Using Finite Mixtures to Robustify Statistical Models

A thesis submitted in fulfilment
of the requirements for the degree
of
Doctor of Philosophy in Statistics
at
The University of Waikato

by

Maheswaran Rohan

February 2011

Abstract

This thesis is concerned with robust estimation of the parameters of statistical models. Although robust estimation is a very good idea, it has some shortcomings when seen from the statistical modelling point of view. For example, there are no easily applicable principles for creating robust estimates in new situations.

In this thesis, we are trying to introduce a unified method to obtain the robustified statistics for various situations such as the linear model and the generalized linear model. We wish to modify the maximum likelihood estimation procedure, which is very sensitive to the outliers. In order to reduce the effect on these estimates by outliers, we add an additional component, which would be of no interest but would contain all outliers, to the regular component forming a finite mixture. In fact, we use the finite mixture model to obtain a "robustified" estimate for a model parameter θ, where the finite mixture form is being used as a mathematical tool to have a tractable form of analysis rather than being used as a serious model for the data. We employ the EM algorithm to obtain the our proposed robustified estimates for the parameters. Our estimates are compared with some other estimates defined in the robust statistics literature.

This thesis examines the robustness of the proposed estimates using the concept of influence function. The estimates are defined iteratively, so that the implicit differentiation method of Jorgensen is used to obtain the influence functions of the estimates. We give example plots of these influence functions which are bounded.

In this thesis, we give mathematical results for all cases and we use well known real data sets to investigate our method. The statistical software **R** is used for all investigation. Finally, we hope that this method may give a unified approach for making parameter estimation in statistical models more robust.

Pre-Presentation of Parts of this Thesis

Some portions of the thesis are presented at statistics conferences. Abstracts for these presentations are available on line.

1. Rohan, M., (2008), New Approach to Robust Estimation, *NZSA conference 2008*
   Available at: http://www.stat.auckland.ac.nz/nzsa2008/.

2. Rohan, M., (2010), Using Finite Mixtures to Compute Robustified Statistics for Regression Parameters, *NZSA conference 2010*
   Available at: http://nzsa_cdl_2010.massey.ac.nz/AbstractBook.pdf.

I would like to thank to Department. of Conservation, where I am working as a full-time statistician, for covering the travelling cost to present these papers at the conferences.

Acknowledgements

I would like to express my deep and sincere thanks to my supervisor Dr. Murray Jorgensen, Department of Statistics, University of Waikato for his support and helpful discussion throughout my research. I would like to acknowledge to Dr Bill Bolstad, Deparment of Statistics, University of Waikato for his guidance during my studies.

I am very grateful to Department of Conservation managers Mr Ian Westbrooke, Dr Rod Hay and Dr Elaine Wright for giving their support and encouraging me to complete my Ph.D degree.

My sincere thanks to Mr Michael Ryan, Statistics New Zealand, Christchurch, and Ms Trish Casci Tribe, University colleague, for their valuable time in proof reading the original draft.

I am obliged to my parents Mr Veerasingam Maheswaran and Mrs Ratnawathy Maheswaran for their guidance and encouragement to complete this thesis. I thank them from the bottom of my heart and I am grateful to my sisters and brother for their support.

Finally, this thesis would not have been completed without the love, patience, and inspiration from my wife, Chelvi, and seven year old daughter, Ashwini. My sincere thanks to them, who rendered the greatest support, encouragement and tolerance during this journey.

Contents

1 Two Approaches to the Problem of Bad Data **5**
- 1.1 Diagnostic Statistics . . . 5
  - 1.1.1 Outliers . . . 6
- 1.2 Robust Statistics . . . 9
  - 1.2.1 Statistical Functionals . . . 10
- 1.3 Maximum Likelihood Estimators . . . 12
  - 1.3.1 The Exponential Family . . . 12
- 1.4 M-estimators . . . 14
  - 1.4.1 M-estimators of Location . . . 17
  - 1.4.2 M-estimators of Location and Scale . . . 19
  - 1.4.3 M-estimators of Regression . . . 20
  - 1.4.4 Generalized Linear Model . . . 23
- 1.5 Goal of the Thesis . . . 24
- 1.6 **R** Software . . . 24

2 Property of Robustness **27**
- 2.1 Influence Function . . . 27
- 2.2 Iteratively Defined Statistics . . . 29
- 2.3 Computing the Influence Function of Iteratively Defined Statistics . . . 31
  - 2.3.1 Jorgensen's Method . . . 31
  - 2.3.2 Jacobian Matrix . . . 32
  - 2.3.3 Example of the Jorgensen Method . . . 32
  - 2.3.4 Numerical Illustration . . . 34
- 2.4 Asymptotic Results for M-estimates . . . 35
- 2.5 Gross Error Sensitivity . . . 37

3 Mixture Models and the EM Algorithm **38**
3.1 Clustering . . . 38
3.2 Mixture Models . . . 39
3.3 Parameter Estimation . . . 40
3.4 The EM Algorithm . . . 41
3.4.1 Application of EM Algorithm to the Mixture Models . 41
3.4.2 Overview of the EM Algorithm . . . 43
3.4.3 Remark on Starting Values . . . 44
3.4.4 Information Matrices . . . 44
3.4.5 Examples using the EM Algorithm . . . 45

4 Robust Location Estimator **49**
4.1 Model . . . 49
4.2 Calculating the Estimator . . . 50
4.3 Influence Function of the Estimator . . . 52
4.4 Mixture Model . . . 52
4.5 Calculating the Robustified Estimator . . . 53
4.5.1 E-Step . . . 54
4.5.2 M-Step . . . 54
4.6 Influence Function for $\tilde{\theta}$ . . . 55
4.6.1 One-Step Influence Function for $\tilde{\theta}$ . . . 56
4.6.2 Jacobian Matrix . . . 57
4.6.3 True Influence Function for $\tilde{\theta}$ . . . 58
4.7 Numerical Results . . . 58
4.7.1 Comparison of $\tilde{\theta}$ with Standard Robust Estimates . . . 59
4.8 Efficiency . . . 61
4.8.1 Choosing λ . . . 62
4.8.2 Comparing Estimates Based on the Efficiency . . . 62

