where $\hat{\theta}_{(i),1:b_i}^* \leq \hat{\theta}_{(i),2:b_i}^* \leq \dots \leq \hat{\theta}_{(i),b_i:b_i}^*$, and

$$a = \lfloor b_i \times \alpha \rfloor,$$

for $i = 1, 2, \dots, n$.

The approximated estimators, $\hat{q}_{\alpha,b_1}^*, \hat{q}_{\alpha,b_2}^*, \dots, \hat{q}_{\alpha,b_n}^*$, can be seen as jackknife replicates, as they are each calculated without the influence of one of the original data values. When the statistic $\hat{\theta}$ is then calculated for each of the b_i bootstrap resamples not containing the i^{th} observation, the bootstrap distribution of $\theta_{(i)}^*$ is approximated. We can therefore apply the jackknife after the implementation of the bootstrap (hence the name *jackknife-after-bootstrap*) to estimate the variance of this quantity, i.e.,

$$\widehat{\text{Var}}_{\text{jack}}(\hat{q}_{\alpha,B}^*) = \left(\frac{n-1}{n} \right) \sum_{i=1}^n \left(\hat{q}_{\alpha,b_i}^* - \hat{q}_{\alpha,b_\bullet}^* \right)^2$$

where

$$\hat{q}_{\alpha,b_\bullet}^* = \frac{1}{n} \sum_{i=1}^n \hat{q}_{\alpha,b_i}^*.$$

Remark 3.1. According to Davison and Hinkley (1997), the number of bootstrap resamples that will not contain the i^{th} observation is approximately $b_i = Be^{-1} \approx 0.368B$, where B is the total number of bootstrap resamples generated. This can be seen from the fact that, since the probability that any single bootstrap resampled value is equal to X_i is $\frac{1}{n}$, i.e., $P(X_i^* = X_j) = \frac{1}{n}$, it stands to reason that the probability that the i^{th} observation is excluded from an entire bootstrap sample is given by $(1 - n^{-1})^n \rightarrow e^{-1}$ as $n \rightarrow \infty$.

This process of using the jackknife-after-bootstrap method to calculate the variance of the bootstrap approximation of the estimate of the α -quantile, $q_{(\alpha)}$, i.e., $\widehat{q}_{(\alpha)}$, of the distribution of the statistic, $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \dots, X_n)$, is summarised in the following algorithm as briefly described in Efron and Tibshirani (1993):

Algorithm 4 *Jackknife-after-bootstrap estimate of the variance of the bootstrap estimate of the α quantile of $\widehat{\theta}$, q_α*

1. Given the observed data set $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, sample n observations with replacement from $\mathbf{X}$ to obtain the bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, \dots, X_n^*\}$.
2. Independently repeat step 1 B times to get B sets of resamples:

$$\mathbf{X}_1^* = \{X_1^*, X_2^*, \dots, X_n^*\} \quad (1)$$

$$\mathbf{X}_2^* = \{X_1^*, X_2^*, \dots, X_n^*\} \quad (2)$$

$$\vdots$$

$$\mathbf{X}_B^* = \{X_1^*, X_2^*, \dots, X_n^*\} \quad (B)$$

3. Let C_i denote the set of indices of the bootstrap resamples that do not contain the i^{th} observation, X_i . The number of indices in C_i is then denoted b_i . Then calculate the bootstrap statistic

$$\widehat{\theta}_{(i),j}^* = \widehat{\theta}(\mathbf{X}_j^*), \quad j \in C_i.$$

We now have the bootstrap replicates $\widehat{\theta}_{(i),1}^*, \widehat{\theta}_{(i),2}^*, \dots, \widehat{\theta}_{(i),b_i}^*$. In terms of their order statistics, we have: $\widehat{\theta}_{(i),1:b_i}^* \leq \widehat{\theta}_{(i),2:b_i}^* \leq \dots \leq \widehat{\theta}_{(i),b_i:b_i}^*$. Repeat this step for each $i = 1, 2, \dots, n$.

4. Calculate the α -quantile when the i^{th} value is deleted as follows:

$$\widehat{q}_{\alpha,b_i}^* = \theta_{(i),a:b_i}^*,$$

where $a = \lfloor b_i \times \alpha \rfloor$. Do this step for all observations $i = 1, 2, \dots, n$.

5. Finally, estimate the jackknife variance over all the quantiles obtained in step 4:

$$\widehat{\text{Var}}_{jack}(\widehat{q}_{\alpha,B}^*) = \left(\frac{n-1}{n} \right) \sum_{i=1}^n (\widehat{q}_{\alpha,b_i}^* - \widehat{q}_{\alpha,b_\bullet}^*)^2,$$

where

$$\widehat{q}_{\alpha,b_\bullet}^* = \frac{1}{n} \sum_{i=1}^n \widehat{q}_{\alpha,b_i}^*.$$

The reason why this jackknife-after-bootstrap method works is because of the jackknife-after-bootstrap sampling lemma from Efron (1992):

Lemma 1. *An i.i.d. sample of size n from the EDF of the sample $X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, denoted $F_{n-1,(i)}$, that is, the sample*

$$F_{n-1,(i)} \xrightarrow{i.i.d.} (X_1^*, X_2^*, \dots, X_n^*)$$

has the same distribution as a bootstrap sample from the EDF of the full sample data, denoted F_n , that is, the sample,

$$F_n \xrightarrow{i.i.d.} (X_1^*, X_2^*, \dots, X_n^*),$$

when none of the X_j^* values in this sample are equal to X_i .

Since all of the necessary theory of resampling methods was given in this chapter, we will now provide modern influential diagnostic methods that are based on resampling methods in the next chapter. Additionally, in Chapter 5, two new methods for the detection of influential observations will also be provided, one based on two resampling methods that will be discussed in the next chapter, and the other one based on a simple approach for estimating the sample distribution of Cook's distance.

Chapter 4

Modern influential diagnostic methods

As explained in the introduction chapter, we will now introduce modern methods for inference for various measures of influence in a regression model. These methods include simulation based methods that already exist in the literature. What follows in this chapter is the necessary motivation and theory for these methods.

Classic influential measure, such as Cook's D, DFFITS and DFBETAS, have complex distributions. Therefore, finding cut-off values for these measures based on their sampling distributions can be very difficult. In the literature it is common to make use of rules based on large sample theory to determine these cut-offs rather than using their sampling distributions (see, for example, Belsley et al., 1980). These large sample cut-off values, however, are typically based on the assumption that the underlying error term is normally distributed. In scenarios where the sample size is much bigger than the number of predictors and the errors of the model used are normally distributed, these cut-off values will work well, but in many scenarios, this conditions is not met, leading to cut-off values that give faulty results. For this reason, other methods have been developed for the purpose of identifying influential values, see, for example, Martin and Roberts (2010) and Marubini and Orenti (2014).

In the next chapter, an improved method will be proposed for the detection of influential observations which will be based on bootstrap critical values of the classic influence diagnostic measure, Cook's distance. This new method is based on the combination of two methods already in the literature, one attributed to Martin and Roberts (2010) and the other to Roberts et al. (2015). The first method by Martin and Roberts (2010) is used for finding bootstrap based cut-off values for a specific regression influence statistic by approximating the sampling distribution of the regression influence statistic. This data dependent cut-off value is then used in the same way as the large sample cut-off values that are traditionally used in this context. However, since multiple observations can be influential at once and have an influence on each other, an improved method for influence value detection was also separately developed by Roberts et al. (2015) to overcome problems like *masking* and *swamping*. Masking can be defined as the situation where, if one deletes

a set of influential observations in the data, one might find that another observation will reveal itself to be influential, which was not identified as such before the deletion of the first observations (i.e., the other observations mask or obscure certain influential cases). Swamping, on the other hand, occurs when a non-influential observation is flagged as being influential on the model as a result of the effect of some other influential observations in the data set (i.e., the other influential observation have the effect of swamping or overwhelming certain non-influential cases, causing them to be ‘dragged’ over into influential status). Note, however, that for the sake of simplicity the cut-off values used in the paper by Roberts et al. (2015) were chosen as the traditional large sample cut-off values.

In an attempt to reconcile these two methods that appear in the literature we will make use of the method by Roberts et al. (2015), but will determine the cut-off values for this procedure by making use of the bootstrap based method described by Martin and Roberts (2010). It is hoped that this new combined method will produce improved performance over either of the two methods. The two methods on which the new combined method is based will be discussed in this chapter as well as a commonly used method for influential value detection, namely, the MC^3 method. In the next chapter we will then explained in detail the new combined method, and a completely new method based on the approximation of Cook’s distance.

4.1 Jackknife-after-bootstrap regression influence diagnostics

Martin and Roberts (2010) created a bootstrap based method for approximating influential measures’ sampling distributions. From this approximated sampling distribution, cut-off values can be calculated and used to determine which values in a data set are influential. Using this method, the cut-off values obtained are then specific to the data set used.

Bootstrap procedures will be used to approximate the sampling distribution of these statistics (and, consequently, the cut-off values) because, while most influential measures have forms that look very similar to some well known statistics which have known distributions, none of these influential measure statistics actually have these same distributions. Indeed, the theoretical distribution of these statistics can be quite complex. For instance, DFFITS and DFBETAS look very much like t -statistics, and one might be forgiven for thinking that they are also t -distributed but, neither of them follow this distribution. Similarly, Cook’s D has the form of an F -statistic, which might leads one to think that it should have an F -distribution, but the reality is that the distribution is much more complicated. Traditionally, these problems are overcome by basing the cut-off values on the asymptotic (large sample theory) distribution of the statistic, as their finite sample distributions are theoretically intractable. However, as already mentioned, this can give untrustworthy results when the sample size is small or when the error term is non-normal. For this reason, the bootstrap method of Martin and Roberts (2010) is a simple alternative method for finding the sampling distribution of these measures of influence (and thus the cut-off values) without having to resort to complex theory.

The model used for the bootstrap method suggested by Martin and Roberts (2010), is the multiple regression model:

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times p)}{\mathbf{X}} \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}} \quad (4.1)$$

where $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ are the response variable values, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$ is the design matrix with the values for the i^{th} case given by $\mathbf{X}_i = (1, X_{i1}, X_{i2}, \dots, X_{i,p-1})$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})^\top$ are the regression coefficient parameters, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^\top$ are the random error terms with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and where $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. In order to draw bootstrap samples from this data, one approach is to make use of the case-based resampling method (see Section 3.2.2). The influence measure statistics would then be calculated for each of these bootstrap samples, and the sampling distribution of the statistic can then be estimated from the replicated bootstrap statistics. However, a major issue facing this approach is the fact that one cannot track where the i^{th} case appears in each bootstrap sample. This is important because these statistics attempt to measure the effect of a single case on the fitted regression model, and so randomly ‘reshuffling’ values in each sample in this naïve manner destroys the ability of each bootstrap sample statistic to measure this. However, the case-deletion nature of many of the influence measure statistics (DF-FITS, DFBETAS, Cook’s D, etc.) directly lend themselves to a ‘jackknife’ methodology. Therefore, the new method proposed by Martin and Roberts (2010) is based on Lemma 1, given in Section 3.4. This lemma states that drawing a bootstrap sample with replacement from $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ will yield the same distributions as when a bootstrap sample is drawn from $x_1, x_2, \dots, x_n$ with none of these values equal to x_i . This results in n sets of bootstrap samples, each one with the property that they do not contain a specific case i , $i = 1, 2, \dots, n$.

The algorithm that Martin and Roberts (2010) suggested, is given as below. Note that in the simulation study that will later be provided in this dissertation, we will be implementing the JaB method on Cook’s distance.

Algorithm 5 *Approximating the sample distribution of an influential measure, θ using the JaB method*

1. Given the data $T_1 = (\mathbf{X}_1^\top, Y_1), T_2 = (\mathbf{X}_2^\top, Y_2), \dots, T_n = (\mathbf{X}_n^\top, Y_n)$, fit the multiple regression model to the data set and compute the influential measure chosen, say $\theta_i = \theta(i; T_1, T_2, \dots, T_n)$, for $i = 1, \dots, n$.
2. Independently sample cases from the original data set to generate B resamples of the form $T_1^*, \dots, T_n^*$.
3. This step is the JaB step. In the B resamples generated in step 2, some of these resamples do not containing the i^{th} data point, T_i . Let b_i denote the number of the B resamples that do not contain T_i . For each of these b_i resamples, calculate the n influence measures, $\theta_i^* = \theta(i; T_1^*, \dots, T_n^*)$, $i = 1, 2, \dots, n$. To obtain an approximate distribution for θ_i for the i^{th} case, combine the nb_i influence measure statistic values that were obtained by the b_i resamples that do not contain the i^{th} case, into one vector, i.e., the resulting vector will have nb_i elements. Repeat this step for each $i = 1, \dots, n$.

Algorithm 5 *Continued . . .*

4. For each case, use the vector/approximated distribution calculated in step 3 and calculate the cut-off value by finding the ω^{th} quantile, where ω is some appropriately chosen proportion value used for the cut-off.
 5. Compare the influential measure's values calculated in step 1 to the obtained cut-off values calculated in step 4 and make a conclusion about the influence of each observation in the data set.
-

Remark 4.1. *Another approach for drawing bootstrap samples from a data set, is residual-based resampling (see Section 3.2.1). For case-deletion, that was used in the algorithm, deleting a single data point is very difficult since there are no clear way of doing this when working with the residual-based method. This would mean that none of the resamples obtained from the residual-based method, would have any cases missing. Case deletion is not sensible here, since no observation would have a case-deletion diagnostic that is meaningful to it. Furthermore, since the process of obtaining bootstrap samples from the residuals, models, which can be potentially be influenced by some observations, need to be fit. For these reasons, the JaB method could not use the residual-based resampling method.*

Remark 4.2. *In step 3 of the algorithm, there was mentioned that some value b_i of the B resamples do not contain a specific observation i . According to Martin and Roberts (2010), the value of b_i is roughly equal to one over Euler's number, e^{-1} . For this reason, if we use 3000 bootstrap replications, it would mean that only about 1000 of the bootstrap samples will not include the i^{th} observation from your original data set. This also means that when using the JaB method, you are going to need about e times more resamples than in the case of the ordinary bootstrap method.*

The aim of this JaB algorithm is to generate a bootstrap estimate of the null distribution of influential measures where the null hypothesis states that the i^{th} data point is not influential. To approximate this null distribution, the algorithm only uses those bootstrap samples that do not contain the i^{th} observation, thereby satisfying the null hypothesis in the sampling procedure, as the influence of the i^{th} observation in each of the individual bootstrap calculations is removed. In contrast, by using a naïve bootstrap approach where *all* of the samples are used (i.e., not only those that excluded the i^{th} observation), then the i^{th} observation would appear in many of the bootstrap resamples (in some cases, it would appear more than once). The result is that, if the i^{th} observation is highly influential, then this case will influence the values of the diagnostic measures calculated for the bootstrap samples and so these values would no longer be representative of the values that would have been obtained under the null hypothesis. The resulting distribution would have heavier tails, resulting in higher cut-offs and, ultimately, lead to obscuring the values that are 'suspicious'. For these reasons the naïve bootstrap is not appropriate and the JaB method is used instead.

Martin and Roberts (2010) used real data as well as simulations to test their new method and to compare it with the traditional methods. From their results it was concluded that, when applied to Cook's distance, mostly all the cut-off values obtained from the JaB method were smaller than the traditional cut-offs and were more asymmetric. The

JaB method's cut-off values were found to be almost the same as those of the traditional cut-off values when a big data set was used, but differed greatly when small data sets were used, or when the error terms were non-normal. This is to be expected since the new method took the features of the data into account. For this reason, the JaB method does not flag points unnecessarily as influential if they are a little bit too far in the tails of the distribution. The result is that this JaB method is able to take the form of the model and asymmetric data into account (since the distribution is based on the observed data), and displays good performance for both large and small samples. This is in contrast to traditional methods which are the same for all models used, can only be applied to normal data, and typically only work well for larger data sets. However, the JaB method does suffer from the drawback that it is computationally much more expensive than the traditional methods; Martin and Roberts (2010) report that a JaB calculation in R with four covariates and a sample size of $n = 50$ executed in 2 minutes, whereas the traditional methods ran the calculation only in a few seconds.

Papers that have already studied the performance of this method include Beyaztas and Alin (2013) and Dalposso et al. (2016), and further modifications of the method are discussed in Beyaztas and Alin (2014), where a delete-2 jackknife-after-bootstrap is discussed, Beyaztas et al. (2014), where methods involving *robust* bootstrap confidence intervals are used, and Muñoz-Pichardo et al. (2019), where additional jackknife-after-bootstrap transformations are employed.

4.2 An adaptive, automatic multiple-case deletion technique for detecting influence in regression

To overcome the problem of masking and swamping in the “Jackknife-after-bootstrap regression influence diagnostics” article of Martin and Roberts (2010), Roberts et al. (2015) proposed a new adapted method based on an influential measure statistic that uses multiple-case deletion techniques to find influential observations (called ‘ADAP’). Roberts et al. (2015) notes that single-case deletion methods can fail because if there are multiple influential observations masking one observation, then that observation could not be unmasked by only deleting a single observation at a time. This new multiple-case deletion method is based on previous work (see, e.g., Martin and Roberts, 2010; Nurunnabi et al., 2011, 2014) and consists of an algorithm which is reasonably quick and easy to calculate, and is capable of finding an unknown number of influential observations even in cases where masking is present in the data.

The idea of multiple-case deletion to reduce effects of masking and swamping, as attempted by the ADAP algorithm, is not a new idea. It was first suggested by Belsley et al. (1980) and then later other similar methods were developed (see for example, Rousseeuw and Van Zomeren (1990), Hadi and Simonoff (1993), Peña and Yohai (1999), Pan et al. (2000), Becker and Gather (2001), Peña (2005), Rahmatullah Imon (2005) and Nurunnabi et al. (2014)). Therefore, many multiple-deletion statistics already exist in the literature, but

are often hampered by extremely long computational times, since the number of combinations in which $j > 1$ observations can be drawn from a sample of size n can be prohibitively large. For example, if a subset of $j = 10$ observations needs to be drawn from a sample of $n = 50$ observations, the number of subsets that can be created is 10 billion whereas, if only one observation needs to be drawn from the 50 observations, only 50 subsets would be created. Furthermore, to make things even worse, it is often unknown how many observations must be removed from the data set in order to accommodate settings like masking. The method by Roberts et al. (2015) overcomes this problem by proposing a two-stage approach: the first stage performs an initial cursory check for influential values and the second stage evaluates each of these values identified in the first stage to confirm the validity of their influential status.

The method involves model fitting on subsets of the whole data set. Roberts et al. (2015) hope that in some of these subsets influential values that mask other observations would be removed from those subsets, giving the masked observations the chance to reveal themselves as influential. Furthermore, this method uses the maximum of a “Cook’s distance”-like statistic calculated in the subsets and so it only needs a small number of all possible subsets to be created from a specific setting of the sample size. For this reason, the method is potentially much faster than other multiple-case deletion methods found in the literature.

The model to be fit for the method proposed by Roberts et al. (2015) is, once again, the linear regression model as defined in (4.1). After fitting the model using, say, ordinary least squares, subsets are generated by deleting $j < n$ observations from the original data set such that the subset consists of $n - j$ observations. Denote the set of indices associated with the observations that were removed by $\mathcal{I} = \{i_1, i_2, \dots, i_j\}$. For each of these subsets, the design matrix and fitted coefficients are also determined and are denoted by $\mathbf{X}_{\{\mathcal{I}\}}$ and $\hat{\beta}_{\{\mathcal{I}\}}$, respectively. Lastly, if an additional observation K is further deleted from the subset consisting of $n - j$ observations, denote the corresponding fitted coefficients by $\hat{\beta}_{\{\mathcal{I}, K\}}$.

With some of the notation now defined, we now describe the two-stages of the ADAP procedure proposed by Roberts et al. (2015), namely the *initial check stage* and the *confirmation stage*.

4.2.1 Initial check stage of ADAP

We start by defining a modified Cook’s distance that can be used to investigate the influence of the K^{th} observation on a specific subset. This “Cook’s distance”-like statistic is given by:

$$C_K^{\mathcal{I}} = \frac{(\widehat{\beta}_{\{\mathcal{I}, K\}} - \hat{\beta}_{\{\mathcal{I}\}})^{\top} \mathbf{X}^{\top} \mathbf{X} (\hat{\beta}_{\{\mathcal{I}, K\}} - \hat{\beta}_{\{\mathcal{I}\}})}{ps^2}, \quad K \in \{1, 2, \dots, n\} \setminus \{\mathcal{I}\}, \quad (4.2)$$

where s^2 estimates σ^2 and is calculated by fitting the linear regression model on the whole data set consisting of n observations. This statistic is almost the same as the usual Cook’s distance statistic when single-case deletion is performed, but Roberts et al. (2015) choose to use the term $(\mathbf{X}^{\top} \mathbf{X})/(ps^2)$ as the scaling element instead of $(\mathbf{X}_{\{\mathcal{I}\}}^{\top} \mathbf{X}_{\{\mathcal{I}\}})/(ps_{\{\mathcal{I}\}}^2)$. The

reason for choosing this specific scaling element is to ensure that it is common to all subsets and to facilitate the comparison of the statistic values across the different subsets.

The next step is then to calculate the statistic given in (4.2) for all $n - j$ observations in the subset generated. This whole procedure of deleting j observations from the original data set and calculating the $n - j$ values of $C_K^{\mathcal{I}}$ must now be repeated M times, where M is some reasonably large value, typically 1000 or larger. In addition, let $M_K < M$ denote the number of times that the K^{th} observation occurs in all M subsets generated in this way, meaning that the value $C_K^{\mathcal{I}}$ is replicated M_K times; denote these values by $C_{K,1}^{\mathcal{I}}, C_{K,2}^{\mathcal{I}}, \dots, C_{K,M_K}^{\mathcal{I}}$ for $K = 1, \dots, n$. These M_K values of $C_K^{\mathcal{I}}$ therefore measure the influence of observation K without the influence of the j deleted observations. Effects like masking might be reduced here since it is possible that the j observations deleted from the original sample may be the values contributing to the masking of observation K . Therefore, in these subsets, observation K would be free from the masking effect of those observations and will thus have a large $C_K^{\mathcal{I}}$ value in that subset indicating the possibility that observation K is an influential observation.

The ADAP procedure then makes use of the maximum of the $M_K C_K^{\mathcal{I}}$ values to determine if the K^{th} observation is influential even when it is masked by other observations. The maximum has a great advantage, because only the largest of the $C_K^{\mathcal{I}}$ values (i.e., the one subset where a masked influential observation is potentially unmasked) would be necessary for identifying a possible influential observation. Unfortunately, the maximum can also have a big drawback in this calculation. Namely, it can incorrectly identify normal observations as influential when for example, the cut-off value is too conservative. Another case where this might happen is when the specific subset's structure is very different from the original data set and the j deleted observations cause some non-influential observation to appear to be influential. This problem is demonstrated in Figure 4.1 (discussed below).

Note that Figure 4.1 was created by first plotting 50 observations generated from the model: $Y = 2X + \epsilon$, and then adding two influential observations to the plot, marked by triangles. Now, suppose $j = 10$ (i.e., 10 observations need to be deleted from the original data set to form a subset), and further suppose that the first subset is generated by deleting the $j = 10$ observations represented by the filled circles in Figure 4.1. In this case, the observation associated with the black square in Figure 4.1, would incorrectly be flagged as influential as a result of the influence exerted by the two influential observations (i.e., the triangles in Figure 4.1). According to Lawrence (1995), this event is called *boosting*. To make sure that the ADAP algorithm does not incorrectly flag some observations as influential, Roberts et al. (2015) added a confirmation stage to the ADAP algorithm. The confirmation stage, discussed next, will therefore reduce effects like boosting and swamping.

4.2.2 Confirmation stage of ADAP

In the first stage of the ADAP procedure certain observations were identified as being influential observations and the remaining observations were identified as being ‘normal’.

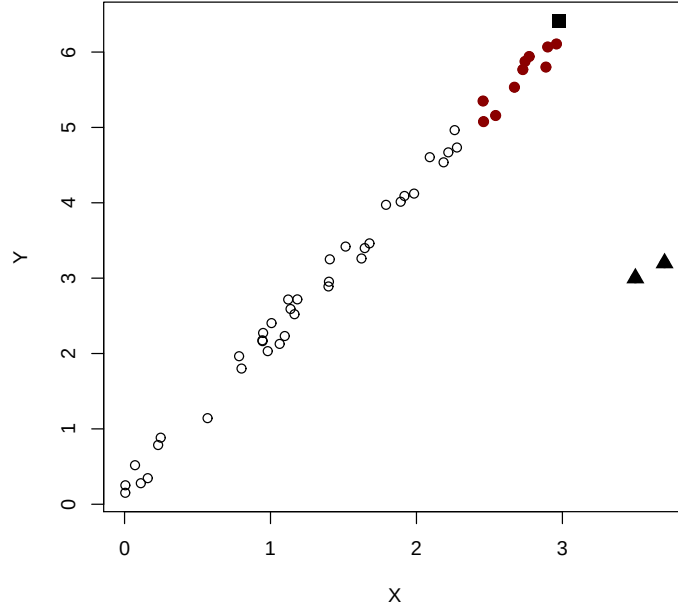


Figure 4.1: An example illustrating why the ADAP algorithm needs the ‘confirmation stage’.

The confirmation stage then uses this classification and splits the data into two separate sets, one containing the ‘normal’ observations (with index set denoted by $\mathcal{C}$) and the other one containing possible influential observations (with index set denoted by $\mathcal{S}$, and defined as $\mathcal{S} = \{K : \max(C_{K,1}^{\mathcal{I}}, C_{K,2}^{\mathcal{I}}, \dots, C_{K,M_K}^{\mathcal{I}}) > 1\}$). These two sets are defined as the *clean set* and the *suspicious set*, respectively. Now, the next step of the confirmation stage is to evaluate each of the observations in the suspicious set to confirm the certainty of their influential status. This evaluation is done by adding every observation included in the suspicious data set individually back into the clean set, one at a time, producing a new subset for each element of $\mathcal{S}$. The model is then refitted on this newly created clean set (contaminated by a single suspicious observation) and the coefficients are calculated. If the coefficients from this contaminated subset only differ a little from the coefficients obtained before inserting one of the suspicious observations, the assumption can be made that this specific observation is not really influential. A new “Cook’s distance”-like statistic is used for this situation, namely the “add-back” Cook’s distance defined as:

$$C_K^+ = \frac{(\hat{\beta}_{+K,\mathcal{C}} - \hat{\beta}_{\mathcal{C}})^{\top} \mathbf{X}_{\mathcal{C}}^{\top} \mathbf{X}_{\mathcal{C}} (\hat{\beta}_{+K,\mathcal{C}} - \hat{\beta}_{\mathcal{C}})}{ps_{\mathcal{C}}^2}, \quad K \in \mathcal{S}, \quad (4.3)$$

where the clean set’s design matrix, estimated error variance and fitted coefficients are denoted by $\mathbf{X}_{\mathcal{C}}$, $s_{\mathcal{C}}$ and $\hat{\beta}_{\mathcal{C}}$, respectively and the fitted coefficients calculated after inserting observation K back into the clean set, is denoted by $\hat{\beta}_{+K,\mathcal{C}}$.

The final step is therefore to compare the values of C_K^+ , $K \in \mathcal{S}$, to the cut-off value (typically chosen to be 1 in this algorithm). If the value is larger than the cut-off, then that provides confirmation that the observation is influential, whereas if the value is not greater than the cut-off, then one might suspect that it is not actually influential, but was identified as influential in the first stage of the process due to, for example, boosting.

The two stages of the ADAP algorithm will now be provided in the next subsection, as well as some additional information on the ADAP method.

4.2.3 Algorithm of the ADAP method

Roberts et al. (2015) summarised this two-stage ADAP method in the following algorithm:

Algorithm 6 *Implementation of the ADAP method for influential value detection*

Initial check stage:

1. From the given data $T_1 = (\mathbf{X}_1^\top, Y_1), T_2 = (\mathbf{X}_2^\top, Y_2), \dots, T_n = (\mathbf{X}_n^\top, Y_n)$, sample $n - j$ observations without replacement to obtain a subset with j observations deleted of the form, $\{T_h : h \in \{1, 2, \dots, n\} \setminus \mathcal{I}\}$, and where $\mathcal{I}$ denotes the set of indices corresponding to the deleted observations, i.e., $\mathcal{I} = \{i_1, i_2, \dots, i_j\}$.
 2. Use the subset generated in the first step to fit the linear regression model using least squares and then calculate, for all $n - j$ observations included in the subset, their corresponding $C_h^\mathcal{I}$ values, $h \in \{1, 2, \dots, n\} \setminus \mathcal{I}$.
 3. Repeat these first two steps M times to obtain M_K values of $C_K^\mathcal{I}$ for every observation $K = 1, \dots, n$. Denote these values by $C_{K,1}^\mathcal{I}, C_{K,2}^\mathcal{I}, \dots, C_{K,M_K}^\mathcal{I}$ for $K = 1, \dots, n$.
 4. For each $K = 1, 2, \dots, n$, determine the maximum of the M_K values of $C_K^\mathcal{I}$ obtained in step 3. These maximum values must now be compared with an appropriate cut-off value for Cook's distance. This cut-off value is chosen to be equal to 1 (for simplicity reasons) by Roberts et al. (2015). Every observation corresponding to a maximum $C_K^\mathcal{I}$ value exceeding one will then be identified as a possible influential observation.
-

Remark 4.3. *Note that even though maximum values of the “Cook’s distance”-like statistic are calculated in the algorithm, a cut-off value of 1 is used in the algorithm. The choice to use 1 as the cut-off value is traditionally used for the Cook’s distances for large samples, but is not a cut-off value for the maximum values of Cook’s distance. A possible better choice would be to find cut-off values based on the maximum Cook’s distance value, rather than cut-off values based on the ordinary values of Cook’s distance. This approach will be considered in the next chapter.*

Remark 4.4. *To further motivate the use of this part of the algorithm, suppose that observation K is an influential observation. If observation K is the only influential observation in the data set, it would influence all the subsets which include observation K , resulting in big values of $C_K^\mathcal{I}$. If, on the other hand, there are some other influential values in the data set too, they might mask observation K . For some of the subsets it might be the case that observation K would not be seen as influential, leading to small $C_K^\mathcal{I}$ values for observation K in those subsets. In other subsets, it might happen (and Roberts et al. (2015) hopes that it would happen) that those influential observations masking observation K are removed from the subset. In these cases observation K would reveal itself as influential, leading to large values of $C_K^\mathcal{I}$.*

Algorithm 6 *Continued ...***Confirmation stage:**

5. Using the information obtained from the fourth step of the ADAP algorithm, split the original data set into a clean set (consisting of all observations with maximum C_K^T value smaller than 1) and suspicious set (consisting of all observations with maximum C_K^T value exceeding one).
6. Using both the clean and suspicious data sets, calculate the values of C_K^+ for all observations included in the suspicious set, i.e., for all $K \in \mathcal{S}$.
7. For every observation $K \in \mathcal{S}$, compare its corresponding C_K^+ value to the cut-off value of 1. If that C_K^+ value is greater than 1, it can be confirmed that observation K is a suspicious observation. On the other hand, if that C_K^+ value is smaller than 1, observation K is defined as a normal observation and must be added back into the clean set.

Remark 4.5. *For further improvement on the ADAP algorithm, these last three steps (the confirmation stage) can be repeated until no further changes occur in the clean and suspicious sets, i.e., no observations are removed from the suspicious set and added to the clean set. However, this additional modification will not be considered further in this text.*

Two very important elements of the ADAP algorithm that need to be specified are M , the number of iterations needed in the algorithm (also the number of subsets that would be generated) and j , the number of observations that need to be deleted to form the M subsets. So, an important matter for the ADAP algorithm is to obtain suitable sizes of these two elements. In order to investigate this matter, Roberts et al. (2015) first made the assumption that δ , the number of inserted influential observations, is known. The aim here is to find the best possible values for M and j , such that all the δ influential observations would be isolated from the rest of the influential observations in at least one of the subsets. In this way, if one or more of the δ influential observations are masked, they would be free from the influence of the observations masking them. In those few subsets then, the masked influential observations would have large C_K^T values, and would ultimately be flagged as being influential. Roberts et al. (2015) stated that the following Bonferroni lower bounds can be used to test the efficiency of different choices for M and j in cases where Cook's distance can identify two influential observations simultaneously:

$$P_B = \delta \times \left(1 - \left(1 - \frac{\binom{n-\delta}{j-\delta+1} + (\delta-1)\binom{n-\delta}{j-\delta+2}}{\binom{n}{j}} \right)^M \right) - \delta + 1, \quad (4.4)$$

where we define P_B as the Bonferroni lower bound for the probability that at least one subset contains at most one influential observation. Using the formula in (4.4), Roberts et al. (2015) suggested that using $j = n/2$ and $M \geq 1000$ is enough to ensure that the probability that the procedure will produce at least one subset that contains at most one influential observation is close to one. This relationship was explored for as many as $\delta = 10$ influential values in the data set and $n \geq 100$, but still holds up well for $n \geq 30$.

Just like the method proposed by Martin and Roberts (2010), the ADAP algorithm has the disadvantage of being more time consuming than traditional methods. In the study of

Roberts et al. (2015), the ADAP algorithm was performed on a data set with $n = 100$ and $p = 10$. For these settings the ADAP algorithm ran roughly 11 seconds. In a second study, an enlargement was made on the data set and more predictor variables were included such that $n = 1000$ and $p = 50$. In this case it took the ADAP algorithm roughly 3.5 minutes to run. The time to execute is therefore relatively acceptable, but if speedup improvements are necessary, methods like parallel processing computation can be implemented for the ADAP algorithm.

Note that the ADAP method has a few extensions, such as the one by Wang et al. (2019), where it was extended to deal with masking and swamping in high-dimensional regression. In the next chapter we present an additional extension to the ADAP method that was briefly proposed (but never investigated) in Roberts et al. (2015, see remark 1, p. 410), where they suggest using the JaB cut-off values developed by Martin and Roberts (2010) in the ADAP algorithm, instead of the cut-off value of 1 as suggested above.

In the next section, we will now briefly discuss a Bayesian method for identifying influential observations, called the MC^3 method. This method is introduced here such that, in the simulation study that will follow in Chapter 6, its performance of identifying influential observations can be compared to the performances of the other methods in this chapter and in the chapter that will follow.

4.3 MC^3 Method

In the literature, there are many ways to perform variable selection and to identify influential observations, however, most procedures focus on these two aspects separately. For the method briefly discussed in this section, found in Hoeting et al. (1996), these two aspects are handled simultaneously.

Hoeting et al. (1996) suggest a Bayesian approach to simultaneously performing model selection and influential case identification by making use of Bayesian model averaging (BMA). The Bayesian approach is based on averaging over all possible models, instead of only looking at the single best model out of all the possible models. According to Madigan et al. (1995), this approach will give very good predictive results, but has the drawback of having to compute posterior probabilities for all possible models (i.e., all combinations of predictor variables and influential observations) to be averaged over, and therefore, will not be practical when there are a (very) large set of possible models. For cases like these, Hoeting et al. (1996) suggested that the Markov chain Monte Carlo (MCMC) model composition (MC^3) method (Madigan et al., 1995) be applied in order to obtain an approximation to the complete solution of the Bayesian averaging approach, i.e., an approximation to the Bayesian model average solution calculated over the space of all possible models. Specifically, the MC^3 method involves a MCMC to approximate the posterior distribution of some quantity under investigation, in this case, the Bayesian model average of all possible models. Madigan et al. (1995) was the first to propose the MC^3 method and later, Raftery et al. (1997) altered it for the use in linear regression models.

This method has a number of advantages as it has been shown that it is generally able to avoid the problems associated with masking of influential observations by employing data screening, where an initial robust technique is first used to obtain a large set of potential outliers. The method then uses this set of possible influential observations and applies the Markov chain on it to obtain the set of posterior probabilities of variables being selected into the model, as well as the posterior probabilities of cases being identified as influential observations. These posterior probabilities will respectively give an indication on which set of variables to include in the regression model, and which observations might possibly be influential. To implement this method in the simulation study that will be provided in Chapter 6, we will be making use of the function MC^3 .REG in the R package BMA, as given in Raftery et al. (2019).

Note that this method is only introduced as a method to be used to compare the performance of the other methods discussed in this dissertation. For further theoretical details on MC^3 , the reader is referred to Madigan et al. (1995), Hoeting et al. (1996), and Raftery et al. (1997).

In the next chapter, two new influential diagnostic measures will be presented.

Chapter 5

Newly proposed influential diagnostic methods

Two new influential diagnostic methods will be discussed in this chapter. The first method is based on the combination of the JaB method and the ADAP method, and the second method involves a simple way to approximate the distribution of Cook's distance in order to obtain cut-off values which can be used to identify influential observations in a data set. The methodology behind these two new methods will now be given in detail.

5.1 Using jackknife-after-bootstrap cut-off values in the ADAP method (the “Combined” method)

The JaB and the ADAP methods that were explained in the previous chapter are both excellent methods for detecting influential observations, but both of them have aspects that can be improved. The results obtained from the JaB method, for instance, might be heavily influenced if masking and swamping occurs in the data. The ADAP method, on the other hand, naïvely uses the traditional large sample cut-off values to determine if the maximum values of the “Cook's distance”-like statistics denote influential observations. This can be improved by finding cut-off values specifically for the maximum of the “Cook's distance”-like statistic's values instead. In this section, these two methods are combined to obtain an improved method for detecting influential observations, and the resulting method will be referred to as the “Combined” method.

In the Combined method, the ADAP method (multiple-case deletions method) proposed by Roberts et al. (2015) will be used to identify influential observation in the data, but instead of comparing the maximum values of the “Cook's distance”-like statistic's values with the traditional large sample cut-off values for Cook's distance, new cut-off values will be generated using the JaB method proposed by Martin and Roberts (2010). These new cut-off values will be obtained using JaB procedures to approximate the sampling distribution of the maximum of these values. Since the JaB method is used here, these bootstrap procedures will be based on Efron's (1992) JaB sampling lemma (see Lemma 1). To implement this lemma, the bootstrap procedures will be based on the null hypothesis stating

that the i^{th} observation is not influential. The approximated sampling distribution would therefore be calculated using only the bootstrap samples not containing the i^{th} observation.

After the cut-off values for the maximum “Cook’s distance”-like statistic’s values are obtained, they are implemented in the initial check stage of the ADAP algorithm to detect possible influential values. Lastly, the initial check stage will be followed by a confirmation stage to confirm which observations are influential. Another improvement on the ADAP algorithm was made here, since instead of using the traditional large sample cut-off values of Cook’s distance for the confirmation stage, the JaB method was implemented again to calculate cut-off values based on the data itself. Note that the cut-off values used in the initial check stage of the ADAP algorithm are not suitable for the confirmation stage, since they are based on the maximum values of the “Cook’s distance”- like statistic’s values; for the confirmation stage, cut-off values based on the ordinary Cook’s distance values are needed instead.

The new improved method can be summarised in the following algorithm:

Algorithm 7 *Implementations of the Combined method for influential value detection*

Calculating cut-off values for initial check stage:

1. Given the data $T_1 = (\mathbf{X}_1^\top, Y_1), T_2 = (\mathbf{X}_2^\top, Y_2), \dots, T_n = (\mathbf{X}_n^\top, Y_n)$, independently sample cases from this original data set to generate B resamples of the form $T_1^*, \dots, T_n^*$.
2. In the B resamples generated in step 1, some of these resamples do not contain the i^{th} data point, T_i . Let b_i denote the number of the B resamples that do not contain the i^{th} data point. For each of these b_i resamples, do the following:
 - 2.1. Randomly sample $n - j$ observations without replacement to obtain a subset with j observations deleted of the form, $\{T_h^* : h \in \{1, 2, \dots, n\} \setminus \mathcal{I}^*\}$, and where $\mathcal{I}^*$ denotes the set of indices corresponding to the deleted observations, i.e., $\mathcal{I}^* = \{i_1, i_2, \dots, i_j\}$.
 - 2.2. Use the subset generated in step 2.1 to fit the linear regression model using least squares and then calculate, for all $n - j$ observations included in the subset, their corresponding $C_h^{\mathcal{I}^*}$ values, $h \in \{1, 2, \dots, n\} \setminus \mathcal{I}^*$.
 - 2.3. Repeat steps 2.1 and 2.2 M times. Since not all subsets contain the bootstrap observation T_k^* , we only obtain $M_k < M$ subsets with this value. Denote the Cook’s distance values that correspond to T_k^* in each of the M_k subsets by $\tilde{C}_{k,1}, \tilde{C}_{k,2}, \dots, \tilde{C}_{k,M_k}$.
 - 2.4. Calculate the maximum value of the Cook’s distances for each $k = 1, 2, \dots, n$, that is, $\tilde{\mathbf{A}}^{max} = [\tilde{A}_1^{max}, \tilde{A}_2^{max}, \dots, \tilde{A}_n^{max}]$ where

$$\tilde{A}_k^{max} = \max \left\{ \tilde{C}_{k,1}, \tilde{C}_{k,2}, \dots, \tilde{C}_{k,M_k} \right\}.$$

At the end of steps 2.1 to 2.4, we have b_i sets of $\tilde{\mathbf{A}}^{max}$, i.e., $\tilde{\mathbf{A}}_1^{max}, \tilde{\mathbf{A}}_2^{max}, \dots, \tilde{\mathbf{A}}_{b_i}^{max}$.

P.T.O.

Algorithm 7 *Continued ...***Calculating cut-off values for initial check stage continued...**

3. To obtain an approximate distribution for the maximum of the Cook's distance for the i^{th} case, combine the nb_i maximum values that were obtained by the b_i resamples that do not contain the i^{th} case, into one vector called $\tilde{\mathbf{A}}_i$, i.e., the resulting vector will have nb_i elements. Repeat this step for each $i = 1, \dots, n$, i.e.,

$$\begin{aligned}\tilde{\mathbf{A}}_1 &= [\tilde{\mathbf{A}}_1^{max}, \tilde{\mathbf{A}}_2^{max}, \dots, \tilde{\mathbf{A}}_{b_1}^{max}] \\ \tilde{\mathbf{A}}_2 &= [\tilde{\mathbf{A}}_1^{max}, \tilde{\mathbf{A}}_2^{max}, \dots, \tilde{\mathbf{A}}_{b_2}^{max}] \\ &\vdots \\ \tilde{\mathbf{A}}_n &= [\tilde{\mathbf{A}}_1^{max}, \tilde{\mathbf{A}}_2^{max}, \dots, \tilde{\mathbf{A}}_{b_n}^{max}]\end{aligned}$$

4. For each $i = 1, 2, \dots, n$, use the vector/approximated distribution calculated in step 3 and calculate the cut-off value by finding the ω^{th} quantile, where $0 \leq \omega \leq 1$ is some appropriately chosen proportion value used for the cut-off (in our case, ω was set to 0.05). Denote this set of cut-off values by CO_i^{ICS} , $i = 1, 2, \dots, n$.

Initial check stage (ICS):

5. From the given data $T_1 = (\mathbf{X}_1^\top, Y_1), T_2 = (\mathbf{X}_2^\top, Y_2), \dots, T_n = (\mathbf{X}_n^\top, Y_n)$, sample $n - j$ observations without replacement to obtain a subset with j observations deleted of the form, $\{T_h : h \in \{1, 2, \dots, n\} \setminus \mathcal{I}\}$, and where $\mathcal{I}$ denotes the set of indices corresponding to the deleted observations, i.e., $\mathcal{I} = \{i_1, i_2, \dots, i_j\}$.
6. Use the subset generated in step 5 to fit the linear regression model using least squares and then calculate, for all $n - j$ observations included in the subset, their corresponding $C_h^{\mathcal{I}}$ values, $h \in \{1, 2, \dots, n\} \setminus \mathcal{I}$.
7. Repeat steps 5 and 6 M times.
Since not all subsets contain the observation T_K , we only obtain $M_K < M$ subsets with this value. Denote the Cook's distance values that corresponds to T_K in each of the M_K subsets by $C_{K,1}, C_{K,2}, \dots, C_{K,M_K}$.
8. For each $K = 1, 2, \dots, n$, determine the maximum of the M_K values of $C_K^{\mathcal{I}}$ obtained in step 3, i.e., calculate $A_K^{max} = \max \{C_{K,1}, C_{K,2}, \dots, C_{K,M_K}\}$. Compare each of these maximum values with their corresponding cut-off value CO_K^{ICS} , $K = 1, 2, \dots, n$. Every observation corresponding to a maximum Cook's distance value exceeding its cut-off value will then be identified as a possible influential observation.

Calculating cut-off values for confirmation stage:

10. Given the data $T_1 = (\mathbf{X}_1^\top, Y_1), T_2 = (\mathbf{X}_2^\top, Y_2), \dots, T_n = (\mathbf{X}_n^\top, Y_n)$, independently sample cases from the original data set to generate B resamples of the form $T_1^*, \dots, T_n^*$.
11. Again, let b_i denote the number of the B resamples that do not contain the i^{th} data point. For each of these b_i resamples, calculate the n ordinary Cook's distance values denoted by C_i^* , $i = 1, 2, \dots, n$. To obtain an approximate distribution for the i^{th} Cook's distance, combine the nb_i Cook's distance values that were obtained by the b_i resamples that do not contain the i^{th} case, into one vector, i.e., the resulting vector will have nb_i elements. Repeat this step for each $i = 1, \dots, n$.
12. For each $i = 1, \dots, n$, use the vector/approximated distribution calculated in step 11 and calculate the cut-off value by finding the ω^{th} quantile, where $0 \leq \omega \leq 1$ is some appropriately chosen proportion value used for the cut-off (in this dissertation, ω was set to 0.05). Denote this set of cut-off values by CO_i^{CS} , $i = 1, 2, \dots, n$.