5 Robust Estimation for Linear Models **65**
5.1 Model . . . 65
5.2 Calculating the Estimator . . . 68
5.3 Influence Function for $\hat{\beta}$ . . . 69
5.4 Mixture Model . . . 70
5.5 Calculating the Robustified Estimator . . . 70
5.5.1 E-Step . . . 71
5.5.2 M-Step . . . 71
5.6 Influence Function for $\tilde{\beta}$ . . . 72

5.6.1 One-Step Influence Function for $\tilde{\beta}$ . . . 72
5.6.2 Jacobian Matrix . . . 73
5.6.3 True Influence Function for $\tilde{\beta}$ . . . 74
5.7 Comparison of Methods Based on the Influence Function of β 75
5.8 Numerical Results . . . 76
5.8.1 Comparison of $\tilde{\beta}$ with Standard Robust Estimates . . . 81
5.9 Efficiency . . . 83
5.10 Appendix: The Empirical Influence Function for One Step Estimator $\breve{\beta}$ . . . 84

6 Robust Estimation for Generalized Linear Models 86
6.1 Introduction to GLM . . . 87
6.2 Model . . . 87
6.3 Calculating the Estimator . . . 89
6.4 Influence function for $\hat{\beta}$ . . . 92
6.4.1 One-Step Influence function for $\hat{\beta}$ . . . 92
6.4.2 Jacobian Matrix . . . 92
6.4.3 True Influence Function for $\hat{\beta}$ . . . 94
6.5 Examples . . . 95
6.5.1 Poisson Model . . . 95
6.5.2 Binomial . . . 97
6.6 Mixture Model . . . 98
6.7 Calculating the Robustified Estimator . . . 98
6.7.1 E-Step . . . 99
6.7.2 M-Step . . . 99
6.8 Influence Function for $\tilde{\beta}$ . . . 102
6.8.1 One-Step Influence Function for $\tilde{\beta}$ . . . 102
6.8.2 Jacobian Matrix . . . 102
6.9 Applications . . . 106
6.9.1 Binomial Models . . . 106
6.9.2 Poisson Models . . . 113

7 Extension of Robust Estimation for Linear Models 120
7.1 Model . . . 120
7.2 Calculating the Estimator . . . 121
7.3 Mixture Model . . . 122
7.4 Calculating the Robustified Estimator . . . 122
7.4.1 E-Step . . . 123

7.4.2 M-Step . . . 123
7.5 Influence Function for $\tilde{\phi}$ . . . 124
7.5.1 One - Step Influence Function for $\tilde{\phi}$ . . . 125
7.5.2 Jacobian Matrix . . . 126
7.6 Numerical Results . . . 129
7.6.1 Comparison of $\tilde{\phi}$ with Standard Robust Estimates . . . 131
7.7 Location and Scale . . . 134

8 Robust Estimation for Non-Linear Models 139
8.1 The Model . . . 139
8.2 Calculating the Estimator . . . 140
8.3 The Mixture Model . . . 144
8.4 Calculating the Robustified Estimator . . . 144
8.5 Numerical Results . . . 146
8.6 Appendix - Treated Puromycin Data . . . 148

9 Another Choice of g 149
9.1 Robust Location Estimate . . . 149
9.2 Robust Regression Estimates . . . 150
9.3 Robust Estimates for Poisson Regression . . . 152
9.4 Robust Location and Scale Estimates . . . 154
9.5 Selection of C_0 . . . 155
9.6 Discussion . . . 158

10 Mixture Estimates for General Case 161
10.1 Model . . . 161
10.2 Calculating the Estimator . . . 161
10.3 Influence Function for $\hat{\theta}$ . . . 162
10.4 Mixture Model . . . 163
10.5 Calculating the Robustified Estimator $\tilde{\theta}$ . . . 164
10.6 One-Step Influence Function for $\tilde{\theta}$ . . . 165
10.7 True Influence Function of $\tilde{\theta}$ . . . 167

11 Summary and Concluding Remarks 169
11.1 Concluding Remarks . . . 170

Chapter 1

Two Approaches to the Problem of Bad Data

Statisticians have long recognized that the analysis of a real data set using a given statistical method should not be automatically carried out without some preliminary checking of the adherence of the data to the assumptions underlying the method. For this reason, a complementary set of statistical methods, known as diagnostic methods, has been developed for detecting violations of the assumptions and for the identification of particular data points whose values are particularly in conflict with the assumptions.

However, in many situations data are very plentiful and skilled human intervention is expensive, so another theme in statistics has been the development of a class of methods, which perform well when traditional assumptions are not met. This theme, which goes by the name of Robust Statistics, does not seek to identify troublesome data but rather seeks estimators whose values are determined by the bulk of the data and are insensitive to large changes in the values of a small proportion of the data. As a consequence, it may be expected that these "robust estimators" are largely unaffected by a small proportion of bad or spurious data.

1.1 Diagnostic Statistics

The theme of diagnostic statistics initially focussed on statistics supplementary to regression output. Work in this field was presented in the two books

[6] and [10] which summarized the material in previously published papers by these authors and others.

It is now common for regression output to present plots and tables based on residuals, and leverages, and distance measures such as Cook's distance, which summarize the effect an individual data point has on the parameter estimates. Pregibon [46] extended these diagnostics to logistic regression modelling and now many types of statistical models have complementary diagnostic statistics of this kind.

Diagnostic statistics are used in various ways. Often they may suggest an alteration to the model such as a transformation of the response variable which may bring the data into greater conformity with the assumed model. If a point is flagged as particularly influential it should, if possible, be verified as correct. If a point has a particularly large residual, it may be regarded as an *outlier* and considered for deletion from the data.

1.1.1 Outliers

Outliers are data values that are remote from the main body of the data. They have long been a problem in the application of statistics and their handling is addressed by several authors. Hawkins [22] discussed outliers from various angles. We introduce the topic by considering the following example:

Example 1.1

The data set introduced by Stigler [49]: sixty-six measurements of the speed of light, were made by Newcomb between July and September 1882. Simon Newcomb measured the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. These measurements were used to estimate the speed of light. The original data is given in Appendix B1 of [18]. It has been analyzed in the Bayesian frame work in pages 77 and 160. The data is plotted in Figure 1.1.