Algorithm 7 *Continued ...*

Confirmation stage: (CS)

13. Using the information obtained from step 8 of the algorithm, split the original data set into a ‘clean’ set (consisting of all observations with maximum Cook’s distance value smaller than its cut-off value) and ‘suspicious’ set (consisting of all observations with maximum Cook’s distance value exceeding its cut-off value).
 14. Using both the clean and suspicious data sets, calculate the values of C_K^+ (see equation (4.3)) for all $K \in \mathcal{S}$.
 15. For every observation $K \in \mathcal{S}$, compare its C_K^+ value to its corresponding cut-off in the set CO_K^{CS} , $K = 1, 2, \dots, n$. If that C_K^+ value is greater than its cut-off value, it can be confirmed that observation K is influential. On the other hand, if that C_K^+ value is smaller than its cut-off value, observation K is defined as a normal observation and must be added back into the clean set.
-

Remark 5.1. *As in the ADAP algorithm, this new algorithm can be improved by repeating the confirmation stage until no further changes occur in the clean and suspicious sets, i.e., no observations are removed from the suspicious set and added to the clean set.*

One big disadvantage of the combined method is that it is incredibly time consuming. For a data set with $n = 500$, and $p = 15$ for example, it took approximately 57.80 minutes to run, using the computer program R on an Intel Core™ i7-6700 PC with 16GB of RAM. This time was obtained even after some modifications were made to the code in order to improve the execution time of this method. For a relatively small data set with $n = 100$ and just with a few predictor values, $p = 5$, it still took the combined method approximately 3.38 minutes to execute. The reason why the combined method is so time consuming is, first of all, because it consists of four stages: The two stages where cut-off values are generated using the JaB algorithm, and then the two stages of the ADAP algorithm. Furthermore, for the first set of cut-off values ($\mathbf{CO}^{ICS}$), 1000 iterations for each of the 300 bootstrap resamples have to be generated while doing the JaB procedure, which takes a long time to complete. Lastly, since the ADAP procedure were used as the method for detecting influential observations, and not the traditional influential measures, the combined method is automatically longer than would be the case of traditional influential measures.

As a result of this drawback of the combined method, even if it performs very well and gives accurate results, it would not always be the best choice to use for detecting influential observations in a data set, since this method is so computationally expensive. A simpler and quicker method for approximating the distribution of Cook’s distance will be proposed in the next section. This distribution will then further be used to calculate cut-off values for Cook’s distance, which can be used to detect influential observations.

5.2 A new method based on a simple approximation of the distribution of Cook's distance (the “New” method)

As already mentioned in the literature study, some rules of thumb are used for cut-off values for Cook's Distance. These rules of thumb are not always very accurate and might miss some influential observations in a data set or flag normal observations as influential. In this section, a new method is considered for determining the cut-off values by considering a simple approximation of the distribution of the Cook's distance values. This method will be referred to as the “New” method going forward.

For the New method, the alternate form of Cook's distance given in (2.18) will be used which is defined in terms of leverages. This Cook's Distance formula is

$$\text{Cook's Distance} = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right), i = 1, 2, \dots, n, \quad (5.1)$$

where r_i are the standardized residuals (as defined in (2.12)), p is the number of coefficient parameters in the linear regression model, and h_{ii} are the leverages.

The way in which the new rough cut-off values will be generated is by simply replacing r_i^2 and h_{ii} in expression (5.1) with the $(1 - \alpha)$ quantile values for the distributions of r_i^2 and h_{ii} , respectively, under the assumption that the observations do not exhibit abnormal behaviour, i.e., are not extreme/outliers. However, according to the *Quantile equivariance theorem* Hosseini (2010), stated in Theorem 5.1 below, replacing the quantiles in this expression will only be valid if (5.1) is monotone with respect to r_i^2 and h_{ii} .

Before stating the *Quantile equivariance theorem*, first consider the following necessary new notation: Suppose the distribution function of the random variable X is denoted by F , then according to Hosseini (2010), the left quantile function for the random variable X is defined as

$$Q_X^L(p) = \inf \{x : F(x) \geq p\},$$

and the right quantile function is defined as

$$Q_X^R(p) = \inf \{x : F(x) > p\} = \sup \{x : F(x) \leq p\}.$$

Theorem 5.1 (Quantile Equivariance Theorem). *Suppose $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is non-decreasing.*

a. *If ϕ is left continuous then*

$$Q_{\phi(X)}^L(p) = \phi(Q_X^L(p)).$$

b. *If ϕ is right continuous then*

$$Q_{\phi(X)}^R(p) = \phi(Q_X^R(p)).$$

Note that the expression (5.1) is certainly monotone with respect to $r_i^2 \geq 0$, and so all that needs to be shown is that it is monotone with respect to h_{ii} . In order to show this latter monotonicity, we only need to determine if the term $\left(\frac{h_{ii}}{1-h_{ii}}\right)$ is monotone. From the graph below, it is clear that the term $\left(\frac{h_{ii}}{1-h_{ii}}\right)$ has the following monotone intervals, $(-\infty, 1)$ and $(1, \infty)$, that is, it is monotone everywhere except at the asymptote $h_{ii} = 1$.

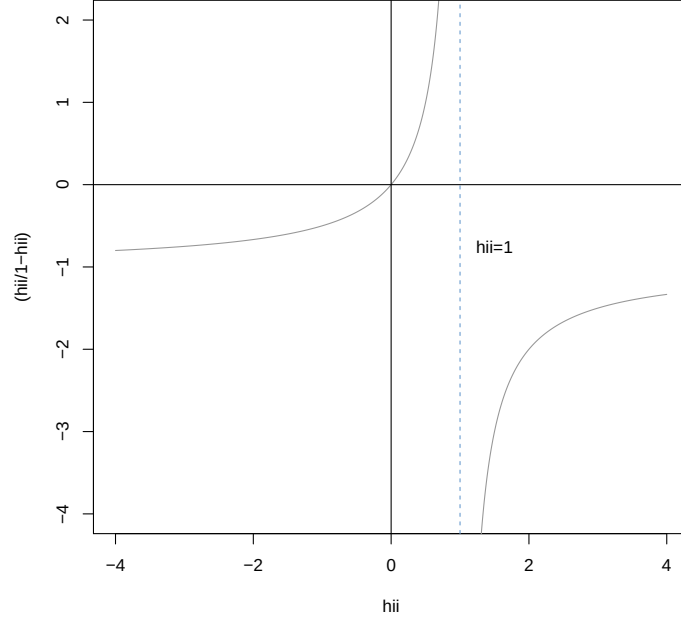


Figure 5.1: Plot of h_{ii} against the function $h_{ii}/(1-h_{ii})$ to show that the function $h_{ii}/(1-h_{ii})$ is monotone everywhere except at $h_{ii} = 1$.

To prove this formally, consider the following function

$$f(x) = \frac{x}{1-x}. \quad (5.2)$$

For (5.2) to be monotone, its derivative should be greater than zero for all x . The derivative is given by

$$f'(x) = \frac{1}{(1-x)^2} > 0, \quad \forall x \neq 1.$$

The term $\left(\frac{h_{ii}}{1-h_{ii}}\right)$ is therefore monotone everywhere except at $h_{ii} = 1$ (as was also seen in Figure 5.1).

Further, it is known that $0 \leq h_{ii} \leq 1$ (see Appendix B.1), so (5.1) would be monotone for $0 \leq h_{ii} < 1$ for all values of $r_i^2 \geq 0$. We will therefore focus attention only on this interval of the h_{ii} values since this is the only interval which will allow one to employ the idea of using the $(1-\alpha)$ quantile values of the r_i^2 and h_{ii} statistics.

Approximating the quantiles of r_i^2

To obtain an approximation of the $(1-\alpha)$ quantile of the distribution of r_i^2 , we start by first considering the distribution of r_i . According to Deakin and Hunter (2018), standardized

residuals of the form

$$w_i = \frac{v_i}{\hat{\sigma}_v},$$

where v_i is some simple residual and $\hat{\sigma}_v$ is the least squares estimate of the standard deviation of v_i , follow a τ -distribution. Note that $r_i = e_i / \sqrt{MSE(1 - h_{ii})}$ takes on this exact form and therefore will also follow a τ -distribution.

Further, it can be shown that the τ -distribution and the t -distribution are related via the following simple expression

$$X = T \sqrt{\frac{\nu - 1}{\nu + T^2}} \sim t_{\nu-1},$$

with $T \sim \tau_\nu$, see, for example, Thompson (1935) and Chu (1956). It is therefore easy to use t -distributed random variables to construct random values from the τ -distribution using this relationship, that is,

$$T = \frac{X\sqrt{\nu}}{\sqrt{\nu - 1 + X^2}} \sim \tau_\nu, \quad (5.3)$$

where $X \sim t_{\nu-1}$ (Pope, 1976; Deakin and Hunter, 2018). Lastly, Pope (1976) stated that, when using r_i , i.e., standardized residuals from linear regression models, the degrees of freedom must be chosen as $\nu = n - p$, where p is the number of parameters in the linear regression model.

In the above discussion we equated the distribution of r_i to the distribution of T , a τ -distributed random variable, and have also shown that this random variable can be expressed in terms of t -distributed random variables. We can now extend this result to also express the distribution of r_i^2 in terms of t -distributed random variables, which will be done by simply noting that the distribution of r_i^2 is identical to the distribution of T^2 . Furthermore, note that the random variable T^2 can be expressed in terms of t -distributed random variables by taking the square of both sides of the expression in (5.3):

$$T^2 = \frac{X^2\nu}{\nu - 1 + X^2} \sim \tau_\nu^2, \quad (5.4)$$

where $X \sim t_{\nu-1}$.

Finally, a method that can be used to approximate the quantiles of the distribution of T^2 is to make use of simple Monte Carlo simulations. This involves generating a large number random variates, say MC , from a $t_{\nu-1}$ distribution, applying the transformation in (5.4) to convert these values to the τ^2 -distribution, and then finding the $(1 - \alpha)$ sample quantile of these MC transformed data values. We will use $\tau_{\nu, MC}(1 - \alpha)$ to denote the resulting $(1 - \alpha)$ approximate quantile obtained from MC Monte Carlo simulated values.

Approximating the quantiles of h_{ii}

We start the derivation to find the approximate quantile of h_{ii} by first defining D_i^2 , the Mahalanobis distance between the i^{th} predictor vector $\mathbf{X}_i$ and the centroid of those predictor variables, $\bar{\mathbf{X}}$:

$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}),$$

where $\mathbf{X}_i$ is a $(p \times 1)$ column vector of the design matrix in a linear regression model with X_{ij} denoting the j^{th} element in this vector, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, $\bar{\mathbf{X}}$ is the $(p \times 1)$ mean vector with elements $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$, $j = 1, 2, \dots, p$, and $\mathbf{S}$ is the $(p \times p)$ sample covariance matrix with $S_{jk} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$ denoting the element in the j^{th} row and k^{th} column, $j, k = 1, 2, \dots, p$. The relationship between the h_{ii} values and the squared Mahalanobis distance, according to Rousseeuw and Leroy (1987), is then given by:

$$h_{ii} = \frac{D_i^2}{n-1} + \frac{1}{n}. \quad (5.5)$$

Using the fact that it is possible to approximate the distribution of D_i^2 , we can now approximate the $(1-\alpha)$ quantile of the distribution of h_{ii} by making use of this well-known relationship. According to Friendly et al. (2013), the squared Mahalanobis distance has a large-sample χ_p^2 distribution in the case where we have multivariate normal variables and so, if this condition is met, one can approximate the $(1-\alpha)$ quantile of D_i^2 using the quantile $\chi_p^2(1-\alpha)$. Then, noting that the expression in (5.5) is a monotone, linear function of D_i^2 , we can once again use Theorem 5.1 and simply substitute D_i^2 with $\chi_p^2(1-\alpha)$ in (5.5) to obtain the $(1-\alpha)$ quantile for h_{ii} . This quantile will be denoted by $h_{ii}(1-\alpha)$.

Combining the approximate quantiles to obtain the final cut-off

The proposed cut-off value based on these quantiles is then defined as,

$$C_\alpha = \frac{\tau_{\nu, MC}(1-\alpha)}{p} \left(\frac{h_{ii}(1-\alpha)}{1 - h_{ii}(1-\alpha)} \right). \quad (5.6)$$

In conclusion, the cut-off value given in (5.6) will be a useful cut-off value for determining the influence of the Cook's distance values. However, note that the calculation of this cut-off value must satisfy the following conditions:

- The predictor variables are assumed to come from a multivariate normal distribution.
- The sample size, n , must have the following properties:

- $n \neq 0$ and $n \neq 1$,
- $n > \frac{(D^2+2)+\sqrt{D^2(D^2+4)}}{2}$, $D^2 > 0$ (note that this result is proved in Appendix F).

In the chapter that follows many of the methods for detecting influential values discussed in this dissertation will be investigated in an extensive simulation study.

Chapter 6

Simulation study

6.1 Introduction

In this chapter all the methods discussed in the previous chapters, which includes the MC^3 method and the traditional Cook's distance, will be investigated in an extensive simulation study using simulated data sets with various configurations of simulated influential and outlying values. In particular, the performance of the improved method — based on the combination of the JaB and the ADAP methods (see Section 5.1), called the Combined method — as well as our new method (see Section 5.2), called the New method, are considered and compared to the performances of the following four more traditional methods: Cook's traditional influential diagnostic measure (see Section 2.3.5), the plain JaB method (see Section 4.1), the ADAP method (see Section 4.2), and the MC^3 method (see Section 4.3). Their performances will then be compared to one another in order to determine which method has the best overall performance for detecting the simulated influential values in the various configurations. The results, collected and summarized in tables provided at the end of the chapter, will then be discussed in great detail, and comments and recommendations will be given. Conclusions regarding all of the methods and some statements on which methods performed the best are relegated to the final chapter.

The method that will be followed in this simulation study will be to generate data from a linear regression model, as given in (4.1), using various choices for the distribution of the error term and then, after the data has been generated, alter some of the observations to make them influential. We will also consider two broad approaches for introducing influential values into these simulated data sets.

1. The first approach will be called the *simple* approach where the data are generated in such a way that simple single-case deletion techniques (such as the traditional Cook's distance using the standard cut-off values) will be able to detect these influential observations, or at least some of them.
2. The second approach, called the *masked* approach, will attempt to generate the influential values such that they are *masked* from one another when they are *all* present, but are easily detected if only one of them features in the data set.

In both cases, a total of δ influential values are added into the data sets. In the simple

approach, the manner in which this is accomplished involves the trivial idea of simply increasing or decreasing the response variables' values of the $2\delta/3$ many selected observations with the highest leverages and the $\delta/3$ lowest leverage values by some arbitrarily large value. To obtain this increase or decrease in the response variables' values, the sign of each observation in the data sets' residual values are calculated and then also the maximum over all of these residuals. The maximum residual is then multiplied by five and added to those specific response variables with highest and lowest leverage values, in the direction of the sign of their corresponding residuals (i.e., if the sign is negative we subtract the maximum residual value and if the sign is positive, we add this value). Furthermore, we also increased and decreased these response variables' corresponding predictor variables. This was done as follows: For a response variable Y_i , we added to X_{ij} three times the maximum value over the column $\mathbf{X}_i$ in alternating directions (i.e., for X_{i1} we subtract this value and for X_{i2} we add the value). After this stage, we choose only the data sets for which Cook's distance are able to identify at least $\lfloor \delta/3 \rfloor$ of the inserted influential observations.

The masking approach, on the other hand, employs a more complex approach, since there are many checks that need to take place to ensure that the influential values are properly masked in the resulting data set. To explain how the influential values are generated in the masking approach, we start by introducing some additional concepts to help explain the process.

First note that, for an observation to be influential using the traditional approach, its corresponding Cook's distance value needs to be greater than 1, that is,

$$D_i = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right) > 1. \quad (6.1)$$

Now, from the formula of the standardized residual, r_i , given in (2.12), it is easy to see that (6.1) can be written as

$$\frac{e_i^2}{s^2 p} \frac{h_{ii}}{(1 - h_{ii})^2} > 1. \quad (6.2)$$

Then, to simplify some of the calculations that will follow, we introduce the variable K , defined as

$$K := \frac{1}{s^2 p} \frac{h_{ii}}{(1 - h_{ii})^2}.$$

The expression in (6.2) can then be written in the following simplified form:

$$e_i^2 K > 1.$$

Then, after some simple algebra, the above expression can be written as follows:

$$e_i > \sqrt{1/K} \quad \text{or} \quad e_i < -\sqrt{1/K},$$

that is,

$$|e_i| > \sqrt{1/K}.$$

Therefore, if we want to make a response value influential, we only need to ensure that the corresponding residuals exceeds $\sqrt{1/K}$ in absolute value. Denote the value that needs to be added to an observation's residual value to make it influential by A , defined as:

$$A := \sqrt{1/K} - |e_i|.$$

Note, however, that if an observation is already influential before adding the value A to its residual, then the value A becomes negative, leading to subtracting the size of A from the observation's residual. In this case, the influential observation will be changed into a non-influential observation. For this reason, we add the absolute value of A , i.e., $|A|$, to the observation's residual that needs to become influential.

Now, the way in which the δ influential observations are inserted in the masking approach, first involves determining which observations have the largest leverages. For each of these observations, their corresponding $|A|$ values are calculated as well as the sign of their residuals. Next, the corresponding $|A|$ value is added to each observations' residual in the direction of the sign of the residual (i.e., if the sign is negative we subtract the $|A|$ value and if the sign is positive, we add the $|A|$ value). Finally, only data sets that comply with the following condition are used:

Condition 1. *No observations must be flagged as being influential observations when the model is fitted on the data set which includes all the influential observations, but, these values must be flagged as influential when the model is fitted to the data set containing only one of these influential observations (i.e., a data set where all but one of the influential observations are deleted).*

Remark 6.1. *Condition 1 states two things: First is that none of the influential observations must be flagged as influential in the case where the model is fitted to the whole data set. Second, when all but one of the δ influential observations are deleted from the data set, then that remaining influential observation will definitely be flagged as influential. This does not mean, however, that this is the only way in which a simulated influential observation will be flagged as influential. It might also happen that if, for example, only two of the δ influential observations are deleted from the data set, then some of the remaining influential observations can be flagged as influential. This condition will therefore have the effect that single-case deletion techniques will struggle to detect the inserted simulated influential observations (or they might not be able to detect these observations at all).*

Summarising results and the confusion matrix: Once the data sets have been generated using either of the two methods described above, each of the six influence diagnostic methods will be applied to these data sets to attempt to detect the simulated influential values. Each of these diagnostic methods will then identify a number of cases in each data set as being influential or not. However, since we know which values were simulated to be influential, we can determine how accurately each method performs for each data set. To summarize this information, we will make use of the notation used in a *confusion matrix*, which needs to be explained first. First note the general form of a confusion matrix given in Table 6.1.

Table 6.1: General form of a confusion matrix.

	Predicted: No	Predicted: Yes
Actual: No	True negative (TN)	False positive (FP)
Actual: Yes	False negative (FN)	True positive (TP)

In Table 6.1, the columns divide the data into two groups namely, cases that the fitted regression model predicted as ‘*Not influential*’ values (i.e., ‘No’) and cases that were predicted as ‘*Influential*’ (i.e., ‘Yes’ values). The same holds for the row names, but here the two groups indicate the actual influential nature of the cases and are similarly designated as being ‘No’ (i.e., the value was not simulated to be influential) or ‘Yes’ (i.e., the value was simulated as being influential). In the tables that will be provided in the next section, we will make use of the terms TP and FP to denote the number of cases classified in the data as *True Positives* and *False Positives*, respectively. The TP value denotes the number of cases in the data set that were *correctly* identified by the diagnostic method as a ‘Yes’ value, that is, the method diagnosed the case as being influential when the observed value is in fact simulated to be influential. The term FP on the other hand, denotes the number of cases for which the method diagnosed a case as being influential (i.e., the method concludes a ‘Yes’), but the value was not simulated to be influential (i.e., the case is actually a ‘No’). The terms TN and FN denote the number of true negatives (i.e., correctly classifying a ‘No’ as a ‘No’) and the number of false negative (i.e., incorrectly classifying a ‘Yes’ as a ‘No’), respectively. Furthermore, we will also make use of the notation SN and SP to denote the *Sensitivity*, and *Specificity*, respectively, and which are defined as follows:

$$SN = \frac{TP}{TP + FN}, \quad \text{and} \quad SP = \frac{TN}{TN + FP}.$$

In words, SN describes the proportion of *correctly diagnosed* ‘Yes’ cases relative to the total number of observed ‘Yes’ cases, and SP is the number of *correctly diagnosed* ‘No’ cases relative to the total number of observed ‘No’ cases.

In order to perform the simulation study now, all the simulation settings have to be discussed first. These settings include the model that will be used, elements that must be chosen for each of the different influence diagnostic methods, as well as some settings relative to the data sets that will be generated. All these necessary settings will be provided in the next section.

6.2 Simulation settings

The settings for the simulation study to follow are given below. The linear regression model given in (4.1) is used in these simulations to randomly generate data sets. The choice of the regression coefficient parameters $(\beta_0, \dots, \beta_{p-1})$, used in this data generation, are all set equal to 1, i.e., for a given case i , the regression model used to generate data is

$$Y_i = 1 + X_{1i} + X_{2i} + \dots + X_{p-1,i} + \varepsilon_i.$$

Furthermore, the covariates used in this regression model, $\mathbf{X}$, are randomly generated from a $N(0, 1)$ distribution, whereas the error terms, $\varepsilon_i, i = 1, 2, \dots, n$, are randomly generated from one of the five distributions given in Table 6.2. The specific parameter settings for each of these distributions used in this simulation (along with convenient abbreviations for these settings) is given in Table 6.3 and the shapes of the distributions are illustrated in Figure 6.1.

Table 6.2: Density functions used for error distributions.

Distribution Name	Density function
Standard normal, $N(0, 1)$	$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
t_ν	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
Log-normal(μ, σ)	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
Laplace(μ, b)	$f(x) = \frac{1}{2b} \exp\left(-\frac{ x-\mu }{b}\right)$
Contaminated normal ($\mu_1, \sigma_1, \mu_2, \sigma_2, p$)	$f(x) = p\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-p)\phi\left(\frac{x-\mu_2}{\sigma_2}\right)$

Table 6.3: Error distributions used in the simulation study.

Abbreviation	Distribution Name	Parameter settings
N	Standard normal, $N(0, 1)$	—
T3	t_ν	$\nu = 3$
LN	Log-normal	$\mu = 0, \sigma = 1$
LP	Laplace	$\mu = 0, b = 3$
CN1	Contaminated normal	$\mu_1 = -1.3, \sigma_1 = 1, \mu_2 = 1.3, \sigma_2 = 1, p = 0.5$
CN2	Contaminated normal	$\mu_1 = 0, \sigma_1 = \frac{1}{\sqrt{13}}, \mu_2 = 0, \sigma_2 = \frac{5}{\sqrt{13}}, p = 0.5$

The reason for choosing the specific distributions listed in Table 6.2 for generating the error terms is to have a variety of error distributions consisting of a variety of distributional properties, such as normality, heavy-tailedness, right-tail skewness, and multiple modes. The many data sets simulated in this study will also vary in sample size (n), the number of regression parameters included in the model used to generate the data (p), as well as the number of inserted influential values (δ). The specific values for these settings used in the study are:

- $n = 100, 200$ and 500 ,
- $p = 5$ and 15 , and
- $\delta = 5$ and 10 .

A total of 100 data sets are generated for each combination of error distribution, n , p , and δ in this study.

Furthermore, each of the influential diagnostic methods have elements that need to be specifically chosen for them. The JaB method, for example, is performed using $B = 1\,000$ bootstrap replications and the ADAP method is performed using $M = 1\,000$ Monte Carlo

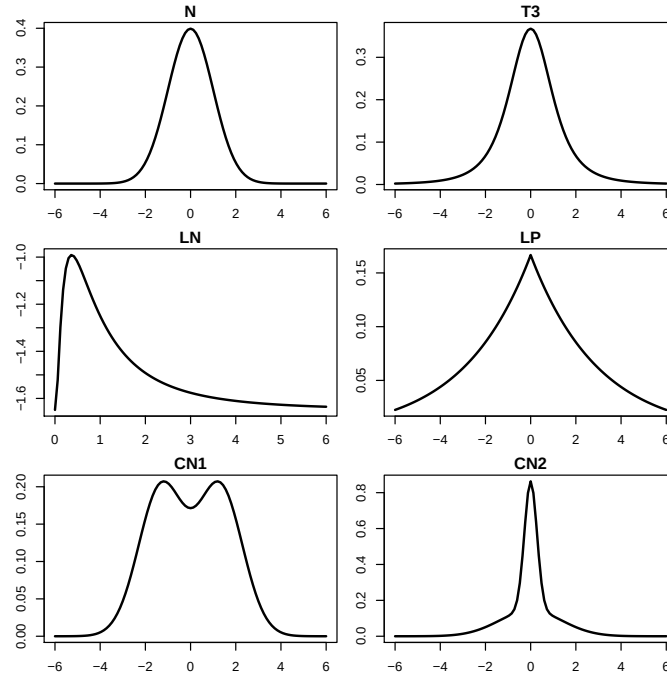


Figure 6.1: An illustration of the error distributions given in Table 6.3.

replications. The Combined method, on the other hand, uses $B = 300$ and $M = 1\,000$. The reason for choosing $B = 300$ in this case is because of the fact that there will be approximately B/e resamples that do not contain the i^{th} observation (see Remark 4.2 on page 33). Choosing $B = 300$ therefore means that approximately 100 of these bootstrap resamples will not contain the i^{th} observation, meaning that we will have at least 100 ‘useful’ bootstrap resamples per observation for the simulation; it was decided that, given the number of Monte Carlo simulations and the extraordinarily long run-times of this simulation, that this would be sufficient for the study. For simplicity, the New method will make use of an α of 0.05. Lastly, the MC^3 method is performed with 1 000 iterations and the New method proposed in this dissertation is executed using 1 000 000 Monte Carlo simulations. The R code for the way in which we implemented these methods, can be found on GitHub using the following link: <https://github.com/ClarisaBooyesen/Dissertation-Codes>.

Finally, in the section that follows, all the results of the simulation study will be given, followed by a detailed discussion of these results as well as an overall summary.

6.3 Results

In this section, the results obtained in the simulation study will be provided in the form of tables. Tables 6.4 to 6.9, include the results on how well the six influential diagnostic methods discussed in the previous chapters (i.e., the Traditional method, JaB method, ADAP method, Combined method, New method, and the MC^3 method) have performed in finding the δ influential observations inserted into the data sets simulated using the *simple approach*. Tables 6.10 to 6.15, on the other hand, contain the same results, but here the methods performances’ are investigated when they are applied to the data set

that is simulated using the *masked approach*. All of these tables summarise the results by making use of the averages over the true positive (TP) and false positive (FP) counts, and the estimated sensitivity (SN) and specificity (SP) values calculated over 100 simulated data sets. Below each of the quantities their corresponding sample variance over the 100 data sets is also reported in parentheses. Furthermore, the calculated cut-off values used in each of the methods (along with their sample variances across the 100 data sets) are also provided in separate tables. The Traditional method and New method cut-offs are given in Table 6.16, the JaB method cut-offs in Table 6.17, and the Combined method cut-off are stated in Table 6.18. Since the JaB method and the Combined method consists of cut-off values for all n observations in the data set, we thought it would be a good idea to split these cut-off values into two sets, namely, one set containing the cut-off values corresponding to the δ simulated influential observations, and the second set containing the remaining cut-off values. Furthermore, instead of stating all of these values separately, we calculated the means over these two sets such that they can easily be compared to one another. In Tables 6.17 and 6.18, we denote the means of the cut-off values over the δ simulated influential observations with the term “Delta” and the mean of the remaining cut-off values with the term “All”. No table is provided for the cut-off value of the ADAP method, since it only uses a single cut-off value in the initial check stage, as well as in the confirmation stage, namely the cut-off value of 1. Lastly, Table 6.19 contains the median (over the different error distribution data sets) of run-time (in seconds) that each of the influential diagnostic methods took to execute when applied to the different simulated data sets.