Figure 1.1 shows that two observations were quite atypical by the virtue of being far from the bulk of the data: they are *outliers*. The term outlier

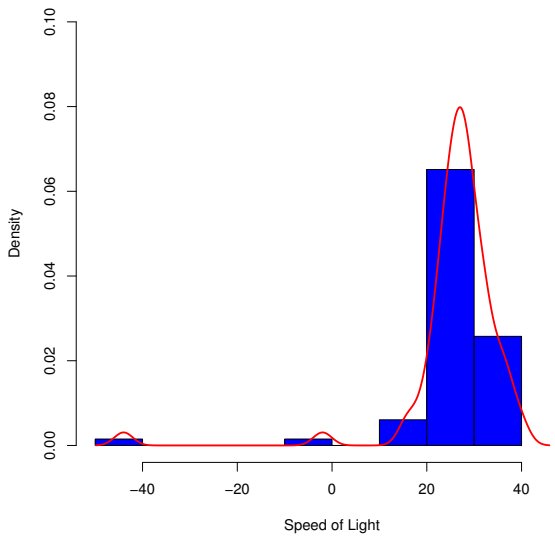

Figure 1.1: Density of speed of light.

is used collectively for discordant observations and contaminants. An observation that appears surprising or discrepant to the investigator is considered as a discordant observation [29] and an observation that is not a realization from the target distribution is considered as a contaminant.

If the contaminating distribution, generating 'contaminants', has tails which are heavier than the distribution which generates the 'good' observations, then there will be a tendency for the contaminants to be outliers. For example, the basic distribution might be standard normal and the contaminating distribution a normal distribution with mean zero and large variance $\sigma^2(>> 1)$. A mixture F of these two distributions is known as a *contaminated normal* and we may write symbolically

$$F = (1-p)\,\mathcal{N}(0,1) + p\,\mathcal{N}(0,\sigma^2)$$

where p is the probability that a given observation comes from the contaminating distribution. The pattern of these three distributions can be viewed in Figure 1.2, where $p = 0.05$ and $\sigma^2 = 9$ are chosen.

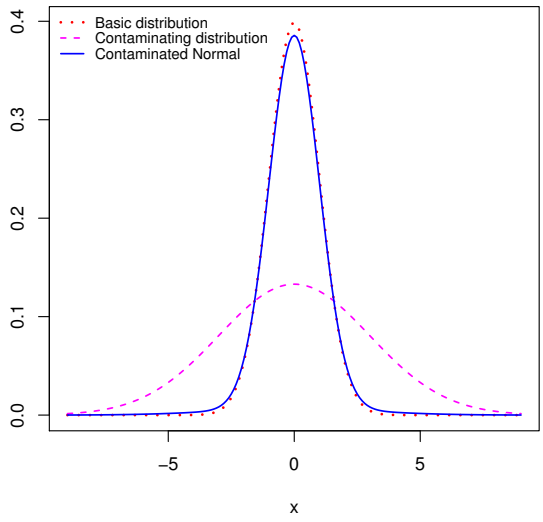

Figure 1.2: Density of contaminated normal.

Tukey [27] used this method to generate artificial contaminated data and later Huber [27] generated contaminated data for examining the efficiency of the estimates such as mean absolute deviation and standard deviation with divisor n.

Outliers for roughly normal data can be detected by a number of methods such as z-score, modified z-score, and modified box plot, [29], [3]. In the regression analysis, the outlier may be defined based on the the residual, the difference between actual response value and predicted response value. Note that it is different from an influential point, which is an outlier in the space of explanatory variable points. Leverages, which are discussed in section 1.4.3, may be used to identify the influential points in regression analysis. Cook distance is commonly used to identify the outliers. Anderson ([1], page 83) explains to identify outliers in the case of GLM. Note that outliers may be incorrectly identified when we assume an incorrect model.

1.2 Robust Statistics

Many classical statistical methods are vulnerable in the presence of outliers, parameter estimates particularly are very sensitive to outliers. It is possible to use other statistics, such as the median, which are insensitive to outliers. However, such statistics may have low efficiency (large variance) compared with classical estimates. The simplest situation occurs when we seek a measure of the centre or location of a set of data. For example, the sampling variance of the mean is smaller than the sampling variance of the median when the data is sampled from a normal distribution.

John W. Tukey was interested in looking for location estimators that shared the median's property of being insensitive to outliers but with smaller variance than the median when the data was sampled from a normal distribution. This project started to attract serious academic attention when several statisticians, including Frank Hampel, Peter Bickel and Peter Huber, were invited to join John Tukey at Princeton University for the academic year 1971-72 to study this and related questions. Many estimators were studied during this year, which became known as the "Princeton Robustness Year". Some of the results of this collaboration were initially published in [2], [27], and [21].

Huber [27] defines a statistic as being *robust* when its distribution is not sensitive to "small" departures from assumptions about how the data is obtained. For example, the data is a sample from a normal distribution. A related concept is *resistance* which signifies that a statistic is not greatly changed by small changes to the data. Both definitions can be made precise in different ways depending on how small is interpreted. In fact the two properties tend to go together and except in very technical contexts it is not necessary to make a sharp distinction between them.

The standard estimation procedures for robust statistics were developed in the 1970s. In simple language they handle good and bad (outlier) data together in such a way as to limit the effect of the bad data.

There are three main classes of robust procedures in practice. They are:

- *M-estimators*: based on the maximum likelihood;
- *L-estimators*: based on the linear combination of ordered observations;

and

- *R-estimators*: derived from the rank tests.

Although robust statistics is a precise approach for obtaining reasonable estimates for possibly contaminated data, in practical situations it has shortcomings, for example:

- There are so many choices and no clear idea of which method to use;
- Computational problems can arise with some methods; and
- Although asymptotic results for robust statistics are available, the general impression is that robust techniques can only be used for moderate sized data sets.

Robust estimation procedures are somewhat different in concept from model-based parametric estimation. Robust statistics are often algorithmically, defined as *statistical functionals*, whereas model-based parametric estimates are designed to estimate the parameters of an assumed distribution. Statistical functionals can be applied to both theoretical and empirical distributions. They can also be applied to distributions in the neighborhood of a given distribution.

1.2.1 Statistical Functionals

Let $y_1, y_2, \ldots, y_n$ be a sample from a population with parameter θ and distribution function $F \in \mathcal{F}$, where $\mathcal{F}$ is a collection of distribution functions including the empirical distribution function F_n to be defined immediately below and let $T_n = T_n(y_1, y_2, \ldots, y_n)$ be a statistic for a parameter θ. If T_n can be written as a functional T of the empirical distribution function F_n, then we can say that T is a *statistical functional*.

The empirical distribution function $F_n(y)$ based on the sample observations $y_1, y_2, \ldots, y_n$ can be defined as follows

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I_{[y_i, \infty)}(y) \tag{1.1}$$

where $I_{[y_i,\infty)}(y)$ is the indicator function for the set of real numbers greater than or equal to y_i. In robust statistics, a parameter θ is defined as a functional T of the distribution function F, which is often approximated by F_n. Hence an estimator, $\hat{\theta}$, of the parameter θ can be defined as a functional T of the empirical distribution function F_n. For example, the median can be written as $T_n = T(F_n) = F_n^{-1}(\frac{1}{2})$, where $F_n^{-1}(\frac{1}{2}) = \inf\{y \mid F_n(y) \geq \frac{1}{2}\}$.

Note that the empirical distribution function is unaffected if the sample is repeated α times, where $\alpha \in \mathcal{Z}^+$. Hence statistics which may be written as functionals of the empirical distribution function, are not affected if the sample is repeated α times.

The following numerical example describes this situation. If $Y = \{3, 6, 7\}$, $\alpha = 2$, then the new data set is $YY = \{3, 6, 7, 3, 6, 7\}$, and the empirical distribution functions for Y and YY are the same. Note that $mean(Y) = 5.33 = mean(YY)$, but $var(Y) = 4.33 \neq 3.47 = var(YY)$ due to the $(n-1)$ divisor. Being unable to study the usual sample variance directly within the function of distribution frame work is not a great limitation as we can study the divisor-n variance as an alternative. Note that some statistics cannot be written in the form of a statistical functional.

Statistical functionals $T(F)$are often defined expectations

$$T(F) = \begin{cases} \int g(y)dF(y) & \text{if } F \text{ is a continuous distribution} \\ \\ \sum_i g(y_i)p_i & \text{if } F \text{ is a discrete distribution} \end{cases}$$

where g is a real valued function and p_i is the point probability mass at y_i. For example, if $g(y) = y^2$, then $T(F)$ becomes the second moment and is denoted as $E_F(Y^2)$ and

$$T(F_n) = \int y^2 dF_n(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

Note that robust statistics is mainly focused on the consistency property, rather than unbiased property. The estimator $T(F_n)$ is said to be *consistent* at F if $T(F_n) \to T(F)$ in probability.

1.3 Maximum Likelihood Estimators

We are mainly focused on the M-estimator in our research. This will be discussed in section 1.4. We begin to discuss the maximum likelihood estimation (MLE) here because the M-estimator is considered as an extension of MLE, and also because our robustified estimators will be defined using MLE .

Let $f(y, \theta)$ be the joint probability density function of random vector $Y = (Y_1, \ldots, Y_n)$, where θ is a parameter. The likelihood $L(\theta, y)$ has the same form as the joint density function $f(y, \theta)$ with a different order of the arguments. If the Y_i's are identically and independently distributed, then the likelihood $L(\theta, y)$

$$L(\theta, y) = \Pi_{i=1}^{n} f(y_i, \theta)$$

The maximum likelihood estimate of θ is the value $\hat{\theta}$, that maximizes $L(\theta, y)$. Subject to regularity conditions, this implies that $\frac{\partial L(\theta,y)}{\partial \theta} = 0$ when $\theta = \hat{\theta}$. Thus,

$$L(\hat{\theta}, y) \geq L(\theta, y) \ \ \forall\, \theta$$

In addition, the variance of $\hat{\theta}$ can be shown under regularity conditions to be given by

$$\text{var}(\hat{\theta}) = \left[-\frac{\partial^2 L(\theta, y)}{\partial \theta^2} \right]^{-1}$$

For mathematical simplification, $l(\theta, y) = \log L(\theta, y)$ is often considered instead of $L(\theta, y)$. Please refer to [15] for more details.

1.3.1 The Exponential Family

Many of the standard distributions belong to the exponential family.

Definition: *Exponential Family*
The function $f(y, \theta)$ is said to be a member of the *exponential family* if it can be written as

$$f(y, \theta) = exp[a(y)b(\theta) + c(\theta) + d(y)]$$

where a and d are functions of y, and b and c are functions of the parameter θ. The density function which belongs to the exponential family is said to be a *canonical form* if $a(y) = y$, and $b(\theta)$ is called a natural parameter.

Dobson [15] discussed the properties of exponential family distributions, including mean and variance.

$$\begin{aligned} E[a(Y)] &= -\frac{c'(\theta)}{b'(\theta)} \\ V[a(Y)] &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \end{aligned}$$

For example, consider the Gamma distribution with a scale parameter θ, which is the parameter of interest, and a known shape parameter λ, its probability density function $f_Y(y, \theta)$ can be written in the exponential family form

$$\begin{aligned} f_Y(y,\theta) &= \frac{y^{\lambda-1}\,\theta^{\lambda}\,e^{-y\,\theta}}{\Gamma(\lambda)} \\ &= \exp[y \times (-\theta) + \lambda \log(\theta) + (\lambda - 1)\log(y) - \log(\Gamma(\lambda))] \\ &= \exp[a(y)b(\theta) + c(\theta) + d(y)] \end{aligned}$$

where

$$\begin{aligned} a(y) &= y, \text{ indicates canonical form,} \\ b(\theta) &= -\theta, \\ c(\theta) &= \lambda \log(\theta), \text{ and} \\ d(y) &= (\lambda - 1)\log(y) - \log(\Gamma(\lambda)) \end{aligned}$$

Hence, $E(Y) = -\frac{\lambda/\theta}{-1} = \frac{\lambda}{\theta}$ and $V(Y) = \frac{0-(-\lambda/\theta^2)\times(-1)}{(-1)^3} = \frac{\lambda}{\theta^2}$. For the maximum likelihood estimator of θ

$$\begin{aligned} l(\theta, y) &= \log L(\theta, y) = \log\left[\Pi_{i=1}^{n} f(y_i, \theta)\right] \\ &= \sum_{i=1}^{n} [y_i(-\theta) + \lambda \log\theta + (\lambda - 1)\log y_i - \log\Gamma(\lambda)] \\ \frac{dl(\theta, y)}{d\theta} &= \sum_{i=1}^{n}\left[-y_i + \lambda\frac{1}{\theta}\right] = 0 \\ \hat{\theta} &= \frac{n\lambda}{\sum_{i=1}^{n} y_i} \end{aligned}$$

1.4 M-estimators

An M-estimator, T_n, is defined in (1.2)

$$T_n = \arg\min_{\theta} \sum_{i=1}^{n} \rho(y_i, \theta) \tag{1.2}$$

where ρ, often referred to as objective function (loss function), is a function chosen to provide the estimator with good properties when the data come from a distribution with density proportional to $\exp(-\rho(y, \theta))$ or close to it.