Table 6.4: Average TP, FP, SN and SP for the various methods across 100 data sets generated with an $N(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
N	Traditional	TP		2.880 (1.541)	3.030 (0.029)	1.890 (0.665)	3.030 (0.029)	3.660 (1.883)	3.040 (0.039)	2.420 (0.913)	3.000 (0.000)	4.710 (0.430)	4.050 (0.674)	3.300 (1.788)	3.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.576 (0.062)	0.303 (0.000)	0.378 (0.027)	0.303 (0.000)	0.732 (0.075)	0.304 (0.000)	0.484 (0.037)	0.300 (0.000)	0.942 (0.017)	0.405 (0.007)	0.660 (0.072)	0.300 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP		4.940 (0.057)	5.560 (1.037)	4.630 (0.417)	6.840 (1.590)	4.930 (0.126)	9.890 (0.119)	4.860 (0.202)	8.590 (1.295)	5.000 (0.000)	9.980 (0.020)	4.950 (0.068)	9.370 (0.639)
		FP		0.060 (0.077)	0.000 (0.000)	0.110 (0.099)	0.030 (0.029)	1.310 (1.327)	0.150 (0.169)	1.350 (1.402)	1.420 (1.074)	7.700 (4.798)	11.230 (3.371)	7.080 (10.276)	12.600 (3.636)
		SN		0.988 (0.002)	0.556 (0.010)	0.926 (0.017)	0.684 (0.016)	0.986 (0.005)	0.989 (0.001)	0.972 (0.008)	0.859 (0.013)	1.000 (0.000)	0.998 (0.000)	0.990 (0.003)	0.937 (0.006)
		SP		0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.993 (0.000)	0.999 (0.000)	0.993 (0.000)	0.993 (0.000)	0.984 (0.000)	0.977 (0.000)	0.986 (0.000)	0.974 (0.000)
	ADAP	TP		4.970 (0.050)	10.000 (0.000)	2.300 (1.606)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	3.950 (1.220)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.800 (0.283)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.994 (0.002)	1.000 (0.000)	0.460 (0.064)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.790 (0.049)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.960 (0.011)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP		4.560 (0.309)	3.110 (1.190)	3.150 (0.795)	2.820 (2.654)	5.000 (0.000)	9.970 (0.029)	5.000 (0.000)	9.810 (0.297)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.990 (1.000)	0.000 (0.000)	0.370 (0.518)	0.000 (0.000)	9.950 (5.442)	6.260 (4.194)	6.010 (6.010)	2.350 (2.068)
		SN		0.912 (0.012)	0.311 (0.012)	0.630 (0.032)	0.282 (0.027)	1.000 (0.000)	0.997 (0.000)	1.000 (0.000)	0.981 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.995 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.980 (0.000)	0.987 (0.000)	0.988 (0.000)	0.995 (0.000)
	New	TP		4.780 (0.335)	6.550 (2.856)	4.500 (0.576)	6.520 (1.909)	4.890 (0.220)	9.430 (0.591)	4.790 (0.309)	7.74 (1.730)	5.000 (0.000)	9.920 (0.074)	4.910 (0.123)	8.860 (1.112)
		FP		0.000 (0.000)	0.000 (0.000)	0.020 (0.020)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.090 (0.083)	0.06 (0.057)	0.020 (0.020)	0.000 (0.000)	0.060 (0.077)	0.630 (0.619)
		SN		0.956 (0.013)	0.655 (0.029)	0.900 (0.023)	0.652 (0.019)	0.978 (0.009)	0.943 (0.006)	0.958 (0.012)	0.774 (0.017)	1.000 (0.000)	0.992 (0.001)	0.982 (0.005)	0.886 (0.011)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)
	MC^3	TP		4.530 (1.039)	3.400 (1.030)	2.890 (1.150)	3.790 (1.279)	4.960 (0.079)	5.870 (9.589)	4.100 (1.182)	5.730 (9.411)	5.000 (0.000)	10.000 (0.000)	4.820 (0.311)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.050 (0.068)	0.010 (0.010)	0.070 (0.658)	0.020 (0.020)	0.220 (0.214)	0.000 (0.000)	0.220 (0.254)	0.000 (0.000)
		SN		0.906 (0.042)	0.340 (0.010)	0.578 (0.046)	0.379 (0.013)	0.992 (0.003)	0.587 (0.096)	0.820 (0.047)	0.573 (0.094)	1.000 (0.000)	1.000 (0.000)	0.964 (0.012)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.5: Average TP, FP, SN and SP for the various methods across 100 data sets generated with an t_3 error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
T3	Traditional	TP	3.180 (2.149)	3.040 (0.039)	1.960 (0.887)	3.100 (0.091)	3.740 (2.215)	3.120 (0.167)	2.410 (1.578)	3.040 (0.039)	4.870 (0.336)	5.820 (4.109)	3.570 (3.076)	3.030 (0.029)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.636 (0.086)	0.304 (0.000)	0.392 (0.035)	0.310 (0.001)	0.748 (0.089)	0.312 (0.002)	0.482 (0.063)	0.304 (0.000)	0.974 (0.013)	0.582 (0.041)	0.714 (0.123)	0.303 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	4.970 (0.090)	5.650 (1.462)	4.760 (0.285)	7.510 (1.263)	4.990 (0.010)	10.000 (0.000)	4.990 (0.010)	9.860 (0.122)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.050 (0.109)	0.000 (0.000)	0.120 (0.107)	0.000 (0.000)	0.840 (1.631)	0.030 (0.029)	1.010 (0.879)	0.250 (0.250)	5.390 (20.099)	5.570 (6.369)	4.050 (5.018)	5.700 (6.192)
		SN	0.994 (0.004)	0.565 (0.015)	0.952 (0.011)	0.751 (0.013)	0.998 (0.000)	1.000 (0.000)	0.998 (0.000)	0.986 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.995 (0.000)	0.999 (0.000)	0.989 (0.000)	0.989 (0.000)	0.992 (0.000)	0.988 (0.000)
	ADAP	TP	5.000 (0.000)	6.960 (6.907)	4.090 (1.679)	9.530 (2.635)	5.000 (0.000)	7.480 (8.838)	4.980 (0.040)	9.530 (2.615)	5.000 (0.000)	9.450 (2.028)	5.000 (0.000)	9.550 (2.351)
		FP	0.000 (0.000)	0.010 (0.010)	0.020 (0.020)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	0.696 (0.069)	0.818 (0.067)	0.953 (0.026)	1.000 (0.000)	0.748 (0.088)	0.996 (0.002)	0.953 (0.026)	1.000 (0.000)	0.945 (0.020)	1.000 (0.000)	0.955 (0.024)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.670 (0.284)	3.340 (1.782)	2.780 (1.244)	2.590 (2.568)	5.000 (0.000)	9.600 (0.889)	5.000 (0.000)	9.880 (0.288)	5.000 (0.000)	9.980 (0.020)	5.000 (0.000)	9.990 (0.010)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.780 (0.739)	0.000 (0.000)	0.240 (0.245)	0.000 (0.000)	6.340 (6.611)	5.330 (4.122)	3.860 (4.546)	2.310 (2.054)
		SN	0.934 (0.011)	0.334 (0.018)	0.556 (0.050)	0.259 (0.026)	1.000 (0.000)	0.960 (0.009)	1.000 (0.000)	0.988 (0.003)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.999 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.987 (0.000)	0.989 (0.000)	0.992 (0.000)	0.995 (0.000)
	New	TP	4.940 (0.138)	8.510 (2.232)	4.700 (0.313)	7.600 (1.899)	4.980 (0.020)	9.990 (0.010)	4.960 (0.099)	9.680 (0.381)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.030 (0.050)	0.000 (0.000)	0.050 (0.048)	0.020 (0.020)	0.190 (0.539)	0.000 (0.000)	0.170 (0.183)	0.020 (0.020)
		SN	0.988 (0.006)	0.851 (0.223)	0.940 (0.013)	0.760 (0.019)	0.996 (0.001)	0.999 (0.000)	0.992 (0.004)	0.968 (0.004)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP	5.000 (0.000)	5.440 (8.188)	3.740 (2.558)	4.630 (4.619)	5.000 (0.000)	6.390 (10.604)	4.980 (0.040)	4.360 (6.475)	4.690 (0.378)	8.410 (3.376)	4.540 (0.918)	6.990 (8.394)
		FP	0.610 (0.442)	0.140 (0.122)	0.380 (0.278)	0.060 (0.057)	1.990 (1.566)	0.880 (0.531)	2.020 (1.555)	0.490 (0.273)	7.740 (44.922)	8.470 (88.211)	9.370 (67.993)	6.690 (81.772)
		SN	1.000 (0.000)	0.544 (0.082)	0.748 (0.102)	0.463 (0.046)	1.000 (0.000)	0.639 (0.106)	0.996 (0.002)	0.436 (0.065)	0.938 (0.015)	0.841 (0.034)	0.908 (0.037)	0.699 (0.084)
		SP	0.994 (0.000)	0.998 (0.000)	0.996 (0.000)	0.999 (0.000)	0.990 (0.000)	0.995 (0.000)	0.990 (0.000)	0.997 (0.000)	0.984 (0.000)	0.983 (0.000)	0.981 (0.000)	0.986 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.6: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Log-normal (0,1) error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
LN	Traditional	TP	2.990 (2.172)	3.050 (0.048)	2.190 (1.004)	3.050 (0.048)	4.000 (2.000)	3.230 (0.260)	2.640 (1.768)	3.050 (0.048)	4.930 (0.147)	6.060 (5.087)	3.880 (2.854)	3.080 (0.095)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.598 (0.087)	0.305 (0.000)	0.438 (0.040)	0.305 (0.000)	0.800 (0.080)	0.323 (0.003)	0.528 (0.071)	0.305 (0.000)	0.986 (0.006)	0.606 (0.051)	0.776 (0.114)	0.308 (0.001)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	4.970 (0.029)	5.360 (0.920)	4.930 (0.126)	7.700 (1.667)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.940 (0.057)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.050 (0.068)	0.000 (0.000)	0.160 (0.176)	0.000 (0.000)	0.660 (0.631)	0.000 (0.000)	1.020 (0.868)	0.150 (0.129)	4.880 (15.925)	5.390 (4.543)	4.500 (5.162)	5.020 (2.828)
		SN	0.994 (0.001)	0.536 (0.009)	0.986 (0.005)	0.770 (0.017)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.994 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	0.999 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.997 (0.000)	1.000 (0.000)	0.995 (0.000)	0.999 (0.000)	0.990 (0.000)	0.989 (0.000)	0.991 (0.000)	0.990 (0.000)
	ADAP	TP	5.000 (0.000)	8.660 (5.903)	4.710 (0.491)	8.920 (4.923)	5.000 (0.000)	9.440 (2.633)	5.000 (0.000)	8.690 (6.559)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	8.620 (5.592)
		FP	0.010 (0.010)	0.000 (0.000)	0.070 (0.086)	0.040 (0.039)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	0.866 (0.059)	0.942 (0.020)	0.892 (0.049)	1.000 (0.000)	0.944 (0.026)	1.000 (0.000)	0.869 (0.066)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.862 (0.056)
		SP	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.700 (0.232)	2.960 (2.423)	2.460 (1.019)	2.090 (1.658)	5.000 (0.000)	9.930 (0.288)	5.000 (0.000)	9.710 (1.178)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.990 (0.010)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.780 (0.537)	0.000 (0.000)	0.320 (0.260)	0.000 (0.000)	5.220 (4.941)	4.920 (3.872)	4.050 (3.907)	2.690 (2.539)
		SN	0.940 (0.009)	0.296 (0.024)	0.492 (0.041)	0.209 (0.017)	1.000 (0.000)	0.993 (0.003)	1.000 (0.000)	0.971 (0.012)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.989 (0.000)	0.990 (0.000)	0.992 (0.000)	0.995 (0.000)
	New	TP	4.950 (0.088)	8.530 (2.029)	4.870 (0.175)	7.910 (2.345)	5.000 (0.000)	9.990 (0.010)	4.990 (0.010)	9.870 (0.134)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.010 (0.010)	0.000 (0.000)	0.060 (0.077)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.090 (0.143)	0.000 (0.000)	0.200 (0.202)	0.040 (0.039)
		SN	0.990 (0.004)	0.853 (0.020)	0.974 (0.007)	0.791 (0.023)	1.000 (0.000)	0.999 (0.000)	0.998 (0.000)	0.987 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP	5.000 (0.000)	6.870 (11.650)	4.070 (2.328)	5.150 (7.422)	5.000 (0.000)	9.130 (4.882)	5.000 (0.000)	5.350 (9.745)	3.080 (1.428)	5.550 (4.836)	2.810 (2.034)	4.250 (3.563)
		FP	1.030 (0.413)	0.210 (0.168)	0.980 (0.585)	0.190 (0.155)	3.300 (4.737)	1.460 (0.837)	3.070 (2.571)	0.890 (0.442)	27.920 (118.579)	27.980 (114.545)	27.540 (115.948)	22.700 (103.424)
		SN	1.000 (0.000)	0.687 (0.116)	0.814 (0.093)	0.515 (0.074)	1.000 (0.000)	0.913 (0.049)	1.000 (0.000)	0.535 (0.097)	0.616 (0.057)	0.555 (0.048)	0.562 (0.081)	0.425 (0.036)
		SP	0.989 (0.000)	0.998 (0.000)	0.990 (0.000)	0.998 (0.000)	0.983 (0.000)	0.992 (0.000)	0.984 (0.000)	0.995 (0.000)	0.944 (0.000)	0.943 (0.000)	0.944 (0.000)	0.954 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.7: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Laplace (0, 3) error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
LP	Traditional	TP	2.890 (2.281)	3.060 (0.057)	2.070 (0.975)	3.120 (0.127)	4.050 (2.149)	3.220 (0.214)	2.380 (1.248)	3.030 (0.029)	4.790 (0.774)	5.710 (4.067)	3.740 (3.245)	3.040 (0.039)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.578 (0.091)	0.306 (0.001)	0.414 (0.039)	0.312 (0.001)	0.810 (0.086)	0.322 (0.002)	0.476 (0.050)	0.303 (0.000)	0.958 (0.031)	0.571 (0.041)	0.748 (0.130)	0.304 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	5.000 (0.000)	5.690 (1.408)	4.930 (0.086)	7.460 (1.503)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.970 (0.029)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.020 (0.020)	0.000 (0.000)	0.150 (0.149)	0.000 (0.000)	0.710 (0.673)	0.000 (0.000)	1.020 (0.727)	0.150 (0.169)	7.250 (15.765)	6.600 (3.677)	6.190 (4.883)	5.920 (3.286)
		SN	1.000 (0.000)	0.569 (0.014)	0.986 (0.003)	0.746 (0.015)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.997 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.995 (0.000)	0.999 (0.000)	0.985 (0.000)	0.987 (0.000)	0.987 (0.000)	0.988 (0.000)
	ADAP	TP	5.000 (0.000)	9.590 (1.517)	3.500 (2.232)	4.970 (5.060)	5.000 (0.000)	9.990 (0.010)	4.990 (0.010)	5.350 (6.391)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	5.790 (8.188)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	0.959 (0.015)	0.700 (0.089)	0.497 (0.051)	1.000 (0.000)	0.999 (0.000)	0.998 (0.000)	0.535 (0.064)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.579 (8.188)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.770 (0.219)	2.740 (1.770)	2.870 (1.387)	2.060 (1.552)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.630 (1.064)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.710 (0.592)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.770 (0.603)	0.000 (0.000)	0.320 (0.301)	0.000 (0.000)	8.860 (6.465)	6.690 (4.903)	5.540 (4.069)	3.400 (2.283)
		SN	0.954 (0.009)	0.274 (0.018)	0.574 (0.055)	0.206 (0.016)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.963 (0.011)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.971 (0.006)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.982 (0.000)	0.986 (0.000)	0.989 (0.000)	0.993 (0.000)
	New	TP	5.000 (0.000)	8.830 (1.981)	4.810 (0.236)	7.760 (2.002)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.850 (0.129)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.030 (0.029)	0.000 (0.000)	0.180 (0.392)	0.000 (0.000)	0.100 (0.111)	0.010 (0.010)
		SN	1.000 (0.000)	0.883 (0.020)	0.962 (0.009)	0.776 (0.020)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.985 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP	4.960 (0.160)	7.980 (8.929)	3.030 (2.979)	4.590 (5.052)	5.000 (0.000)	9.830 (1.052)	4.830 (0.587)	4.580 (6.650)	4.080 (1.145)	7.670 (4.526)	4.110 (1.331)	3.880 (4.228)
		FP	0.240 (0.225)	0.010 (0.010)	0.200 (0.202)	0.020 (0.020)	1.170 (0.951)	0.630 (0.397)	1.000 (0.889)	0.230 (0.199)	12.450 (86.896)	17.120 (149.925)	11.710 (91.885)	8.650 (82.896)
		SN	0.992 (0.006)	0.798 (0.089)	0.606 (0.119)	0.459 (0.051)	1.000 (0.000)	0.983 (0.011)	0.966 (0.023)	0.458 (0.067)	0.816 (0.046)	0.767 (0.045)	0.822 (0.053)	0.388 (0.042)
		SP	0.997 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.994 (0.000)	0.997 (0.000)	0.995 (0.000)	0.999 (0.000)	0.975 (0.000)	0.965 (0.001)	0.976 (0.000)	0.982 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.8: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(-1.3, 1, 1.3, 1, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
CN1	Traditional	TP	3.040 (1.312)	3.030 (0.029)	1.980 (0.767)	3.090 (0.083)	3.990 (1.687)	3.060 (0.057)	2.520 (0.878)	3.020 (0.020)	4.630 (0.538)	4.140 (0.909)	3.070 (1.783)	3.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.608 (0.052)	0.303 (0.000)	0.396 (0.031)	0.309 (0.001)	0.798 (0.067)	0.306 (0.001)	0.504 (0.035)	0.302 (0.000)	0.926 (0.022)	0.414 (0.009)	0.614 (0.071)	0.300 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	4.960 (0.039)	5.630 (0.842)	4.620 (0.379)	6.780 (1.608)	4.970 (0.050)	9.820 (0.169)	4.900 (0.131)	8.510 (1.424)	4.990 (0.010)	9.950 (0.048)	4.960 (0.039)	9.280 (0.567)
		FP	0.030 (0.029)	0.000 (0.000)	0.180 (0.149)	0.060 (0.057)	1.090 (1.295)	0.270 (0.260)	1.320 (1.250)	1.230 (0.987)	7.660 (8.126)	11.190 (3.287)	6.870 (5.023)	12.700 (3.889)
		SN	0.992 (0.002)	0.563 (0.008)	0.924 (0.015)	0.678 (0.016)	0.994 (0.002)	0.982 (0.002)	0.980 (0.005)	0.851 (0.014)	0.998 (0.000)	0.995 (0.000)	0.992 (0.002)	0.928 (0.006)
		SP	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	0.999 (0.000)	0.994 (0.000)	0.999 (0.000)	0.993 (0.000)	0.994 (0.000)	0.985 (0.000)	0.977 (0.000)	0.986 (0.000)	0.974 (0.000)
	ADAP	TP	4.930 (0.147)	9.920 (0.640)	2.370 (1.508)	10.000 (0.000)	4.990 (0.010)	10.000 (0.000)	4.220 (1.244)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.840 (0.277)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.986 (0.006)	0.992 (0.006)	0.474 (0.060)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.844 (0.050)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.968 (0.011)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.550 (0.412)	3.360 (1.142)	2.930 (0.975)	3.030 (2.413)	5.000 (0.000)	9.970 (0.029)	4.970 (0.050)	9.910 (0.083)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.940 (0.865)	0.000 (0.000)	0.270 (0.280)	0.000 (0.000)	9.610 (5.675)	6.160 (3.772)	6.130 (5.771)	2.360 (1.829)
		SN	0.910 (0.016)	0.336 (0.011)	0.586 (0.039)	0.303 (0.024)	1.000 (0.000)	0.997 (0.000)	0.994 (0.002)	0.991 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.995 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.981 (0.000)	0.987 (0.000)	0.988 (0.000)	0.995 (0.000)
	New	TP	4.910 (0.123)	6.930 (3.298)	4.470 (0.595)	6.310 (1.792)	4.970 (0.050)	9.380 (0.521)	4.800 (0.323)	7.770 (2.118)	4.990 (0.010)	9.860 (0.142)	4.920 (0.135)	8.600 (1.051)
		FP	0.000 (0.000)	0.000 (0.000)	0.030 (0.029)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.020 (0.020)	0.080 (0.095)	0.060 (0.057)	0.010 (0.010)	0.040 (0.039)	0.900 (0.859)
		SN	0.982 (0.005)	0.693 (0.033)	0.894 (0.024)	0.631 (0.018)	0.994 (0.002)	0.938 (0.005)	0.960 (0.013)	0.777 (0.021)	0.998 (0.000)	0.986 (0.001)	0.984 (0.005)	0.860 (0.011)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)
	MC^3	TP	4.560 (1.017)	3.500 (1.081)	2.860 (1.354)	3.830 (1.536)	4.990 (0.010)	6.680 (10.725)	4.480 (0.798)	5.930 (10.268)	5.000 (0.000)	9.950 (0.250)	4.850 (0.311)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.040 (0.039)	0.030 (0.029)	0.020 (0.020)	0.040 (0.039)	0.310 (0.236)	0.020 (0.020)	0.150 (0.129)	0.000 (0.000)
		SN	0.912 (0.041)	0.350 (0.011)	0.572 (0.054)	0.383 (0.015)	0.998 (0.000)	0.668 (0.107)	0.896 (0.032)	0.593 (0.103)	1.000 (0.000)	0.995 (0.003)	0.970 (0.012)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.9: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(0, \frac{1}{\sqrt{13}}, 0, \frac{5}{\sqrt{13}}, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Simple” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
CN2	Traditional	TP	2.820 (1.846)	3.030 (0.029)	2.000 (0.869)	3.050 (0.048)	3.690 (2.135)	3.070 (0.066)	2.280 (1.052)	3.010 (0.010)	4.550 (0.856)	4.100 (0.798)	3.200 (1.737)	3.010 (0.010)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.564 (0.074)	0.303 (0.000)	0.400 (0.035)	0.305 (0.000)	0.738 (0.085)	0.307 (0.001)	0.456 (0.042)	0.301 (0.000)	0.910 (0.034)	0.410 (0.008)	0.640 (0.069)	0.301 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	4.910 (0.143)	5.600 (0.949)	4.560 (0.431)	7.020 (1.899)	4.960 (0.079)	9.780 (0.234)	4.870 (0.195)	8.790 (1.238)	4.980 (0.040)	9.950 (0.068)	4.950 (0.048)	9.380 (0.561)
		FP	0.060 (0.118)	0.000 (0.000)	0.210 (0.208)	0.040 (0.039)	1.020 (1.575)	0.210 (0.168)	1.200 (0.970)	1.220 (1.284)	8.670 (12.910)	11.010 (3.141)	7.070 (6.854)	12.310 (4.236)
		SN	0.982 (0.006)	0.560 (0.009)	0.912 (0.017)	0.702 (0.019)	0.992 (0.003)	0.978 (0.002)	0.974 (0.008)	0.879 (0.012)	0.996 (0.002)	0.995 (0.001)	0.990 (0.002)	0.938 (0.006)
		SP	0.999 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.995 (0.000)	0.999 (0.000)	0.994 (0.000)	0.994 (0.000)	0.982 (0.000)	0.978 (0.000)	0.986 (0.000)	0.975 (0.000)
	ADAP	TP	4.980 (0.020)	9.920 (0.256)	2.220 (1.608)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	3.940 (1.289)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.740 (0.578)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.996 (0.001)	0.992 (0.003)	0.444 (0.064)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.788 (0.052)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.948 (0.023)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.610 (0.301)	3.680 (1.331)	2.970 (0.696)	3.120 (2.329)	5.000 (0.000)	9.980 (0.020)	4.970 (0.050)	9.860 (0.122)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.020 (1.050)	0.000 (0.000)	0.280 (0.305)	0.000 (0.000)	10.620 (8.460)	6.710 (4.349)	6.510 (5.141)	2.190 (2.458)
		SN	0.922 (0.012)	0.368 (0.013)	0.594 (0.028)	0.312 (0.023)	1.000 (0.000)	0.998 (0.000)	0.994 (0.002)	0.986 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.995 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.979 (0.000)	0.986 (0.000)	0.987 (0.000)	0.996 (0.000)
	New	TP	4.810 (0.317)	6.930 (3.116)	4.370 (0.599)	6.690 (2.095)	4.940 (0.138)	9.360 (0.596)	4.810 (0.277)	7.840 (2.378)	4.960 (0.059)	9.870 (0.175)	4.880 (0.187)	8.800 (1.232)
		FP	0.010 (0.010)	0.000 (0.000)	0.020 (0.020)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.080 (0.095)	0.110 (0.139)	0.010 (0.010)	0.100 (0.111)	0.700 (0.980)
		SN	0.962 (0.013)	0.693 (0.031)	0.874 (0.024)	0.669 (0.021)	0.988 (0.006)	0.936 (0.006)	0.962 (0.011)	0.784 (0.024)	0.992 (0.002)	0.987 (0.002)	0.976 (0.007)	0.880 (0.012)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)
	MC^3	TP	4.670 (0.870)	3.700 (1.788)	2.800 (0.990)	4.020 (1.818)	4.940 (0.118)	5.940 (10.239)	4.070 (1.379)	6.020 (9.495)	5.000 (0.000)	9.910 (0.446)	4.760 (0.467)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.020 (0.020)	0.020 (0.020)	0.020 (0.020)	0.040 (0.039)	0.000 (0.000)	0.180 (0.189)	0.010 (0.010)	0.140 (0.122)	0.010 (0.010)
		SN	0.934 (0.035)	0.370 (0.018)	0.560 (0.040)	0.402 (0.018)	0.988 (0.005)	0.594 (0.102)	0.814 (0.055)	0.602 (0.095)	1.000 (0.000)	0.991 (0.004)	0.952 (0.019)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.10: Average TP, FP, SN and SP for the various methods across 100 data sets generated with an $N(0, 1)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
N	Traditional	TP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP		5.000 (0.000)	7.760 (1.275)	5.000 (0.000)	8.750 (0.876)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.020 (0.020)	0.000 (0.000)	0.160 (0.156)	0.010 (0.010)	2.230 (1.128)	0.000 (0.000)	1.880 (1.198)	0.000 (0.000)	11.490 (5.949)	7.670 (5.799)	9.310 (4.519)	5.630 (4.599)
		SN		1.000 (0.000)	0.776 (0.013)	1.000 (0.000)	0.875 (0.009)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.998 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.989 (0.000)	1.000 (0.000)	0.990 (0.000)	1.000 (0.000)	0.977 (0.000)	0.984 (0.000)	0.981 (0.000)	0.989 (0.000)
	ADAP	TP		5.000 (0.000)	10.000 (0.000)	4.890 (0.989)	9.900 (0.091)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.680 (2.543)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.570 (2.106)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		1.000 (0.000)	1.000 (0.000)	0.978 (0.004)	0.990 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.968 (0.025)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.957 (0.021)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP		4.850 (0.129)	2.440 (1.582)	1.860 (1.091)	0.260 (0.295)	5.000 (0.000)	9.980 (0.198)	5.000 (0.000)	9.610 (0.321)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.480 (1.060)	0.000 (0.000)	0.400 (0.364)	0.000 (0.000)	11.510 (5.303)	7.610 (4.988)	6.330 (4.203)	3.830 (3.031)
		SN		0.970 (0.005)	0.244 (0.016)	0.372 (0.044)	0.026 (0.003)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.961 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.992 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.977 (0.000)	0.984 (0.000)	0.987 (0.000)	0.992 (0.000)
	New	TP		5.000 (0.000)	9.770 (0.199)	4.980 (0.020)	9.150 (1.159)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.030 (0.029)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		1.000 (0.000)	0.977 (0.002)	0.996 (0.001)	0.915 (0.116)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP		5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.980 (0.040)
		FP		0.070 (0.066)	0.000 (0.000)	0.040 (0.039)	0.010 (0.010)	0.080 (0.074)	0.020 (0.020)	0.060 (0.057)	0.020 (0.020)	0.180 (0.169)	0.090 (0.083)	0.110 (0.099)	0.140 (1.031)
		SN		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)
		SP		0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.11: Average TP, FP, SN and SP for the various methods across 100 data sets generated with an t_3 error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
T3	Traditional	TP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP		5.000 (0.000)	7.590 (1.295)	4.990 (0.010)	8.820 (0.836)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.110 (0.099)	0.010 (0.010)	0.420 (0.387)	0.010 (0.010)	3.270 (0.926)	0.110 (0.099)	2.930 (1.035)	0.110 (0.099)	14.030 (3.969)	9.510 (2.818)	11.590 (4.527)	7.720 (3.173)
		SN		1.000 (0.000)	0.759 (0.013)	0.998 (0.000)	0.882 (0.008)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.999 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.983 (0.000)	0.999 (0.000)	0.985 (0.000)	0.999 (0.000)	0.972 (0.000)	0.981 (0.000)	0.977 (0.000)	0.984 (0.000)
	ADAP	TP		5.000 (0.000)	10.000 (0.000)	4.940 (0.077)	9.960 (0.039)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.830 (0.971)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.430 (3.056)
		FP		0.010 (0.010)	0.030 (0.029)	0.000 (0.000)	0.010 (0.010)	0.050 (0.048)	0.000 (0.000)	0.020 (0.020)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		1.000 (0.000)	1.000 (0.000)	0.988 (0.003)	0.996 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.983 (0.010)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.943 (0.031)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP		4.880 (0.107)	2.570 (1.561)	1.740 (0.942)	0.250 (0.210)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.760 (0.225)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.960 (0.968)	0.000 (0.000)	0.770 (0.563)	0.000 (0.000)	12.330 (4.244)	7.880 (3.703)	7.470 (3.928)	4.860 (3.354)
		SN		0.976 (0.004)	0.257 (0.016)	0.348 (0.038)	0.025 (0.002)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.976 (0.002)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.990 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.975 (0.000)	0.984 (0.000)	0.985 (0.000)	0.990 (0.000)
	New	TP		5.000 (0.000)	9.780 (0.214)	4.990 (0.010)	9.380 (0.642)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.060 (0.057)	0.030 (0.029)	0.150 (0.129)	0.010 (0.010)	0.260 (0.215)	0.060 (0.057)	0.270 (0.199)	0.070 (0.066)	0.560 (0.451)	0.140 (0.122)	0.640 (0.516)	0.220 (0.214)
		SN		1.000 (0.000)	0.978 (0.002)	0.998 (0.000)	0.938 (0.006)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.999 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)
	MC^3	TP		5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.630 (0.579)	8.490 (5.081)	4.470 (0.676)	8.980 (2.343)
		FP		1.510 (1.202)	0.920 (0.519)	1.130 (1.084)	0.420 (0.387)	2.910 (2.244)	2.330 (1.637)	2.630 (2.599)	1.770 (1.169)	10.170 (46.062)	11.030 (102.898)	12.550 (86.109)	10.790 (89.157)
		SN		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.926 (0.023)	0.849 (0.051)	0.894 (0.027)	0.898 (0.023)
		SP		0.984 (0.000)	0.990 (0.000)	0.988 (0.000)	0.995 (0.000)	0.985 (0.000)	0.988 (0.000)	0.987 (0.000)	0.991 (0.000)	0.979 (0.000)	0.977 (0.000)	0.975 (0.000)	0.978 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.12: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Log-normal (0,1) error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
LN	Traditional	TP		0.000 (0.000)	0.000 (0.000)	0.000 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.000 (0.000)	0.000 (0.000)	0.002 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP		5.000 (0.000)	7.320 (1.493)	5.000 (0.000)	8.980 (0.747)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.240 (0.184)	0.020 (0.020)	0.390 (0.321)	0.020 (0.020)	3.740 (0.982)	0.180 (0.149)	3.430 (0.934)	0.240 (0.184)	15.500 (3.909)	11.170 (3.233)	13.630 (3.892)	8.720 (3.355)
		SN		1.000 (0.000)	0.732 (0.015)	1.000 (0.000)	0.898 (0.007)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.997 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.981 (0.000)	0.999 (0.000)	0.982 (0.000)	0.999 (0.000)	0.969 (0.000)	0.977 (0.000)	0.972 (0.000)	0.982 (0.000)
	ADAP	TP		5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.980 (0.198)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.650 (1.947)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.380 (3.895)
		FP		0.080 (0.074)	0.040 (0.039)	0.120 (0.127)	0.030 (0.029)	0.030 (0.029)	0.010 (0.010)	0.050 (0.048)	0.020 (0.020)	0.010 (0.010)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)
		SN		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.965 (0.019)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.938 (0.039)
		SP		0.999 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP		4.850 (0.129)	2.610 (2.038)	1.720 (1.133)	0.290 (0.269)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.560 (0.411)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.010 (0.010)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	2.110 (0.947)	0.000 (0.000)	1.100 (0.535)	0.010 (0.010)	12.390 (4.159)	8.120 (3.299)	8.460 (4.554)	5.170 (3.456)
		SN		0.970 (0.005)	0.261 (0.020)	0.344 (0.045)	0.029 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.956 (0.004)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.989 (0.000)	1.000 (0.000)	0.994 (0.000)	1.000 (0.000)	0.975 (0.000)	0.983 (0.000)	0.983 (0.000)	0.989 (0.000)
	New	TP		5.000 (0.000)	9.750 (0.270)	4.990 (0.010)	9.340 (0.651)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.190 (0.176)	0.050 (0.048)	0.280 (0.244)	0.030 (0.029)	0.450 (0.331)	0.110 (0.099)	0.680 (0.462)	0.180 (0.149)	1.290 (0.733)	0.450 (0.351)	1.180 (0.654)	0.430 (0.328)
		SN		1.000 (0.000)	0.975 (0.003)	0.998 (0.000)	0.934 (0.007)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.998 (0.000)	0.999 (0.000)	0.997 (0.000)	1.000 (0.000)	0.998 (0.000)	0.999 (0.000)	0.997 (0.000)	0.999 (0.000)	0.997 (0.000)	0.999 (0.000)	0.998 (0.000)	0.999 (0.000)
	MC^3	TP		5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	2.300 (2.152)	4.810 (7.004)	2.490 (2.333)	4.940 (5.512)
		FP		3.280 (2.103)	2.080 (1.044)	2.290 (2.087)	1.060 (1.067)	6.750 (6.290)	4.590 (3.113)	5.820 (4.513)	4.010 (2.636)	30.200 (191.697)	29.120 (138.693)	29.630 (167.993)	31.030 (121.646)
		SN		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.460 (0.086)	0.481 (0.070)	0.498 (0.093)	0.494 (0.055)
		SP		0.965 (0.000)	0.977 (0.000)	0.976 (0.000)	0.988 (0.000)	0.965 (0.000)	0.976 (0.000)	0.970 (0.000)	0.979 (0.000)	0.939 (0.001)	0.941 (0.001)	0.940 (0.001)	0.937 (0.001)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.13: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Laplace (0, 3) error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
LP	Traditional	TP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP		5.000 (0.000)	7.810 (1.186)	5.000 (0.000)	8.700 (0.859)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.070 (0.066)	0.000 (0.000)	0.340 (0.227)	0.000 (0.000)	2.910 (1.295)	0.050 (0.048)	2.540 (1.261)	0.020 (0.020)	14.290 (3.380)	9.530 (2.777)	12.010 (5.465)	7.120 (3.844)
		SN		1.000 (0.000)	0.781 (0.000)	1.000 (0.000)	0.870 (0.009)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		0.999 (0.000)	1.000 (0.000)	0.996 (0.000)	1.000 (0.000)	0.985 (0.000)	1.000 (0.000)	0.987 (0.000)	1.000 (0.000)	0.971 (0.000)	0.981 (0.000)	0.976 (0.000)	0.985 (0.000)
	ADAP	TP		5.000 (0.000)	10.000 (0.000)	4.920 (0.095)	9.920 (0.074)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.830 (0.567)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.460 (2.998)
		FP		0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN		1.000 (0.000)	1.000 (0.000)	0.984 (0.004)	0.992 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.983 (0.006)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.946 (0.030)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP		4.840 (0.136)	2.820 (1.745)	1.810 (0.863)	0.310 (0.277)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.670 (0.244)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.760 (0.932)	0.000 (0.000)	0.520 (0.414)	0.000 (0.000)	12.060 (5.269)	7.910 (4.123)	7.470 (5.646)	4.050 (3.220)
		SN		0.968 (0.005)	0.282 (0.017)	0.362 (0.035)	0.031 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.967 (0.002)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.991 (0.000)	1.000 (0.000)	0.997 (0.000)	1.000 (0.000)	0.976 (0.000)	0.984 (0.000)	0.985 (0.000)	0.992 (0.000)
	New	TP		5.000 (0.000)	9.740 (0.255)	4.980 (0.020)	9.220 (0.719)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP		0.010 (0.010)	0.010 (0.010)	0.110 (0.099)	0.020 (0.020)	0.020 (0.020)	0.000 (0.000)	0.070 (0.086)	0.000 (0.000)	0.100 (0.091)	0.010 (0.010)	0.120 (0.107)	0.000 (0.000)
		SN		1.000 (0.000)	0.974 (0.003)	0.996 (0.001)	0.922 (0.007)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP		1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP		5.000 (0.000)	10.000 (0.000)	4.990 (0.010)	9.980 (0.020)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.120 (0.895)	7.680 (5.695)	4.010 (1.000)	7.900 (4.333)
		FP		0.620 (0.602)	0.470 (0.433)	0.580 (0.630)	0.310 (0.277)	1.400 (1.576)	0.870 (0.983)	1.370 (1.347)	0.660 (0.914)	13.720 (105.194)	15.660 (143.701)	14.340 (122.227)	13.610 (121.857)
		SN		1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	0.998 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.824 (0.036)	0.768 (0.057)	0.802 (0.040)	0.790 (0.043)
		SP		0.993 (0.000)	0.995 (0.000)	0.994 (0.000)	0.997 (0.000)	0.993 (0.000)	0.995 (0.000)	0.993 (0.000)	0.997 (0.000)	0.972 (0.000)	0.968 (0.001)	0.971 (0.000)	0.972 (0.001)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.14: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(-1.3, 1, 1.3, 1, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
CN1	Traditional	TP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.000 (1.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	5.000 (0.000)	7.740 (1.043)	5.000 (0.000)	8.560 (0.895)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.020 (0.020)	0.000 (0.000)	0.220 (0.194)	0.000 (0.000)	1.950 (1.543)	0.000 (0.000)	1.820 (1.563)	0.000 (0.000)	10.060 (9.714)	6.080 (6.903)	8.100 (5.990)	5.310 (5.852)
		SN	1.000 (0.000)	0.774 (0.010)	1.000 (0.000)	0.856 (0.009)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.990 (0.000)	1.000 (0.000)	0.991 (0.000)	1.000 (0.000)	0.980 (0.000)	0.988 (0.000)	0.984 (0.000)	0.989 (0.000)
	ADAP	TP	5.000 (0.000)	10.000 (0.000)	4.930 (0.066)	9.840 (1.045)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.880 (0.491)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.430 (3.904)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	1.000 (0.000)	0.986 (0.003)	0.984 (0.010)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.988 (0.005)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.943 (0.039)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.880 (0.127)	2.570 (1.864)	1.770 (0.947)	0.250 (0.230)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.590 (0.345)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.600 (1.172)	0.000 (0.000)	0.390 (0.382)	0.000 (0.000)	11.530 (8.050)	7.080 (5.832)	6.270 (4.664)	3.820 (2.897)
		SN	0.976 (0.005)	0.257 (0.019)	0.354 (0.038)	0.025 (0.002)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.959 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.992 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.977 (0.000)	0.986 (0.000)	0.987 (0.000)	0.992 (0.000)
	New	TP	5.000 (0.000)	9.790 (0.208)	4.980 (0.020)	8.950 (1.078)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.050 (0.048)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	0.979 (0.002)	0.996 (0.001)	0.985 (0.011)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.15: Average TP, FP, SN and SP for the various methods across 100 data sets generated with a Contaminated normal $(0, \frac{1}{\sqrt{13}}, 0, \frac{5}{\sqrt{13}}, 0.5)$ error distribution (estimated variances are stated in parentheses below each entry). Values inserted using the “Masked” approach.