For the minimization of (1.2),

$$\sum_{i=1}^{n} \psi(y_i, \theta) = 0 \tag{1.3}$$

where ψ is a known real valued function and defined in (1.4)

$$\psi(y_i, \theta) = \frac{\partial \rho(y_i, \theta)}{\partial \theta} \tag{1.4}$$

This method is called M-estimation, because it is a generalized form of MLE. That means $-\rho(y, \theta)$ need not necessarily be the log-density of y. If $\rho(y_i, \theta) = -\log f(y_i, \theta)$, it is known as a maximum likelihood estimator (MLE). For MLE, $\psi(y_i, \theta)$ is often called the score function.

From (1.3), we can write $E_{F_n}[\psi(y, T_n)] = 0$. Hence the statistical functional corresponding to T_n is defined as a solution of (1.5).

$$E_F[\psi(y, T(F))] = \int \psi(y, T(F)) dF = 0 \tag{1.5}$$

For the regression situation, an M-estimator can be defined as being a value of θ satisfying

$$\sum_{i=1}^{n} \psi(y_i, x_i, \theta) = 0 \tag{1.6}$$

where $\psi(y_i, x_i, \theta) = \frac{\partial \rho(y_i, x_i, \theta)}{\partial \theta} = \frac{\partial \rho\left(\frac{y_i - x_i, \theta}{\sigma}\right)}{\partial \theta}$.

Iterative computation is required to solve (1.3), because (1.3) is a set of non-linear equations. Holland and Welsch [24] give a good survey of the use of iterative re-weighted least squares (IRLS) for robust M-estimation of the parameters. These estimates down-weight observations with large residuals in order to achieve resistance to small proportions of discordant observations.

We need to choose starting values to run IRLS. It is important to choose good starting values, because of the non-convex ρ functions, [24]. Holland and Welsch suggest various starting values for various cases in their paper [24]. Note that the application of IRLS in robust estimation of linear regression is different from its application in the standard generalized linear model (GLM), where the response variable is often adjusted. For example, the M-estimator for location can be expressed as a weighted mean. Let $w(y, \theta)$ be a weight function.

$$w\left(y,\theta\right)=\begin{cases}\frac{\psi(y,\theta)}{y} & \text{if } y\neq 0\\ \\ \psi'(0) & \text{if } y=0\end{cases} \tag{1.7}$$

Since ρ is our choice, a number of ρ functions are defined in the robust literature. For example:

L_p Estimator
: If we choose $\rho(y,\theta)=|y-\theta|^p$, the resultant M-estimator is known as L_p estimator. If $p=1$, the location L_1 estimate is the median, and if $p=2$ it is the mean.

Huber ρ function
: Huber [25] defined a ρ function based on the density with a Gaussian in the center and double exponential in the tails. The Huber $\rho\left(r\right)$ function is defined in (1.8).

$$\rho\left(r\right)=\begin{cases}\frac{r^2}{2} & \text{if } |r|\leq c\\ \\ c\left|r\right|-\frac{c^2}{2} & \text{if } |r|>c\end{cases} \tag{1.8}$$

Bi-square ρ function
: Tukey introduced another common ρ function for M-estimators, defined

in (1.9). It is known as the *Tukey Bi-square* (Bi-weight) estimator.

$$\rho(r) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \frac{r^2}{c^2}\right)^3\right] & \text{if } |r| \leq c \\ \frac{c^2}{6} & \text{if } |r| > c \end{cases} \tag{1.9}$$

The values 'c' used in (1.8) and (1.9) are called *tuning constants*, which, roughly speaking, govern whether a point is treated as typical or as an outlier. With smaller value of the tuning constant more points will be treated as outliers. With a large value of the tuning constant, only residuals that are large relative to variance will be treated as outlying and downweighted. The tuning constant is chosen as a trade off between robustness of the estimator and efficiency of the estimator at normal errors. Page 27 of [35] explained this situation with various contamination proportions. In practice, we often choose $c = 1.345$ for (1.8), and $c = 4.685$ for (1.9) in order to have 95% efficiency at the normal. Corresponding ψ functions for (1.8) and (1.9) are given below.

$$\psi(r) = \begin{cases} r & \text{if } |r| \leq c \\ c\,\text{sign}(r) & \text{if } |r| > c \end{cases} \tag{1.10}$$

$$\psi(r) = \begin{cases} r\left[1 - \left(\frac{r}{c}\right)^2\right]^2 & \text{if } |r| \leq c \\ 0 & \text{if } |r| > c \end{cases} \tag{1.11}$$

Since the ψ function in (1.11), after initially increasing, decreases to zero as $r \to \pm\infty$, the estimator is known as *re-descending*. In general, re-descending M-estimators are more efficient than Huber estimators but may be more sensitive to the choice of starting values when computed by IRLS.

From (1.10) and (1.7) we can say that Huber assigned unit weight for the center and small weights (closer to zero, but not exactly to zero) to observations that are outlying from the centre. Different weights are assigned for other ρ-functions [24].

Definition: *Translation invariance*
An estimator, T, is called *translation invariance* if

$$T(y_1 + a, \ldots, y_n + a) = T(y_1, \ldots, y_n) + a$$

Definition: *Scale equivariant*
An estimator, T, is called *scale equivariant* if

$$T(ay_1, \ldots, ay_n) = aT(y_1, \ldots, y_n)$$

where a is a constant.

A drawback of the location M-estimator is that it is translation invariance but it is not scale equivariant [31]. This problem may be solved by one of the following approaches:

1. Define $r = \frac{y-\theta}{\sigma}$ if σ is known.
2. Compute the robust estimate, $\hat{\sigma}$, for the scale parameter σ before estimating the location parameters and define $r = \frac{y-\theta}{\hat{\sigma}}$.
3. Compute robust estimate for location and scale together by defining $r = \frac{y-\theta}{\sigma}$ (Refer to section 1.4.2).

The median absolute deviation (MAD) is considered as a better robust estimate for scale parameter, σ, than the sample standard deviation, because it is insensitive to outliers.

$$MAD = \text{median}|y_i - \hat{\theta}|$$

where $\hat{\theta}$ is the median. It is advisable to determine the scale parameter σ first if we are interested in computing the location parameter. Another choice for scale is *normalized* MAD (MADN) ([35], page 33)

$$MADN = \frac{MAD}{0.675}$$

Holland and Welsch [24] list some familiar ρ and ψ functions for M-estimators. Next, we will use M-estimator procedures to compute the robust estimators for linear models.

1.4.1 M-estimators of Location

We can assume that the outcome y_i is generated around the true value θ, which is the parameter to be estimated. Consider the model (1.12)

$$y_i = \theta + \epsilon_i \ , i = 1 \ldots n \tag{1.12}$$

where the errors $\epsilon_1, \ldots, \epsilon_n$ are independent random variables, which have the same distribution function. If the data come from exactly normal distribution, the optimal estimate for θ is mean. Our goal here is to seek for estimates that are reasonably good in some sense when the distribution is approximately normal.

The $\Delta = \sum_i \rho(\frac{y_i - \theta}{\sigma})$ needs to be minimized to find the estimate of θ. Note that we are not seeking to find the estimator for the scale parameter, σ. At the minimum

$$\begin{aligned} \frac{d\Delta}{d\theta} &= 0 \\ \sum_i \frac{d}{d\theta}\left[\rho\left(\frac{y_i - \theta}{\sigma}\right)\right] &= 0 \end{aligned}$$

The estimator $\hat{\theta}$ of the parameter θ satisfies the following equation:

$$\sum_i \psi\left(\frac{y_i - \hat{\theta}}{\sigma}\right) = 0 \tag{1.13}$$

Based on (1.7), the estimating equation (1.13) can be written as follows:

$$\begin{aligned} \sum_i (y_i - \hat{\theta})\left[w\left(\frac{y_i - \hat{\theta}}{\sigma}\right)\right] &= 0 \\ \hat{\theta}\sum_i w_i - \sum_i w_i y_i &= 0 \end{aligned}$$

This leads to

$$\hat{\theta} = \frac{\sum_i w_i x_i}{\sum_i w_i} \tag{1.14}$$

where $w_i = w\left(\frac{y_i - \hat{\theta}}{\sigma}\right)$. The common and well known choices for weight function w are the Huber and the Tukey weight functions. The expression (1.14) is not an explicit form, because w in the right hand side of the expression (1.14) depends on θ. Therefore, solving this problem requires an iteration process. That means, a given initial value for θ will give new weights, which gives new θ and so on. Let $\hat{\theta}^{(m)}$ be the m^{th} iteration of the process with

initial value $\theta^{(0)}$.

$$\hat{\theta}^{(m)} = \frac{\sum_i w_i^{(m-1)} y_i}{\sum_i w_i^{(m-1)}}$$

If $\hat{\theta}^{(m)} \approx \hat{\theta}^{(m+1)}$, then $\hat{\theta} = \hat{\theta}^{(m+1)}$.

1.4.2 M-estimators of Location and Scale

Until now, we have considered only M-estimators for the location parameter with a previously estimated scale parameter. M-estimators for the parameters location θ and scale σ can be computed simultaneously. The estimates $\hat{\theta}$ and $\hat{\sigma}$ for the parameters θ and σ are defined in (1.15).

$$(\hat{\theta}, \hat{\sigma})^t = \arg\min_{\theta,\, \sigma} \left\{ \sum_{i=1}^{n} \sigma\rho\left(\frac{y_i - \theta}{\sigma}\right) + a\sigma \right\} \tag{1.15}$$

where a is a positive appropriate constant and t indicates the transpose. A necessary condition for a minimum is that $\hat{\theta}$ and $\hat{\sigma}$ satisfy

$$\sum_i \psi\left(\frac{y_i - \theta}{\sigma}\right) = 0 \tag{1.16}$$

$$\sum_i \chi\left(\frac{y_i - \theta}{\sigma}\right) = a \tag{1.17}$$

where $\chi(r) = r\psi(r) - \rho(r)$. To obtain robust statistics, the ψ function may be defined with standard robust functions such as Huber, Tukey bi-weight, etc.

Huber [25] defined another possible function $\chi(r) = \psi(r)^2$ and

$$a = 2\Phi(c) - 1 - 2c\varphi(c) + 2c^2(1 - \Phi(c))$$

where φ and Φ are the standard normal density and distribution function respectively, c is the Huber turning constant, and ψ is a Huber function defined in (1.10). It is known as *Huber's proposal 2*. MLE for θ and σ are obtained if $\psi(r) = r$ and $\chi(r) = n - r\psi(r)$.

1.4.3 M-estimators of Regression

The robust estimation of the parameters for the multiple linear regression model is briefly discussed in this section. The least squares method is not appropriate to estimate the regression parameters for contaminated data, because unusual observations have a heavy impact on the least square estimates of the regression coefficients. In the regression analysis, there are two types of unusual observations. One is outliers, where observations are distant from bulk of the data, and the other one is influential points (points of leverage) which are outlying in the space of predictor variables.