				$n = 100$				$n = 200$				$n = 500$			
				$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
				$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
CN2	Traditional	TP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	JaB	TP	5.000 (0.000)	7.760 (1.376)	5.000 (0.000)	8.810 (0.822)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	10.000 (0.000)
		FP	0.040 (0.039)	0.000 (0.000)	0.290 (0.269)	0.000 (0.000)	3.170 (1.334)	0.010 (0.010)	2.390 (1.250)	0.000 (0.000)	14.340 (3.196)	9.420 (3.034)	11.740 (5.346)	7.220 (3.143)	7.220 (3.143)
		SN	1.000 (0.000)	0.776 (0.014)	1.000 (0.000)	0.881 (0.008)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.997 (0.000)	1.000 (0.000)	0.984 (0.000)	1.000 (0.000)	0.988 (0.000)	1.000 (0.000)	0.971 (0.000)	0.981 (0.000)	0.976 (0.000)	0.985 (0.000)	0.985 (0.000)
	ADAP	TP	5.000 (0.000)	10.000 (0.000)	4.970 (0.029)	9.920 (0.074)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.910 (0.325)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	9.470 (3.039)	9.470 (3.039)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	1.000 (0.000)	0.994 (0.001)	0.992 (0.001)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.991 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.947 (0.030)	0.947 (0.030)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	Combined	TP	4.920 (0.074)	2.310 (1.246)	1.840 (1.388)	0.190 (0.196)	5.000 (0.000)	9.990 (0.010)	5.000 (0.000)	9.610 (0.321)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	10.000 (0.000)
		FP	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	1.960 (1.190)	0.000 (0.000)	0.480 (0.373)	0.000 (0.000)	12.150 (4.189)	7.650 (3.220)	7.210 (4.511)	4.310 (2.923)	4.310 (2.923)
		SN	0.984 (0.003)	0.231 (0.012)	0.368 (0.056)	0.019 (0.002)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.961 (0.003)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.990 (0.000)	1.000 (0.000)	0.998 (0.000)	1.000 (0.000)	0.975 (0.000)	0.984 (0.000)	0.985 (0.000)	0.991 (0.000)	0.991 (0.000)
	New	TP	5.000 (0.000)	9.750 (0.290)	4.980 (0.020)	9.320 (0.624)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	10.000 (0.000)
		FP	0.010 (0.010)	0.000 (0.000)	0.090 (0.083)	0.010 (0.010)	0.030 (0.029)	0.000 (0.000)	0.040 (0.059)	0.000 (0.000)	0.010 (0.010)	0.000 (0.000)	0.040 (0.039)	0.000 (0.000)	0.000 (0.000)
		SN	1.000 (0.000)	0.975 (0.003)	0.996 (0.001)	0.932 (0.006)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
		SP	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	MC^3	TP	5.000 (0.000)	10.000 (0.000)	5.000 (0.000)	10.000 (0.000)	4.990 (0.010)	9.910 (0.143)	4.990 (0.010)	9.960 (0.099)	2.200 (1.677)	3.810 (5.226)	2.340 (1.499)	4.140 (4.223)	4.140 (4.223)
		FP	0.720 (0.951)	0.360 (0.394)	0.510 (0.616)	0.280 (0.264)	1.570 (2.066)	0.980 (2.303)	1.290 (1.380)	0.840 (8.580)	29.930 (149.844)	28.230 (170.482)	27.760 (121.417)	30.680 (110.321)	30.680 (110.321)
		SN	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.000)	0.991 (0.001)	0.998 (0.000)	0.996 (0.001)	0.440 (0.067)	0.381 (0.052)	0.468 (0.060)	0.414 (0.042)	0.414 (0.042)
		SP	0.992 (0.000)	0.996 (0.000)	0.995 (0.000)	0.997 (0.000)	0.992 (0.000)	0.995 (0.000)	0.993 (0.000)	0.996 (0.000)	0.940 (0.001)	0.942 (0.001)	0.944 (0.000)	0.937 (0.000)	0.937 (0.000)

Note that the best TP values (highest) and FP values (lowest) are highlighted in a black, bold font, and the worst TP values (lowest) and FP values (highest) are highlighted using a grey, bold font.

Table 6.16: Cut-off values used in the New and Traditional methods.

	$n = 100$				$n = 200$				$n = 500$			
	$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
New	0.106	0.106	0.091	0.091	0.049	0.049	0.038	0.038	0.019	0.019	0.014	0.014
Traditional	0.877	0.877	0.963	0.963	0.873	0.873	0.959	0.959	0.871	0.871	0.957	0.957

Table 6.17: Average of the JaB method cut-off values for the various methods across 100 generated data sets (estimated variances are stated in parentheses below each entry).

		$n = 100$				$n = 200$				$n = 500$			
		$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
		$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
N	All	0.084 (0.000)	0.166 (0.000)	0.065 (0.000)	0.104 (0.000)	0.010 (0.000)	0.051 (0.000)	0.012 (0.000)	0.032 (0.000)	0.003 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
	Delta	0.040 (0.054)	0.169 (0.000)	0.057 (0.000)	0.101 (0.000)	0.010 (0.000)	0.012 (0.000)	0.012 (0.000)	0.017 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
T3	All	0.092 (0.000)	0.164 (0.000)	0.068 (0.000)	0.110 (0.000)	0.010 (0.000)	0.056 (0.000)	0.012 (0.000)	0.036 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
	Delta	0.047 (0.000)	0.167 (0.000)	0.059 (0.000)	0.106 (0.000)	0.010 (0.000)	0.018 (0.000)	0.012 (0.000)	0.021 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
LN	All	0.102 (0.000)	0.163 (0.000)	0.070 (0.000)	0.112 (0.000)	0.010 (0.000)	0.061 (0.000)	0.013 (0.000)	0.038 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
	Delta	0.058 (0.000)	0.166 (0.000)	0.059 (0.000)	0.109 (0.000)	0.010 (0.000)	0.023 (0.000)	0.013 (0.000)	0.023 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
LP	All	0.087 (0.000)	0.163 (0.000)	0.067 (0.000)	0.107 (0.000)	0.011 (0.000)	0.051 (0.000)	0.013 (0.000)	0.033 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
	Delta	0.045 (0.000)	0.166 (0.000)	0.058 (0.000)	0.103 (0.000)	0.011 (0.000)	0.014 (0.000)	0.013 (0.000)	0.018 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
CN1	All	0.081 (0.000)	0.165 (0.000)	0.065 (0.000)	0.103 (0.000)	0.009 (0.000)	0.051 (0.000)	0.012 (0.000)	0.032 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
	Delta	0.038 (0.000)	0.169 (0.000)	0.056 (0.000)	0.099 (0.000)	0.010 (0.000)	0.012 (0.000)	0.012 (0.000)	0.017 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
CN2	All	0.088 (0.000)	0.165 (0.000)	0.068 (0.000)	0.106 (0.000)	0.011 (0.000)	0.053 (0.000)	0.013 (0.000)	0.033 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
	Delta	0.044 (0.000)	0.168 (0.000)	0.059 (0.000)	0.102 (0.000)	0.011 (0.000)	0.015 (0.000)	0.013 (0.000)	0.018 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)

Table 6.18: Average of the Combined method cut-off values for the various methods across 100 generated data sets. The initial check stage cut-off values appear in the row labelled “Init.” and the confirmation stage cut-off values appear in the row labelled “Conf.” (estimated variances are stated in parentheses below each entry).

			$n = 100$				$n = 200$				$n = 500$			
			$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
			$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
N	Init.	All	2.030 (0.168)	2.997 (0.087)	4.408 (0.080)	4.936 (0.116)	0.187 (0.000)	0.635 (0.063)	0.364 (0.001)	0.812 (0.022)	0.033 (0.000)	0.028 (0.000)	0.043 (0.000)	0.040 (0.000)
		Delta	3.188 (0.172)	2.961 (0.086)	4.762 (0.131)	4.929 (0.119)	0.188 (0.000)	1.220 (0.013)	0.368 (0.001)	1.183 (0.008)	0.032 (0.000)	0.028 (0.000)	0.042 (0.000)	0.040 (0.000)
	Conf.	All	0.082 (0.000)	0.166 (0.000)	0.065 (0.000)	0.103 (0.000)	0.010 (0.000)	0.051 (0.000)	0.012 (0.000)	0.032 (0.000)	0.003 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
		Delta	0.040 (0.000)	0.169 (0.000)	0.056 (0.000)	0.100 (0.000)	0.010 (0.000)	0.012 (0.000)	0.012 (0.000)	0.017 (0.000)	0.003 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
T3	Init.	All	2.044 (0.249)	3.033 (0.090)	4.478 (0.049)	4.902 (0.087)	0.191 (0.001)	0.658 (0.062)	0.372 (0.001)	0.837 (0.026)	0.032 (0.000)	0.028 (0.000)	0.043 (0.000)	0.041 (0.000)
		Delta	3.125 (0.251)	2.998 (0.091)	4.829 (0.074)	4.893 (0.090)	0.193 (0.001)	1.213 (0.015)	0.376 (0.001)	1.192 (0.012)	0.031 (0.000)	0.028 (0.000)	0.043 (0.000)	0.041 (0.000)
	Conf.	All	0.091 (0.000)	0.163 (0.000)	0.068 (0.000)	0.108 (0.000)	0.010 (0.000)	0.055 (0.000)	0.012 (0.000)	0.036 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
		Delta	0.048 (0.000)	1.664 (0.000)	0.058 (0.000)	0.105 (0.000)	0.010 (0.000)	0.018 (0.000)	0.012 (0.000)	0.021 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
LN	Init.	All	2.212 (0.181)	3.008 (0.088)	4.565 (0.130)	4.875 (0.108)	0.190 (0.001)	0.684 (0.064)	0.383 (0.001)	0.866 (0.020)	0.031 (0.000)	0.027 (0.000)	0.043 (0.000)	0.041 (0.000)
		Delta	3.106 (0.152)	2.972 (0.088)	4.904 (0.163)	4.863 (0.111)	0.193 (0.001)	1.218 (0.012)	0.391 (0.001)	1.180 (0.009)	0.030 (0.000)	0.027 (0.000)	0.042 (0.000)	0.041 (0.000)
	Conf.	All	0.102 (0.000)	0.162 (0.000)	0.070 (0.000)	0.114 (0.000)	0.010 (0.000)	0.061 (0.000)	0.013 (0.000)	0.038 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
		Delta	0.058 (0.000)	0.166 (0.000)	0.058 (0.000)	0.110 (0.000)	0.010 (0.000)	0.024 (0.000)	0.013 (0.000)	0.023 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
LP	Init.	All	2.012 (0.015)	3.015 (0.087)	4.439 (0.070)	4.964 (0.074)	0.194 (0.001)	0.622 (0.061)	0.373 (0.001)	0.833 (0.020)	0.033 (0.000)	0.028 (0.000)	0.045 (0.000)	0.040 (0.000)
		Delta	3.110 (0.123)	2.979 (0.087)	4.795 (0.096)	4.959 (0.075)	0.195 (0.001)	1.218 (0.010)	0.377 (0.001)	1.188 (0.010)	0.032 (0.000)	0.028 (0.000)	0.044 (0.000)	0.040 (0.000)
	Conf.	All	0.086 (0.000)	0.164 (0.000)	0.067 (0.000)	0.104 (0.000)	0.011 (0.000)	0.053 (0.000)	0.013 (0.000)	0.033 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
		Delta	0.045 (0.000)	0.167 (0.000)	0.058 (0.000)	0.101 (0.000)	0.011 (0.000)	0.014 (0.000)	0.013 (0.000)	0.019 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
CN1	Init.	All	1.989 (0.171)	3.077 (0.119)	4.468 (0.100)	4.948 (0.074)	0.184 (0.001)	0.590 (0.059)	0.362 (0.001)	0.794 (0.020)	0.032 (0.000)	0.027 (0.000)	0.043 (0.000)	0.040 (0.000)
		Delta	3.123 (0.146)	3.043 (0.119)	4.828 (0.166)	4.940 (0.076)	0.186 (0.001)	1.218 (0.015)	0.366 (0.001)	1.187 (0.007)	0.031 (0.000)	0.027 (0.000)	0.042 (0.000)	0.040 (0.000)
	Conf.	All	0.084 (0.000)	0.166 (0.000)	0.065 (0.000)	0.101 (0.000)	0.009 (0.000)	0.052 (0.000)	0.012 (0.000)	0.033 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)	0.002 (0.000)
		Delta	0.038 (0.000)	0.169 (0.000)	0.056 (0.000)	0.098 (0.000)	0.010 (0.000)	0.012 (0.000)	0.012 (0.000)	0.017 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
CN2	Init.	All	2.023 (0.168)	3.035 (0.090)	4.443 (0.071)	4.943 (0.098)	0.193 (0.001)	0.591 (0.050)	0.368 (0.001)	0.814 (0.020)	0.034 (0.000)	0.029 (0.000)	0.045 (0.000)	0.040 (0.000)
		Delta	3.159 (0.147)	3.000 (0.089)	4.807 (0.112)	4.936 (0.101)	0.194 (0.001)	1.219 (0.011)	0.371 (0.001)	1.174 (0.009)	0.033 (0.000)	0.029 (0.000)	0.044 (0.000)	0.040 (0.000)
	Conf.	All	0.088 (0.000)	0.165 (0.000)	0.067 (0.000)	0.107 (0.000)	0.011 (0.000)	0.053 (0.000)	0.013 (0.000)	0.033 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
		Delta	0.044 (0.000)	0.168 (0.000)	0.059 (0.000)	0.103 (0.000)	0.011 (0.000)	0.015 (0.000)	0.013 (0.000)	0.018 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)

Table 6.19: Median run-time (in seconds) over the different error distribution data sets for each influential diagnostic method.

	$n = 100$				$n = 200$				$n = 500$			
	$p = 5$		$p = 15$		$p = 5$		$p = 15$		$p = 5$		$p = 15$	
	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$	$\delta = 5$	$\delta = 10$
JaB	15.16	14.92	21.01	21.02	32.86	32.76	45.81	45.77	579.21	928.43	1012.90	979.31
ADAP	3.69	3.66	3.64	3.63	4.34	4.35	5.65	5.40	8.60	8.55	14.53	14.48
Combined	202.92	159.69	315.60	320.64	468.27	370.89	787.09	716.56	1626.89	1632.12	3138.51	3468.11
New	0.35	0.42	0.38	0.41	0.43	0.43	0.43	0.47	0.44	0.43	0.42	0.41
MC^3	0.35	0.32	0.48	0.44	2.94	2.52	3.28	2.93	20.78	14.56	22.16	18.35

Note that Intel Core™ i7-6700 PCs were used with 16GB of RAM.

6.4 Discussion

We now present the discussion on the results given in Tables 6.4 to 6.15 of the previous section. These tables all include information about how well each of the six methods are able to detect the influential values inserted into simulated data sets, and so the discussion will center on comparing the performance of these methods to one another in each of the various simulation settings considered in the study. Some comments regarding the calculated cut-off values (presented in Tables 6.16, 6.17, and 6.18) for these methods will also be provided, along with a brief look at the execution time of each method (given in Table 6.19).

6.4.1 Results when influential values are inserted using the *simple* method

We will start by looking at Tables 6.4 to 6.9, where influential values were added into the data sets by making use of the *simple* method described in Section 6.1. The first set of results that will be looked at are those given in Table 6.4, that were obtained by applying the six methods to the data sets generated with normal errors. Tables 6.5 to 6.9 will then be compared to Table 6.4, that is, Table 6.4 is used as a base-line for the other tables.

Data generated with normal errors: From Table 6.4, it can be seen that the Traditional method results in some of the worst performance in this study when looking at the TP values. In most cases, the Traditional method identifies less than 50% of the inserted simulated influential observations. The only cases in which these TP values are a little bit better, are when both p and δ are small. Furthermore, for these specific cases the TP values seem to increase as n increases. When looking at the FP values, on the other hand, the opposite performance is noticed, as this method produces the best FP performance in the study. From these values it can be seen that no non-influential observation is incorrectly flagged as influential.

Moving on to the results of the JaB method, it is noticed that this method performs relatively well when looking at the TP values, although it is not the best. Additionally, it seem like the JaB method struggles a little bit in terms of TP when n is small and δ is large. In general, it seems that the JaB method performs better for cases with lower

values of p and that the TP values increase a little as n increases. When looking at the FP values, it is clear that the JaB method struggles a lot when n is large, with FP values greater than 7 and 11 on average for δ equal to 5 and 10, respectively.

The next method that will be discussed is the ADAP method. This method produces the best performances in terms of TP values except for those cases where $p = 15$ and $\delta = 5$, although there is a general tendency for the TP values to improve as n increases. The method has the worst performance overall when $n = 100$, $p = 15$, and $\delta = 5$, where we find that the TP indicates that the method was only able to detect, on average, about 46% of the simulated influential observations. However, the ADAP method results in some of the best performances in terms of FP values when compared to all of the other methods, except for the Traditional method.

From the results of the Combined method, it is noted that the performance of this method is worse for larger δ values than for smaller δ values. This result is the most noticeable when n is small, since in these cases, the Combined method was only able to identify on average between 28% and 32% of the simulated influential observations, however, the rest of the TP values seem very good. Note that the TP values for this method produce the best results (compared to the other methods) in the case where $n = 500$. From the FP values of the Combined method, we see that the performance of the method decreases as n increases and for n large, very bad performances is obtained, except for the case where $p = 15$ and $\delta = 10$.

Considering now the results related to the newly proposed method (denoted “New” in the table) we find that the TP values seem to increase as n increases, however, they are still not perfect. Additionally, the performance of this method seems to decrease mildly in terms of TP when there is an increase in p , especially when δ is large as well. From the FP values of the New method, overall good performance is obtained with a very small decrease in performance as n increases.

The MC^3 method’s performance in terms of TP , increases as n increases. For large n the TP values are very good, however, when n is small the MC^3 method does not perform that well, specifically when δ is large. The FP values of this method, on the other hand, are very good overall.

When looking at the variances of all the TP values in Table 6.4, it seems that there is an overall decrease in the variances as n increase. The variances of the FP values, on the other hand, will generally increase (or decrease) as the FP values increase (or decrease).

Next, we will compare the results in Tables 6.5 to 6.9 to the results in Table 6.4, and discuss the differences between these different settings.

Data generated with t -distributed errors: The first table that will be compared to the normal case is Table 6.5, which consists of the results obtained when the six methods were applied to those data sets generated with t -distributed error terms. All the result

in this table seem almost the same as in the normal case, except for the following few differences: No noticeable changes have occurred in the results of the Traditional method, and its performance is still very poor in terms of the TP values, although it has very good FP values. In the results of the JaB method, there is a general, but slight, improvement in performance when looking at the TP and FP values, however, the FP improvement is more significant, since in some cases the FP values have reduced by more than a half. The JaB method's performance is therefore better than in the normal case, but it still has the property that it performs poorly in terms of FP when n is large. Moving on to the ADAP method, we see that its performance is still very good, although there is a slight decrease when looking at the TP values in the case of large n . In terms of FP , very good performance occurred as well, with a negligible amount of errors occurring in cases where n is small. For the Combined method, similar results were obtained as in the normal case, and again bad performance occurred for those cases where n is small and δ large. For the FP values, on the other hand, a slight improvement in the performance of this method occurred for large n . The performance of the New method, in terms of TP values has improved for small and large n , where in the case of large n , this method has identified *all* the simulated influential observations. In terms of the FP values of the New method, no obvious differences have occurred relative to Table 6.4, except for a mild improvement in the case when n , p , and δ are all large. For the last method, that is, the MC^3 method, a marginally improvement can be seen in terms of TP when $n = 100$, however as n increases a decrease in performance of this method is found. From the FP values of this method, it can be seen that an overall decrease in performance has occurred, especially in the case of large n , where we find that this decrease is substantial.

Data generated with log-normal errors: Now we will provide a discussion on the comparison between the results obtained in the normal case and the results obtained in the case where the errors were generated from the log-normal distribution. These results can be found in Table 6.6. The only significant differences between these two sets of results are as follows: A slight overall improvement in performance can be seen when looking at the TP values of the Traditional method, however, they are still very poor. In the results of the JaB method, again an overall improvement can be seen in the TP values, where some of these values are amongst the best TP values of this table. Also, just as in the t -distributed case, there is a big improvement in the FP values relative to the normal case, although in the cases where n is large, the FP performance of the JaB method is still the worst overall. For the ADAP method, an interesting result was found in the TP results in that, when these values are compared to those in the normal case, it appears that there is an improvement in performance in the case of small δ values, however, when δ is large, the performance has degraded somewhat. When looking at the FP values of the ADAP method, only a very small loss in performance was again observed. Moving on to the Combined method's TP values, some loss in performance can be seen in most of the cases where n is small, but, on the other hand, an improvement was observed in the FP value of this method in the cases where n is large. For the New method, an overall increase in performance can be seen in terms of the TP values, however, the FP values of this method are comparable to the corresponding values in the normal case. In the results of the MC^3 method, a few differences have occurred, for example, when n is equal to 100

and 200 a great improvement can be seen. In some of these cases, the TP values have doubled. However, for large n very poor performance is obtained, especially when both p and δ are also large. Furthermore, this log-normal setting has some of the worst FP values of the MC^3 method.

Data generated with Laplace errors: The results obtained from the Laplace distribution, can be found in Table 6.7. Once again, we note that the results in this table are relatively similar to those in Table 6.4, and the most noticeable of these deviations are as follows: Minor improvements can be seen in the TP values of the Traditional method, however, the FP values are similar to the normal case. For the JaB method, the same strange occurrence in the TP values that occurred in the log-normal case can be seen here as well, where an improvement in performance occurred in case where δ is small and a decrease in performance occurred where δ is large. Again, the FP values of this method improved for this error distribution as well, especially in those cases where n and δ are large. Furthermore, when looking at the TP values of the ADAP method, it is noticed that for those cases where both p and δ are large, the ADAP method again under-performs. For these cases, the ADAP method is only able to identify on average about 50% of the simulated influential observations. In the FP values of the ADAP method, as well as the TP and FP values of the Combined method, no significant deviations from the normal case occurred. For the New method, an overall improvement in performance has occurred in terms of the TP values. In the cases where n is large, the New method was able to detect all the simulated influential observations, and we also note that the TP and FP values of this method (for large n values) have some of the best values among the other methods. Additionally, this method has the best performance over all the other methods in the case where n is small and both p and δ are large. Lastly, very good performance in terms of the FP values occurred overall for the New method as well. For the MC^3 method, an overall improvement in the TP values can be seen for $n = 100$ and $n = 200$, however, for $n = 500$ the performance has deteriorated, especially in the cases where δ is large. From the FP values, the performance of this method also degraded and has a large loss in performance when $n = 500$.

Data generated with contaminated normal errors: The last two tables that will be discussed are the two tables containing the results of the six methods applied to the data sets generated with contaminated normal error terms. Table 6.14 contains the results of the first contaminated normal distributed error data sets (i.e., the contamination of two normal distributions with different means, but the same variances, producing a bi-modal distribution). From this table no significant differences are seen when compared to the normal case.

Table 6.15, on the other hand, contains the results of the second contaminated normal distributed error data sets (i.e., the contamination of two normal distributions with the same means, but different variances). Once again, very little has changed from the normal case, except for the following few minor differences: A marginally decrease in performance can be seen for large n values in terms of the FP values of the New method as well as in the Combined method.