Consider the classical multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \ i = 1, \ldots, n$$

or, in matrix form.

$$y = X\,\beta + \epsilon \tag{1.18}$$

where

$$\begin{aligned} y &= [y_i]_{n\times 1} \text{ is a vector of response values} \\ X &= [x_{ij}]_{n\times(k+1)} \text{ is a data matrix with full rank and } x_{i0} = 1 \ \forall i \\ \beta &= [\beta_j]_{(k+1)\times 1} \text{ is a vector of parameters to be estimated} \\ \epsilon &= [\epsilon_i]_{n\times 1} \text{ is a random error vector with the following properties).} \end{aligned}$$

Note that $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I$ where I is an $n \times n$ identity matrix. The estimates for the parameter of regression coefficients can be obtained using the least squares (LS) method

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t y \tag{1.19}$$

$$\text{var}(\hat{\beta}) = (X^t X)^{-1}\sigma^2 \tag{1.20}$$

where $\hat{\sigma}^2 = \hat{\epsilon}^t\hat{\epsilon}/(n-(k+1))$ and $\hat{\epsilon} = y - \hat{y}$.

$H = X\,(X^t X)^{-1}\,X^t$ is called the *hat matrix* (or projection matrix). A diagonal element, h_i of zero, indicates no influence on the fit by the i^{th} row vector $\mathbf{x_i}$ of the design matrix, X. In contrast, high leverage points can be determined by looking at the diagonal elements of H. A rule of thumb is

that if $h_i > \frac{2(k+1)}{n}$ then attention to row vector $\mathbf{x_i}$ of the design matrix, X, needs to be given, where h_i is a i^{th} diagonal element of H, $(k+1)$ is a number of parameters in the model and n is the sample size. Rousseeuw and Leroy ([48], page 217) discuss the usefulness of the hat matrix in detail.

Our main intention is to find the robust estimators for the regression coefficients. Huber [26] introduced the M-estimator method to calculate robust estimates for the regression parameters. It is a direct modification of M-estimation for location.

We are able to compute the estimator, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ by minimizing Δ with respect to β:

$$\begin{aligned}
\Delta &= \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{\sigma}\right) \\
&= \sum_{i=1}^{n} \rho\left(\frac{y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}}{\sigma}\right) \\
&= \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i\ \beta}{\sigma}\right)
\end{aligned}$$

where $\mathbf{x}_i = (x_{i0} = 1, x_{i1}, x_{i2}, ...x_{ik})$ is a i^{th} row vector of the matrix X. For minimization of Δ, $\frac{\partial \Delta}{\partial \beta} = 0$, which gives

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i\, \beta}{\sigma}\right) x_{ij} = 0; \;\; j = 0, 1, \ldots k \tag{1.21}$$

The equations (1.21) can be written by replacing the $\psi(r)$ function by $rw(r)$.

$$\begin{aligned}
\sum_{i=1}^{n} w_i \times [y_i - \mathbf{x}_i\beta] \times x_{ij} &= 0 \;\; \forall j = 0, 1, \ldots k \\
\Longrightarrow (\mathbf{x}^j)^t W y - (\mathbf{x}^j)^t W X \beta &= 0 \;\; \forall j = 0, 1, \ldots k
\end{aligned}$$

where $\mathbf{x}^j$ is a j^{th} column of design matrix X and W is a $n \times n$ diagonal matrix whose elements are w_i

$$w_i = w\left(\frac{y_i - \mathbf{x}_i\hat{\beta}}{\sigma}\right)$$

Hence, we can write in matrix form

$$X^t W y - X^t W X \beta = \mathbf{0} \tag{1.22}$$

The iteration process is required, because w_i depends on the parameter β. Finally, $\hat{\beta}$ may be defined as the limit of an IRLS process in which y is the dependent variable and the weight function is $w(r) = \frac{\psi(r)}{r}$. It is important to choose starting a value for β; say $\beta^{(0)} = (\beta_0^{(0)}, \beta_1^{(0)}, ...\beta_k^{(0)})$. Green [19] gives an example that shows that good starting values for parameters may have an effect on the speed of convergence, and that the choice of parameterization can strongly influence the sensitivity of the IRLS to the choice of starting values.

The following steps will lead a solution for (1.22) when the scale parameter σ is not of interest.

Step 1: Compute MAD / MADN for scale parameter, σ.

Step 2: Give initial values for β.

Step 3: Compute $r_i = \frac{y_i - \mathbf{x}_i \, \beta}{\sigma}$.

Step 4: Compute weights for each observation using the appropriate objective function.

Step 5: Use (1.22) to compute β with weights computed in Step 4.

Step 6: Go to Step 3 and repeat steps 4 and 5 until convergence is achieved for the estimated coefficients.

Huber [27] discusses the asymptotic covariance variance matrix, $\mathcal{V}$, given below, for the robust $\hat{\beta}$

$$\mathcal{V}(\hat{\beta}) = \frac{E(\psi^2)}{\left[E(\psi')\right]^2} \left(X X^t\right)^{-1}$$

Therefore, the estimated covariance matrix, $\mathcal{V}_{est}$, for the robust $\hat{\beta}$ is given by:

$$\mathcal{V}_{est}(\hat{\beta}) = \frac{n^{-1} \sum_i \left[\psi(r_i)\right]^2}{\left[n^{-1} \sum_i \psi'(r_i)\right]^2} \left(X X^t\right)^{-1}$$

Up to now we have assumed in the regression analysis that the response variable y may have outliers and the design matrix X contains no rows with high leverage. The design matrix X may be assumed as random, particularly for observational study, even though it is considered to be fixed in the theory. If X contains at least one high leverage point, the estimates are somewhat weak because estimates are highly influenced by those points in X.

An approach to overcome this issue is to down-weight the influential $\mathbf{x}_i$'s to prevent them from dominating the estimating equations. Based on this concept, [36] and [37] proposed a generalized M-estimator, known as the GM-estimator. This topic is extensively discussed in ([35], section 5.11), and ([31], section 4.4)

1.4.4 Generalized Linear Model

In this section, we introduce a model which permits the response variable to have any distribution, with a mean value equal to the function of a linear combination of predictor variables while data is contaminated. It is discussed briefly here and a detailed discussion appears in Chapter 6.

We mainly review the existing robust estimation method for regression coefficients of generalized linear model (GLM). The development of robust inference for GLM is very limited [8]. At present, only logit and Poisson models are commonly investigated in detail in the robust literature [1].

In the early stage, robust methods for GLM are developed in the same way as M-estimation in the linear model. Pregibon [46] found a way to make logistic regression more robust. Later, Bianco and Yohai [7] found that these estimators are not Fisher consistent ($E_\beta[\psi(y, x, \beta)] = 0$). They proposed amending Pregibon's method [46] by adding a correction term. Still, these estimates are not B-robust, defined in section 2.5, if the design matrix X is random. Croux and Haesbroeck [12] modified the version of Bianco and Yohai [7] to overcome this problem by down-weighting high leverage observations.