6.4.2 Results when influential values are inserted using the *masked* method

In this sub-section, the ability of the different methods to detect *masked* influential values in data sets will be discussed. The first set of results that will be considered are those that were obtained when the six methods were applied to the data sets generated with normal errors, found in Table 6.10. This table will then be used as a base-line for comparison to the results obtained when other distributions were used for the error, namely Tables 6.11 to 6.15. Note that, in all these tables, the Traditional method performed very poorly, since it was unable to detect a single influential observation over all of the data sets and all settings considered. The reason why this method performed so poorly is because of the way in which the δ influential values were inserted into these data sets. As mentioned previously, these data sets were generated in such a way that single-case deletion methods would *not* easily be able to find any inserted influential observations. The Traditional method, however, produced very good false positive rates since no observations were incorrectly flagged as influential. Since the Traditional method produced the same results in Table 6.10 to Table 6.15, no further attention will be given on this method when discussing these tables.

Data generated with normal errors: From Table 6.10, it can be seen that the JaB method performs very well when looking at the TP values, except in those cases where n is small, but δ is large. In these cases, that is, where $n = 100$, and δ is equal to 5 and 10, the JaB method flagged, on average, 7.76 and 8.75 out of the 10 inserted influential observations, respectively, which is a relatively good performance. Furthermore, from the FP values of the JaB method it can be seen that, as n increases, these values increase as well. Therefore, the performance of the JaB method noticeably deteriorates, since it flags more non-influential observations as influential as n increases. The JaB method especially suffers in those cases where $n = 500$ in which it flagged, on average, 11.49 non-influential observations as influential. Note also that the variance of the FP values also increases drastically when n increases. Lastly, a strange tendency was spotted in the FP values of the JaB method, where the JaB method performed worse for smaller δ cases than for the larger ones.

The ADAP method's performance is an improvement on the JaB method. For this method, when looking at its TP values, it seems that the performance increases a little as n increases, but their variances also increase. Furthermore, it seems that the ADAP method only loses some performance in terms of TP in those cases where both p and δ are large. When investigating the FP values of the ADAP method, on the other hand, it was noted that the ADAP method works very well and suffers no noticeable problems, meaning that no non-influential observations were incorrectly flagged as influential.

Turning attention to the Combined method, we note that the TP values perform poorly in those cases where n is small, except in the case where $p = 5$ and $\delta = 5$. For the cases where n is large, on the other hand, this method's TP values seem relatively good. Additionally, it seems that over all n , except for $n = 500$, the method suffers when δ is large, producing

lower TP values in these cases. The variances over the TP values are also worse when n is small, but both the TP values and their variances seems to improve with the increase of n . When comparing the TP values of the Combined method to those of the JaB method, it seems like the JaB method performs better than the Combined method, although, when looking at their FP values the opposite is true. From the FP values of the Combined method and their corresponding variances, it can be seen that this method's performance worsens as n increases. However, when n , p , and δ is large, it is noted that the Combined method performs better than the ADAP method in terms of TP , and it performs better than the JaB method in terms of FP .

The next method that will be discussed is the New method which gave very good results in this simulation. From this method's TP values it can be seen that it performs very well, although it displays slightly worse performance in cases where n is small. Very good performance was also obtained in terms of the FP values, with very few data sets producing FP values greater than zero and then only in those settings where n is small.

The last method that was applied to the data sets is the MC^3 method. This method performed very well in terms of its TP values (even better than the ADAP method) for both small and large n , as well as for all combinations of n , p , and δ . However, by looking at the MC^3 method's FP values, it seems that it does not perform that well. Many of the simulation settings produced some data sets where the FP values were greater than zero (albeit very small deviations from zero), with the variance of these FP values being relatively small, but increasing with sample size.

Now we will start comparing Tables 6.11 to 6.15 with Table 6.10.

Data generated with t -distributed errors: The first table that will be discussed is Table 6.11, which consists of the results obtained on the data sets with error terms generated from the t_3 distribution. From this table, everything seems almost the same as in the normal case, except for the following few differences: There is a slight increase in the FP values of the JaB method relative to the corresponding results in Table 6.10, especially for large n , but a slight reduction in the variances of these values occurred. For the ADAP method, a marginal increase occurred in the FP values as well as their corresponding variances, but only for those cases where n is small. The same tendency as in the normal case has occurred in the FP values of the Combined method, but is slightly higher, whereas the variances of these values decreased slightly. For the New method, the FP values and their variances are somewhat worse than in the normal case and they increase a little as n increase. The MC^3 method has taken a big knock in terms of TP values (dropping significantly) as well as their variances (increasing significantly), for those cases where n is large. Furthermore, the FP values and their variances also greatly increased for large n , indicating a drastic deterioration of performance when compared to the results from the normal error data sets in Table 6.10.

Data generated with log-normal errors: The results obtained from the log-normal distribution, can be found in Table 6.12. Once again, we note that the results in this table

are somewhat similar to those in Table 6.10, and only some small differences occurred. The most noticeable of these deviations from Table 6.10 are as follows: The performance of the JaB method has slightly decreased in terms of the FP values, now producing slightly larger values than in the normal case. A slight decrease in performance of the ADAP method has also occurred when considering the FP values and their variances. Similarly, the FP values for the Combined method have also increased, but their corresponding variances have slightly decreased. For the JaB, ADAP and Combined methods, the changes in the FP values and their variances occurs mostly in the cases where n is large. Additionally, it must also be noted that the strange event that occurred in the JaB method in Table 6.10, where poorer performance for smaller δ values was obtained, occurred here for the JaB method as well. However, for the New method the performance did not decrease appreciably, which can be seen when looking at the FP values and their variances; again with the same tendency where the FP performance mildly degrades with an increase in sample size. Interestingly, we now observe that the same strange tendency of poor performance for lower δ values (observed previously in the JaB method) occurs for the new method for this data. From the results of the MC^3 method, it was noted that the TP values again degrade when n is large, and that the corresponding variances of these TP values also increases. Lastly, the FP values and their variances increased overall relative to the normal case, but for $n = 500$, we see a sharp increase, even worse than in the case of the t -distributed error data sets.

Data generated with Laplace errors: Now we will provide a discussion on the comparison between the results obtained in the normal case to those obtained in the case where the errors were generated from the Laplace distribution. The results of the Laplace error distributed data sets can be found in Table 6.13 and the only significant differences between these results and those obtain in the normal case are as follows: The results in this table show that the performance of the JaB method has somewhat degraded in terms of its FP values, with the worst FP values occurring when n is large. For the ADAP method, the difference in performance between the normal and Laplace data sets is minimal and there is nothing worth taking note of. For the Combined method, on the other hand, a very small increase has occurred in the FP values for these data sets, but only in those cases where n is large. Furthermore, very small losses of the performance of the New method occurred relative to the FP values as well as their variances. The same results as in the two previous non-normal cases were obtained again for the MC^3 method, that is, the method suffers when sample sizes are large in terms of TP and, especially, in terms of FP and their variances.

Data generated with contaminated normal errors: In Table 6.14, the results of the first contaminated normal distributed error data sets can be found (i.e., the contamination of two normal distributions with the different means, but the same variances, producing a bi-modal distribution). From this table it can be seen that the only big differences from the normal case is as follows: The JaB method actually performed better than in the normal case when looking at the FP values (but only mildly) and this occurrence is only significant when n is large, although the opposite is true when considering the variances of FP . Therefore, these variances are a little bit worse than in the normal case and increase

with increasing sample size. Similar to the Laplace case of Table 6.13, no obvious changes have occurred for the ADAP, Combined and the New method relative to the normal error data sets. Lastly, the MC^3 method performed better than the normal case when looking at the FP values and their variances, and it did not suffer as significant a drop in performance as in the previous non-normal cases when n is large.

Lastly, Table 6.15 contains the results of the performance of the 6 methods applied to the data sets with error terms generated from the second contaminated normal distribution (i.e., the contamination of two normal distributions with the same means, but different variances). Once again, not much has changed from the normal case, except for the following: The performance of the JaB method seems to be a little bit worse overall when looking at the FP values. The performance of the ADAP method was largely unchanged in terms of both TP and FP , but the performance of the Combined method, on the other hand, has decreased minimally in terms of FP for large n . The FP values of the New method overall are a little bit higher than those in the normal case, but only marginally so. Lastly, the performance of the MC^3 method is much worse than the normal case. The TP values decrease as n increases and the FP values and their variances, together with the variances of TP , increase as n increases. These quantities have the same tendency as in the previous cases, where they suffer a lot in those cases where n is large.

6.4.3 Comparison between the results of the two different ways in which the influential observations were added to the data

When comparing Tables 6.4 to 6.9 (the tables containing the results of the six methods applied to the data sets having inserted influential observations generated using the *simple* method), to Tables 6.10 to 6.15 (tables containing the results of the methods applied to the data sets having inserted influential observations generated using the *masked* approach), it is noted that the only two methods that have worse performance in the *masked* data sets, are the Traditional method and the MC^3 method. This might be an indication that these two methods struggle when masking is present in the data. Furthermore, since the *masked* approach is an artificial method, meaning that it inserted influential observations in such a way that masking is present in the data, it might happen that some methods seem to perform better than they really should. Lastly, since the ways in which the influential observations were added into the data differs greatly in the two sets of tables, it makes it difficult to compare these tables' results with one another and therefore, no further detail will be given here.

6.4.4 Cut-off values for each method

Tables 6.16 to 6.18 were provided for the sake of completeness and simply present the calculated cut-off values of the different methods for each of the different settings of n , p and δ .

In Table 6.16, the cut-off values for both the Traditional method and the New method are provided and, since the cut-off values for these two methods are not directly based on the data, it was not necessary for us to determine these cut-off values separately for each different data set using each different error distribution. From this table, it is noted that

the cut-off values for the New method decreases as n increases and that smaller cut-off values occur for larger p values. The Traditional method's cut-off values do not seem to be affected by n , but for larger p values on the other hand, these cut-off values seems to be larger than for smaller p values. When comparing these two methods' cut-off values with each other, it is clear that the New method has significantly smaller cut-off values than the Traditional method.

From Tables 6.17 and 6.18, no discernible patterns can be seen from the cut-off values, although it appears to be some unknown function of n , p , and δ . Furthermore, from Table 6.18 we can also see that, in general, the cut-off values obtained in the confirmation stage are a little bit smaller than the cut-off values obtained in the initial check stage.

6.4.5 Execution time for each method

To end our discussion, we will discuss the results stated in Table 6.19 which presents the execution times of each of the methods. From this table, it is clear that the Combined method is by far the most time consuming method, with a maximum time of 3924.41 seconds, i.e., 1.09 hours, for one simulation to execute. The next 'slowest' method is the JaB method. While this method is a lot faster than the Combined method, it still has a maximum time value of 2532.09 seconds, i.e., 42.20 minutes. Moving on to the next three methods, we find that there is a large jump in the execution times. The maximum execution time over the three methods that are left, are 36.67 seconds, which is a drastic improvement on the Combined and JaB methods. The method corresponding to this time value is the MC^3 method, which is followed by the ADAP method with a maximum execution time of 23.15 seconds. The last method, and therefore also the method that executed the fastest over all the methods, is the New method with an extraordinary fast executing maximum time value of only 0.61 seconds. The differences in execution times is therefore very large, ranging from the smallest maximum value of 0.61 seconds (corresponding to the New method) to the largest maximum time value of 1.09 hours (for the Combined method). Finally, we note from Table 6.19 that it seems like the execution times for all of the methods follow a non-linear progression over the sample sizes.

In the next chapter, we will now give our overall conclusions drawn from the results discussed in this chapter.

Chapter 7

Conclusion

In this study, we aimed to provide and investigate two new methods for the detection of influential observations, namely the Combined and the New method. The Combined method is based on the combination of the jackknife-after-bootstrap (JaB) method, proposed by Martin and Roberts (2010), and the adaptive, automatic multiple-case deletion technique (ADAP), proposed by Roberts et al. (2015). The New method, on the other hand, is based on a simple approximation of the distribution of Cook's distance. Additionally, we wanted to investigate the performance in detecting influential observations of the traditional Cook's distance influential diagnostic measure, the JaB method, the ADAP method, the MC^3 method, as well as the two new methods. Lastly, we wanted to compare the performance of all these methods to one another and make recommendations.

The procedure followed in this dissertation to develop the proposed Combined method involved making use of the ADAP method for the detection of influential observations and then altering it so that, instead of applying the simple large sample cut-off values of Cook's distance, the JaB method was instead implemented to generate more suitable data-driven cut-off values for both stages of ADAP. Furthermore, we were able to develop a new method (called simply 'New method') by simply approximating the cut-off value used for Cook's distances via a simple approximation of the Cook's distance's distribution. This results in a new classification rule for identifying values as being influential using Cook's distance. These methods' performances were investigated, as well the performances of the traditional Cook's distance and the three modern methods, namely, the JaB method, the ADAP method, and the MC^3 method, in a comprehensive simulation study and their performances were also compared to one another.

In our simulation study, data sets were generated with different sample sizes, n (including $n=100$, $n=200$ and $n=500$), and different number of predictor variables, p (including $p=5$ and $p=15$). These data sets were all generated using the linear regression model and five different distributions to generate the error terms, namely, the normal, t , log-normal, Laplace, and contaminated normal distributions. After these data sets were generated, δ (varying between 5 and 10) of the observations in the data sets were adapted in order for them to become influential. This was done using one of two methods. The first method used a simple, random way to insert influential observations into the data, such

that single-case deletion techniques would also be able to detect these values, or at least some of them. The second method, on the other hand, inserted influential observations into the data sets in such a way that masking occurs.

From the results given in Section 6.3, the plain ADAP method (i.e., the ADAP method using naïve cut-off values) seems to give the best overall performance for small sample sizes based on TP and FP values. For large sample sizes, the ADAP method also results in excellent FP values, and only a small loss in performance in TP values in those cases where the number of inserted influential observations is large. These two sets of results indicate that this is almost certainly the preferred method among all methods considered for detecting influential values, both in the case where influential values are masked as well as in the case when they are not. Unfortunately, the proposed improvement on this method, that is, the Combined method, did not produce stellar results at all; managing to only equal the ADAP method in terms of TP in a few settings, but performing markedly worse in terms of FP , in the majority of the large sample settings considered. However, the one setting where the Combined method was consistently the better method in terms of TP , is when the number of parameters in the linear regression model, sample size, and number of inserted influential values were all large. In addition to this relatively mediocre performance, this method also had a significantly longer run-time than the other methods, and so, while it produced useable results, we would not recommend it for general use. A possible reason for the poor performance can perhaps be attributed to the low value of the number of bootstrap replications used in the study. For possible future research and investigation on this method, one might look into improving the over-all run-time of the JaB and ADAP methods (and therefore the Combined method too), thereby allowing one to increase the number of bootstrap replications used in the jackknife-after-bootstrap method to select cut-off values in both the initial check stage and confirmation stage.

The New method proposed in this dissertation performed relatively well for small sample sizes in terms of TP and FP , although the performance was not as good as the ADAP and the MC^3 methods in these settings, but is definitely better than the Traditional method considered. Therefore, considering these results, we can rate this method as being the ‘next best’ method among the methods investigated. This New method also has the benefit of being much simpler and less time consuming to compute than the ADAP method, so it is recommended as the ‘quick-and-easy’ approach for decent results. For large sample sizes, the New method also demonstrated excellent TP values, even outperforming the ADAP method in many instances.

The JaB method has good TP values for large sample sizes, but suffers for small sample sizes, and it has moderate FP values when the sample sizes are small. However, this method has the worst overall FP performance when sample sizes are large. Therefore, we would potentially recommend the JaB method as the ‘3rd best’ choice for small sizes, and would not recommend it at all for larger sample sizes.

Lastly, under normally distributed error terms, one of the best methods is the MC^3 method, but even slight deviations from normality causes big problems in the performance

of the method when influential values are masked. Specifically, the method produced the worst FP values for smaller sample sizes, and very poor FP values in large sample sizes. Therefore, if there is reason to believe that the data are indeed generated from a normal error linear regression model, then this method will be desirable, but any deviations from normality make the results somewhat untrustworthy.

Bibliography

- ATKINSON, A. C. (1986). Masking unmasked. *Biometrika*, **73** (3), 533–541.
- BECKER, C. AND GATHER, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis*, **36** (1), 119–127.
- BELSLEY, D. A., KUH, E., AND WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, Inc., Hoboken, NJ.
- BEYAZTAS, U. AND ALIN, A. (2013). Jackknife-after-bootstrap method for detection of influential observations in linear regression models. *Communications in Statistics - Simulation and Computation*, **42** (6), 1256–1267.
- BEYAZTAS, U. AND ALIN, A. (2014). Delete-2 jackknife-after-bootstrap in regression. *Quality and Reliability Engineering International*, **30**, 993–1002.
- BEYAZTAS, U., ALIN, A., AND MARTIN, M. A. (2014). Robust BCa–JaB method as a diagnostic tool for linear regression models. *Journal of Applied Statistics*, **41** (7), 1593–1610.
- CHU, J. T. (1956). Errors in normal approximations to the t, τ , and similar types of distribution. *The Annals of Mathematical Statistics*, **27** (3), 780–789.
- COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19** (1), 15–18.
- COOK, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, **74** (365), 169–174.
- COOK, R. D. AND WEISBERG, S. (1982). *Monographs on Statistics and Applied Probability*, chapter Residuals and Influence in Regression. Chapman & Hall, New York.
- DALPOSSO, G. H., URIBE-OPAZO, M. A., AND JOHANN, J. A. (2016). Soybean yield modeling using bootstrap methods for small samples. *Spanish Journal of Agricultural Research*, **14** (3), 1–12.
- DAVISON, A. C. AND HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

- DEAKIN, R. E. AND HUNTER, M. N. (2018). Tau distribution and testing residuals. Technical report, School of Mathematical and Geospatial Sciences, RMIT University, Melbourne.
- EFRON, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, Series B*, **54** (1), 83–127.
- EFRON, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, **18** (2), 135–140.
- EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics*, **32** (2), 407–499.
- EFRON, B. AND TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton.
- FREEDMAN, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9** (6), 1218–1228.
- FRIENDLY, M., MONETTE, G., AND FOX, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, **28** (1), 1–39.
- GRAY, H. L. AND SCHUCANY, W. R. (1972). *The Generalized Jackknife Statistic*. Marcel Dekker Inc., New York.
- GREENE, W. H. (2003). *Econometric Analysis*. 5th edition. Prentice Hall, Upper Saddle River.
- HADI, A. S. AND SIMONOFF, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, **88** (424), 1264–1272.
- HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer, New York.
- HINKLEY, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, **19** (3), 285–292.
- HOERL, A. E. AND KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12** (1), 55–67.
- HOETING, J., RAFTERY, A. E., AND MADIGAN, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, **22** (3), 251–270.
- HOSSEINI, R. (2010). Quantiles equivariance. *eprint arXiv:1004.0533*.
- JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. (2015). *An Introduction to Statistical Learning with Applications in R*. 6th edition. Springer.
- JOHNSON, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics*, **23** (2), 49–54.

- KUTNER, M. H., NACHTSHEIM, C. J., NETER, J., AND LI, W. (2005). *Applied Linear Statistical Models*. 5th edition. McGraw-Hill/Irwin, New York.
- LAWRENCE, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society, Series B*, **57** (1), 181–189.
- MADIGAN, D., YORK, J., AND ALLARD, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, **63** (2), 215–232.
- MARTIN, M. A. AND ROBERTS, S. (2010). Jackknife-after-bootstrap regression influence diagnostics. *Journal of Nonparametric Statistics*, **22** (2), 257–269.
- MARUBINI, E. AND ORENTI, A. (2014). Detecting outliers and/or leverage points: A robust two-stage procedure with bootstrap cut-off points. *Epidemiology Biostatistics and Public Health*, **11** (3), 1–17.
- MILLER, R. G. (1974). The jackknife - A review. *Biometrika*, **61** (1), 1–15.
- MUÑOZ-PICHARDO, J. M., MORENO-REBOLLO, J. L., PINO-MEJÍAS, R., AND CUBILES DE LA VEGA, M. D. (2019). Influence measures in beta regression models through distance between distributions. *AStA Advances in Statistical Analysis*, **103**, 163–185.
- NURUNNABI, A. A. M., HADI, A. S., AND IMON, A. H. M. R. (2014). Procedures for the identification of multiple influential observations in linear regression. *Journal of Applied Statistics*, **41** (6), 1315–1331.
- NURUNNABI, A. A. M., IMON, A. H. M. R., AND NASSER, M. (2011). A diagnostic measure for influential observations in linear regression. *Communications in Statistics - Theory and Methods*, **40** (7), 1169–1183.
- PAN, J.-X., FUNG, W.-K., AND FANG, K.-T. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference*, **83** (1), 153–167.
- PEÑA, D. (2005). A new statistic for influence in linear regression. *Technometrics*, **47** (1), 1–12.
- PEÑA, D. AND YOHAI, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association*, **94** (446), 434–445.
- POLITIS, D. N. AND ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, **22** (4), 2031–2050.
- POPE, A. J. (1976). The statistics of residuals and the detection of outliers. NOAA Technical Report NOS65 NGS 1, Geodetic Research and Development Laboratory, National Geodetic Survey, Rockville.
- QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B*, **11** (1), 68–84.

- RAFTERY, A., HOETING, J., VOLINSKY, C., PAINTER, I., AND YEUNG, K. Y. (2019). *BMA: Bayesian Model Averaging*. R package version 3.18.10.
- RAFTERY, A. E., MADIGAN, D., AND HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92** (437), 179–191.
- RAHMATULLAH IMON, A. H. M. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*, **32** (9), 929–946.
- RAO, C. R. AND TOUTENBURG, H. (1995). *Linear Models: Least Squares and Alternatives*. Springer Series in Statistics. Springer, New York.
- ROBERTS, S., MARTIN, M. A., AND ZHENG, L. (2015). An adaptive, automatic multiple-case deletion technique for detecting influence in regression. *Technometrics*, **57** (3), 408–417.
- ROUSSEEUW, P. J. AND LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York.
- ROUSSEEUW, P. J. AND VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85** (411), 633–639.
- SEN, P. K. (1968). Estimates of the regression coefficient based on Kendall’s tau. *Journal of the American Statistical Association*, **63** (324), 1379–1389.
- SHAO, J. AND TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
- SHAO, J. AND WU, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, **17** (3), 1176–1197.
- THOMPSON, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *Annals of Mathematical Statistics*, **6** (4), 214–219.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58** (1), 267–288.
- TUKEY, J. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, **29** (2), 614–623.
- TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.
- VARIAN, H. (2005). Bootstrap tutorial. *The Mathematica Journal*, **9** (4), 768–775.
- WANG, T., LI, Q., ZANG, Q., AND LI, Z. (2019). A multiple-case deletion approach for detecting influential points in high-dimensional regression. *Communications in Statistics - Simulation and Computation*, **48** (7), 2065–2082.
- WU, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, **18** (3), 1438–1452.

- ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, **57** (298), 348–368.

Appendix A

Derivation of the least squares estimates

In this appendix, the following result will be derived:

The least squares regression estimates, $\mathbf{b}$, are given by

$$\underset{(p \times 1)}{\mathbf{b}} = \underset{(p \times p)}{(\mathbf{X}^\top \mathbf{X})^{-1}} \underset{(p \times n)}{\mathbf{X}^\top} \underset{(n \times 1)}{\mathbf{Y}}.$$

Note: This statement will be derived by making use of the linear regression model in (4.1).

The sum of the squared residuals that must be minimized is given by

$$Q = \underset{(1 \times n)}{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}^\top \underset{(n \times 1)}{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}. \quad (\text{A.1})$$

Using the fact that $\mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{Y}$, (A.1) can be written in the following form:

$$Q = \underset{(1 \times n)}{\mathbf{Y}^\top} \underset{((n \times 1))}{\mathbf{Y}} - 2 \underset{(1 \times p)}{\boldsymbol{\beta}^\top} \underset{(p \times n)}{\mathbf{X}^\top} \underset{(n \times 1)}{\mathbf{Y}} + \underset{(1 \times p)}{\boldsymbol{\beta}^\top} \underset{(p \times n)}{\mathbf{X}^\top} \underset{(n \times p)}{\mathbf{X}} \underset{(p \times 1)}{\boldsymbol{\beta}}.$$

Next, Q must be derived with respect to the regression parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$. These derivations result in the following:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}. \quad (\text{A.2})$$

Now, setting (A.2) equal to zero and solving for $\mathbf{b}$, the least squares estimates of the regression coefficients are obtained, that is,

$$\begin{aligned} -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\mathbf{b} &= 0 \\ \Rightarrow \mathbf{b} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \end{aligned}$$

Appendix B

Proofs of the properties of leverages

In this appendix, the following will be proved:

Two important properties of leverages are

$$0 \leq h_{ii} \leq 1, \quad \forall i = 1, 2, \dots, n, \quad (\text{B.1})$$

and

$$\sum_{i=1}^n h_{ii} = p. \quad (\text{B.2})$$

B.1 Proof of the bounds of leverages

Equation (B.1) will be proven using the following two properties of the hat matrix, $\mathbf{H}$ (defined as in (2.6)):

- $\mathbf{H}$ is idempotent, i.e.,

$$\begin{aligned} \mathbf{H}^2 &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \\ &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \\ &= \mathbf{H}. \end{aligned} \quad (\text{B.3})$$

- $\mathbf{H}$ is symmetric, i.e.,

$$\mathbf{H}^\top = \mathbf{H}. \quad (\text{B.4})$$

Proof of $h_{ii} \geq 0$. From (B.3), it can be concluded that $h_{ii} = (\mathbf{H}^2)_{ii}$ and then by using this statement together with (B.4), it follows that

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2. \quad (\text{B.5})$$

Note that the right hand side of (B.5) consists of the sum of two squared terms, therefore, $h_{ii} \geq 0$. $\square$

Proof of $h_{ii} \leq 1$. Again using B.5, note that the second term on the right hand side of this equation is the sum of positive values and from $h_{ii} - \sum_{j \neq i} h_{ij}^2 = h_{ii}^2$, it can be concluded that h_{ii} must be greater than, or equal to, h_{ii}^2 . The only way in which $h_{ii} \geq h_{ii}^2$, for $h_{ii} \geq 0$, is if $h_{ii} \leq 1$. $\square$

B.2 Proof that the sum of the leverages add up to p

To prove (B.2), the trace of the hat matrix $\mathbf{H}$ will be implemented.

Proof.

$$\begin{aligned}
 \sum_{i=1}^n h_{ii} &= \text{tr}(\mathbf{H}) \\
 &= \text{tr}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\
 &= \text{tr}(\mathbf{X} \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X})^{-1}) \\
 &= \text{tr}(\mathbf{I}_p) \\
 &= p.
 \end{aligned}$$

Here $\mathbf{I}_p$ is the identity matrix with dimensions equal to p , i.e,

$$\mathbf{I}_p = \underset{(p \times p)}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

$\square$

Appendix C

Proof of the variance of the regression residuals

In this appendix, the proof of the following statement will be provided:

The variance of the vector of regression residuals are given by

$$\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}),$$

where $\mathbf{I}$ is the identity matrix.

Before the proof will be provided, first note that $\mathbf{e}$ can be written in terms of the hat matrix, $\mathbf{H}$, as follows:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H}) \mathbf{Y}.$$

Furthermore, recall that $\text{Var}(Y_i) = \sigma^2, i = 1, 2, \dots, n$, leading to the following variance for $\mathbf{Y}$:

$$\text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}.$$

Proof of the variance of $\mathbf{e}$:

$$\begin{aligned} \text{Var}(\mathbf{e}) &= \text{Var}((\mathbf{I} - \mathbf{H})\mathbf{Y}) \\ &= (\mathbf{I} - \mathbf{H}) \sigma^2 \mathbf{I} (\mathbf{I} - \mathbf{H})^\top \\ &= \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^\top. \end{aligned}$$

Since $(\mathbf{I} - \mathbf{H})$ is symmetric, we now have

$$\text{Var}(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}),$$

and since, $(\mathbf{I} - \mathbf{H})$ is idempotent, we have

$$\text{Var}(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H}).$$

□

Appendix D

Proof of the alternative form of *CovRatio*

In Section 2.3.4, the *CovRatio* influence diagnostic measure was defined as follows:

$$CovRatio_i = \frac{s_{(i)}^{2p}}{s^{2p}} \left(\frac{\det(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}}{\det(\mathbf{X}^\top \mathbf{X})^{-1}} \right), \quad i = 1, 2, \dots, n. \quad (D.1)$$

Furthermore, an alternative form of *CovRatio* was also given and will be proved here, that is, we will prove the following:

The *CovRatio* influential diagnostic measure has the alternative form given by

$$CovRatio_i = \frac{1}{\left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p} \right)^p (1 - h_{ii})}, \quad (D.2)$$

where t_i is the studentized residual defined as follows:

$$t_i = \frac{e_i}{s_{(i)} \sqrt{1 - h_{ii}}}, \quad i = 1, 2, \dots, n. \quad (D.3)$$

In the proof of (D.2), we will be making use of two results given by Belsley et al. (1980), namely:

$$(n - p - 1)s_{(i)}^2 = (n - p)s^2 - \frac{e_i^2}{1 - h_{ii}} \quad (D.4)$$

and

$$\det(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}) = (1 - h_{ii}) \det(\mathbf{X}^\top \mathbf{X}). \quad (D.5)$$

Proof that (D.1) can be written as (D.2): First, (D.4) is rewritten in an alternative form that will be used later, given by

$$(n - p - 1) = (n - p) \frac{s^2}{s_{(i)}^2} - \frac{e_i^2}{s_{(i)}^2 (1 - h_{ii})}. \quad (D.6)$$

Note that the last term in D.6 is equal to the squared studentized residuals, therefore, D.6 can be written in the following reduced form:

$$(n - p - 1) = (n - p) \frac{s^2}{s_{(i)}^2} - t_i^2. \quad (\text{D.7})$$

With a few derivations, (D.7) becomes

$$\frac{s^2}{s_{(i)}^2} = \frac{(n - p - 1) + t_i^2}{n - p}. \quad (\text{D.8})$$

Also note that (D.5) can be written as follows:

$$\frac{\det(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}}{\det(\mathbf{X}^\top \mathbf{X})^{-1}} = \frac{1}{1 - h_{ii}},$$

and therefore, the *CovRatio*, defined as in (D.1), becomes

$$CovRatio = \frac{s_{(i)}^{2p}}{s^{2p}} \left(\frac{1}{1 - h_{ii}} \right). \quad (\text{D.9})$$

From (D.9), we get

$$\begin{aligned} \frac{1}{CovRatio} &= \left(\frac{s^2}{s_{(i)}^2} \right)^p (1 - h_{ii}) \\ &= \left(\frac{(n - p - 1) + t_i^2}{n - p} \right)^p (1 - h_{ii}). \end{aligned} \quad (\text{D.10})$$

The second step above was obtained by making use of (D.8). Now, it is easy to see that (D.1) can be written as in (D.2), that is,

$$CovRatio = \frac{1}{\left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p} \right)^p (1 - h_{ii})}.$$

□

Appendix E

Proof of the simple form of Cook's distance

In Section 2.3.5, Cook's distance was defined as follows:

$$D_i = \frac{\sum_{j=1}^n \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{ps^2},$$

and in matrix notation, it was given as

$$D_i = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} \right)^\top \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} \right)}{ps^2}. \quad (\text{E.1})$$

Additionally, a simple form of Cook's distance was given and will be proved here, that is, we will prove the following:

Cook's distance can be written in the following simple form:

$$D_i = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right), \quad (\text{E.2})$$

where r_i , $i = 1, 2, \dots, n$, are the standardized residuals defined as

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \dots, n. \quad (\text{E.3})$$

In order to prove (E.2), first note the following:

$$(\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - h_{ii}},$$

where $\mathbf{X}_{(i)}$ denotes the $((n - 1) \times p)$ design matrix constructed from the $(n \times p)$ design matrix $\mathbf{X}$ but with the i^{th} row deleted. Furthermore,

$$\begin{aligned}
\mathbf{b}_{(i)} &= (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{X}'_{(i)} \mathbf{Y}_{(i)} \\
&= \left[(\mathbf{X}' \mathbf{X})^{-1} + \frac{(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i \mathbf{X}'_i (\mathbf{X}' \mathbf{X})^{-1}}{1 - h_{ii}} \right] [\mathbf{X}' \mathbf{Y} - \mathbf{X}_i Y_i] \\
&= \mathbf{b} - \frac{e_i}{1 - h_{ii}} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i,
\end{aligned}$$

and therefore

$$\mathbf{b} - \mathbf{b}_{(i)} = \frac{e_i}{1 - h_{ii}} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i. \quad (\text{E.4})$$

The proof of the simple form of Cook's distance is given as follows:

Proof.

$$\begin{aligned}
D_i &= \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^\top (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{ps^2} \\
&= \frac{(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_{(i)})^\top (\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_{(i)})}{ps^2} \\
&= \frac{(\mathbf{X}(\mathbf{b} - \mathbf{b}_{(i)}))^\top \mathbf{X}(\mathbf{b} - \mathbf{b}_{(i)})}{ps^2} \\
&= \frac{(\mathbf{b} - \mathbf{b}_{(i)})^\top \mathbf{X}^\top \mathbf{X}(\mathbf{b} - \mathbf{b}_{(i)})}{ps^2}.
\end{aligned}$$

Substituting (E.4) into the above expression, we have

$$\begin{aligned}
D_i &= \frac{e_i^2}{ps^2(1 - h_{ii})^2} \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i \\
&= \frac{e_i^2 h_{ii}}{ps^2(1 - h_{ii})^2} \\
&= \frac{1}{p} \frac{e_i^2}{s^2(1 - h_{ii})} \frac{h_{ii}}{(1 - h_{ii})}.
\end{aligned}$$

Now, since the standardized residuals is defined as in (E.3), we obtain the final, simple form of Cook's distance, given by

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}. \quad (\text{E.5})$$

□

Appendix F

Proofs on restrictions for the New method

We will now provide the proofs for the restrictions stated for the sample size, n , in order to be able to calculate the cut-off values proposed in the New method. We will therefore provide proofs for the following:

The calculation of the cut-off value proposed in the New method must satisfy the following conditions on the sample size, n :

- $n \neq 0$ and $n \neq 1$,
- $n > \frac{(D^2+2)+\sqrt{D^2(D^2+4)}}{2}$, $D^2 > 0$,

where D_i^2 , the Mahalanobis distance between the i^{th} predictor vector $\mathbf{X}_i$ and the centroid of those predictor variables, $\bar{\mathbf{X}}$, given by:

$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}).$$

Proof. Recall that the only leverage values that are valid for the procedure of the New method to work, are in the interval $0 \leq h_{ii} < 1$, and so we must find the corresponding values of n for which the Mahalanobis distances would fall inside this interval. This can be done by solving for n in the following expression:

$$0 \leq \frac{D^2}{n-1} + \frac{1}{n} < 1. \tag{F.1}$$

Before we solve for n in (F.1), first note that n cannot be equal to 0 or 1. We now solve for n for the less than one condition and then, separately, for the greater than or equal to zero condition in (F.1).

The ‘less than one’ condition in (F.1): To solve for n in the first inequality, we note that

it can be written in the following form:

$$\begin{aligned}
& \frac{D^2}{n-1} + \frac{1}{n} < 1 \\
\Rightarrow & \frac{-n^2 + n(D^2 + 2) - 1}{n(n-1)} < 0 \\
\Rightarrow & \frac{n^2 - n(D^2 + 2) + 1}{n(n-1)} > 0.
\end{aligned} \tag{F.2}$$

For (F.2) to be true, we can have one of the following two cases

Case A: Both the numerator and the denominator of (F.2) are greater than 0.

Case B: Both the numerator and the denominator of (F.2) are less than 0.

Case A is therefore defined as the situation where both of the following two conditions hold:

$$n^2 - n(D^2 + 2) + 1 > 0 \tag{F.3}$$

and

$$n(n-1) > 0. \tag{F.4}$$

Case B on the other hand, is defined as the situation where both of the following two conditions hold:

$$n^2 - n(D^2 + 2) + 1 < 0 \tag{F.5}$$

and

$$n(n-1) < 0. \tag{F.6}$$

We start by first solving for n in $n^2 - n(D^2 + 2) + 1 = 0$; this can be done using the quadratic formula, which gives us the following solution:

$$\begin{aligned}
n &= \frac{(2 + D^2) \pm \sqrt{(2 + D^2)^2 - 4}}{2} \\
&= \left\{ \begin{array}{l} \frac{(2+D^2)+\sqrt{4D^2+D^4}}{2} \\ \frac{(2+D^2)-\sqrt{4D^2+D^4}}{2} \end{array} \right. .
\end{aligned} \tag{F.7}$$

Define now $n_1 := \frac{(2+D^2)-\sqrt{4D^2+D^4}}{2}$ and $n_2 := \frac{(2+D^2)+\sqrt{4D^2+D^4}}{2}$. We will now prove that n_1 and n_2 have some simple restrictions when $D^2 \geq 0$, namely:

$$0 \leq n_1 \leq 1 \tag{F.8}$$

and

$$n_2 \geq 1. \tag{F.9}$$

The proofs of the restrictions (F.8) and (F.9) are given below.

Proof that $n_1 \leq 1$ given that $D^2 \geq 0$:

$$\begin{aligned}
D^2 &\geq 0 \\
4D^2 &\geq 0 \\
4D^2 + D^4 &\geq D^4 \\
\sqrt{4D^2 + D^4} &\geq D^2 \\
-D^2 + \sqrt{4D^2 + D^4} &\geq 0 \\
D^2 - \sqrt{4D^2 + D^4} &\leq 0 \\
D^2 + 2 - \sqrt{4D^2 + D^4} &\leq 2 \\
\frac{D^2 + 2 - \sqrt{4D^2 + D^4}}{2} &\leq 1
\end{aligned}$$

$\therefore n_1 \leq 1$, if $D^2 \geq 0$. □

Proof that $n_1 \geq 0$ given that $D^2 \geq 0$: First note that

$$(D^2 + 2)^2 = D^4 + 4D^2 + 4.$$

Then we have

$$\begin{aligned}
(D^2 + 2)^2 = D^4 + 4D^2 + 4 &> D^4 + 4D^2 \\
D^2 + 2 &> \sqrt{D^4 + 4D^2} \\
D^2 + 2 - \sqrt{D^4 + 4D^2} &> 0 \\
\frac{D^2 + 2 - \sqrt{D^4 + 4D^2}}{2} &> 0.
\end{aligned}$$

$\therefore n_1 > 0$, if $D^2 > 0$.

Furthermore, if $D^2 = 0$, then $n_1 = 1$, however, if $D^2 \rightarrow \infty$, then $n_1 = 0$. □

These two proofs can also be seen in the Figure F.1, where we have plotted n_1 for different values of D^2 .

Proof $n_2 \geq 1$ given that $D^2 \geq 0$: From the formula of n_2 , it is clear that the numerator consists of some positive value ($D^2 + \sqrt{4D^2 + D^4}$) plus 2, and that the denominator is only equal to 2. Therefore, the numerator will *always* be greater or equal to the denominator for $D^2 \geq 0$, leading to the conclusion that $n_2 \geq 1$ when $D^2 \geq 0$. This can also be seen in Figure F.2 where we have plotted n_2 for different values of D^2 . □

Using the results (F.8) and (F.9), we can now develop the solutions for n in Case A and Case B.

Case A: We can now obtain solutions for n for the conditions in Case A. The inequality (F.3) would therefore hold if

$$n < n_1 \quad \text{or} \quad n > n_2,$$

and (F.4) would hold if

$$n < 0 \quad \text{or} \quad n > 1.$$

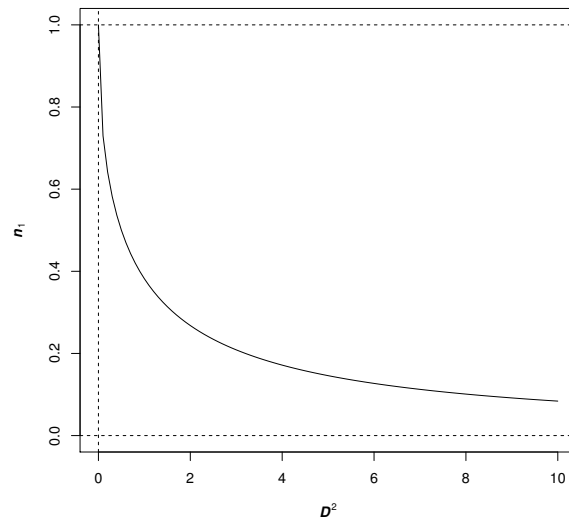


Figure F.1: Plot of n_1 against different values of D^2 to find the bounds of n_1 .

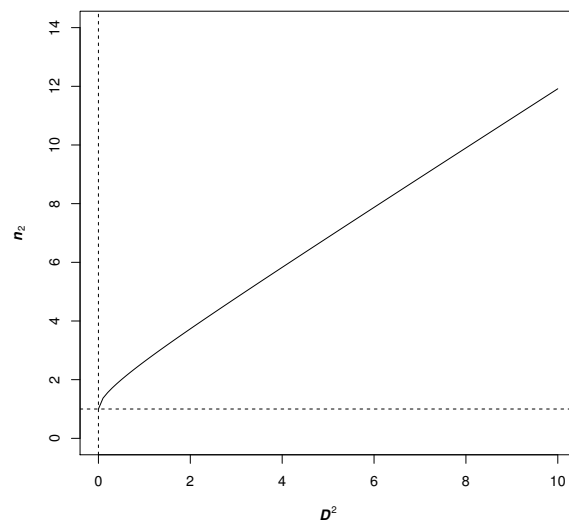


Figure F.2: Plot of n_2 against different values of D^2 to find the bounds of n_2 .

Further, the above inequalities can be re-written as

$$n < n_1 \quad \text{and} \quad n < 0, \quad (\text{F.10})$$

or

$$n < n_1 \quad \text{and} \quad n > 1, \quad (\text{F.11})$$

or

$$n > n_2 \quad \text{and} \quad n < 0, \quad (\text{F.12})$$

or

$$n > n_2 \quad \text{and} \quad n > 1. \quad (\text{F.13})$$

By making use of the results in (F.8) and (F.9), as well as the fact that the sample size must be positive, i.e., $n > 0$, we can simplify the conditions for the solutions for n in Case A given in (F.10) to (F.13): First of all, since $n > 0$, the conditions (F.10) and (F.12) can be removed entirely. Furthermore, since $n_1 \leq 1$ for all $D^2 \geq 0$, the condition (F.11) is an empty set and can be removed also. However, since $n_2 > 1$ for all $D^2 < 0$, (F.13) can be reduced to $n > n_2$. Therefore, for Case A, we have the following simple solution:

$$n > n_2 = \frac{(2 + D^2) + \sqrt{4D^2 + D^4}}{2}. \quad (\text{F.14})$$

Case B: The same can now be done for Case B. For equation (F.5) to hold, we have

$$n_1 < n < n_2,$$

and equation (F.6) would hold if

$$0 < n < 1.$$

Therefore, using the above conditions as well as the conditions stated in (F.8) and (F.9), the only solution for Case B is

$$\frac{(2 + D^2) - \sqrt{4D^2 + D^4}}{2} = n_1 < n < 1.$$

The ‘greater than or equal to zero’ condition in (F.1):

Again, first note that the greater than or equal to zero part in (F.1) can be written in the following form:

$$\begin{aligned} 0 &\leq \frac{D^2}{n-1} + \frac{1}{n} \\ 0 &\leq \frac{n(D^2 + 1) - 1}{n(n-1)}. \end{aligned} \quad (\text{F.15})$$

For (F.15) to be true, either one of the following two cases must hold:

Case C: Both the numerator and the denominator of (F.15) are greater than 0.

Case D: Both the numerator and the denominator of (F.15) are less than 0.

We will now briefly look at the solutions satisfy these two conditions separately.

Case C: Case C is defined as:

$$n(D^2 + 1) - 1 \geq 0 \Rightarrow n \geq \frac{1}{D^2 + 1} \quad (\text{F.16})$$

and

$$n(n - 1) \geq 0. \quad (\text{F.17})$$

Now, define $n_3 := \frac{1}{D^2 + 1}$, and note that $0 \leq n_3 \leq 1$ for $D^2 \geq 0$. Using this result and the conditions in (F.16) and (F.17), we get the following solution for Case C :

$$n \geq 1.$$

Case D: Case D on the other hand, is defined as:

$$n(D^2 + 1) - 1 \leq 0 \Rightarrow n \leq \frac{1}{D^2 + 1} \quad (\text{F.18})$$

and

$$n(n - 1) \leq 0. \quad (\text{F.19})$$

The solution is then simply:

$$0 \leq n \leq n_3 = \frac{1}{D^2 + 1}.$$

Since we know that $0 \leq n_3 \leq 1$, this particular case will not produce a feasible solution for n , except when $n = 1$. However, as stated earlier, the solution $n = 1$ is not valid. Therefore, Case D does not produce any feasible solutions to the problem and will be omitted from consideration in the remainder of the discussion.

We now have solutions to three different scenarios which are labelled as follows:

Case A solution: $A := \{n : n > n_2\}$.

Case B solution: $B := \{n : n_1 < n < 1\}$.

Case C solution: $C := \{n : n \geq 1\}$.

Taking all of the above solutions into account, we therefore have to find the values of n that satisfy the following expression:

$$(A \cup B) \cap (C) \Rightarrow (A \cap C) \cup (B \cap C).$$

Before we will obtain the solution, note that $n_3 \leq 1$ for all $D^2 \geq 0$. We now note the following results:

$$(A \cap C) = \{n : n > n_2\}, \text{ and}$$

$$(B \cap C) = \emptyset.$$

Therefore, our final solution for n is simply

$$n > n_2 = \frac{(D^2 + 2) + \sqrt{D^2(D^2 + 4)}}{2}, \quad D^2 \geq 0.$$

□