Apart from these methods, [32] introduced another method under the framework of M-estimation. It is a more complicated method and it is known

as *conditionally unbiased bounded influence* (CUBIF) estimates. Recently, Cantoni and Ronchetti [8] developed a method based on the quasi-likelihood generalized estimating equations of Preisser and Qaqish [47]. Estimates obtained by this method have bounded influence functions, defined in Chapter 2.

1.5 Goal of the Thesis

The M-estimator is an alternative to the classical estimator, but it is not considered globally for the whole class of generalized linear models in the robust literature [8]. Moreover, due to the number of ρ functions defined by various authors, it is hard to decide which method to use.

We propose a unified method to obtain robustified estimates, which are alternative to the maximum likelihood estimates (MLEs), for the whole class of generalized linear models. We use mixture models (see Chapter 3) to obtain the robustified estimates. We will demonstrate that our proposed estimators have good robustness properties.

For the thesis, we investigate a range of situations from simple to complicated models. Our aim is to show that our method provides a way to make the construction of statistical models for applications more robust. Location parameter estimation is considered in Chapter 4 and followed by linear models in Chapter 5. The more important problem of the generalized linear model is analyzed in Chapter 6. Chapter 7 describes how to estimate the parameters of location and scale together. Chapter 8 focus on how to compute robust statistics for the non-linear model parameters. Chapter 9 explains the computation of estimates using improper method. Finally, we discuss a very general case in Chapter 10. In the last Chapter, we summarize the thesis.

1.6 R Software

The object oriented programme **R** is used for the statistical analysis throughout this study. Therefore we would like to give a brief introduction about **R** here.

The **R** was introduced to the globe by University of Auckland, New Zealand in 1996, officially released in 1999. This is similar to the statistical programme Splus. Due to the free access and the main feature of flexibility, usage of **R** has increased exponentially. It can be freely downloaded from the web site http://www.r-project.org.

R comes with the base libraries and recommended packages. A number of authors contribute to **R** by supplying packages and now there are more than 300 packages available. The packages can be easily installed in **R** as required. For example, many packages for robust statistics are not installed in standard **R**.

For robust estimation, a number of packages are available in **R**. Some of these packages are given below.

1. MASS:

huber(), hubers() - M-estimator for location.

rlm() - Robust fitting of linear models using M-estimation method.

rnls() - Robust non-linear regression.

Other functions eg: cov.rob(), lqs(), etc.

2. robustbase:

lmrob() - Computes fast MM-estimators for linear regression models.

glmrob() - Robust fitting for generalized linear models especially for binomial and Poisson.

anova() - Model selection for both 'lmrob' and 'glmrob'.

Other functions eg: anova.lmrob(), Qn(), Sn() etc.

3. robust: This is exactly similar to the robust package in Splus.

lmRob() - Robust fitting for linear regression models.

glmRob() - Robust fitting for generalized linear models especially for binomial and Poisson. In addition, the CUBIF estimator can be obtained here.

Most of these functions are used in the appropriate places to compare our numerical results. We have written **R** functions based on our method to obtain estimates for parameters.

Chapter 2

Property of Robustness

In this chapter, we describe how to assess the robustness of an estimator quantitatively. Robustness is commonly measured by *influence function*, *breakdown point* and *gross-error sensitivity*. We use only influence function to examine the robustness of the estimator for this research.

2.1 Influence Function

Hampel is the originator of influence function [25]. It is sometimes called the influence curve, particularly when statistical functional T is one dimensional and F is defined over the real numbers.

Definition: *Influence Function*
Let T be a statistical functional and y be a point in a sample space. Then the influence function of T is defined by

$$IF_{T,F}(y) = \lim_{\epsilon \to 0} \frac{T([1-\epsilon]F + \epsilon\Delta_y) - T(F)}{\epsilon} \tag{2.1}$$

where Δ_y is the cumulative distribution function representing the unit probability mass at the point y. Often F will be taken as an empirical distribution function F_n, then the influence function is known as empirical influence function (EIF). The form of the right hand side of the (2.1) may be interpreted as the first derivative of T for an underlying distribution F. In fact, the influence function is known mathematically as *Gâteau derivative* of the functional T for the distribution F in the direction of the distribution Δ_y.

Suppose $\epsilon = 0.1$, we could imagine a sample size $10n$ made up of 9 copies of $y_1, \ldots, y_n$ and n copies of y, then this sample would have $(1-\epsilon)F + \epsilon\Delta_y$ as its empirical distribution function. Alternatively, we could think of a weighted sample in two parts, the first part having F_n and having equally weighted observations whose weights are $1-\epsilon$ and the second part being the single number $\{y\}$ with weight ϵ. For example, if $Y = \{3, 6, 7\}, y = \{100\}$ and $\epsilon = 0.1$, then the sample size is 30 $(= 3 \times 10)$, made up by $YY = \{3, 3, \ldots, 3, 3, 6, 6, \ldots, 6, 6, 7, 7, \ldots, 7, 7, 100, 100, 100\}$, which follows Hence $\bar{YY} = 14.8 = 0.9 \times \bar{Y} + 0.1 \times 100$.

Since $T((1-\epsilon)F + \epsilon\Delta_y)$ is an estimate for the distribution $(1-\epsilon)F + \epsilon\Delta_y$, a member of the contamination neighborhood of F, and $T(F)$ is an estimate for exact distribution F, the difference between these two estimates is caused by contamination in the observations, approximated by $\epsilon\, IF_{T,F}(y)$. In fact, the influence function measures the relative effect on $T(F)$ of a very small (infinitesimal) amount of contamination at y. If $IF_{T,F}(y) = 0$ at the statistic T, then adding new data y to the data set will not change the statistic T.

One finite sample version of (2.1) is called the *sensitivity curve*, SC_n, where we replace F by F_n and ϵ by $\frac{1}{n+1}$. It is also known as Tukey's sensitivity curve.

$$\begin{aligned} SC_n(y) &= (n+1)[T(\frac{n}{n+1}F_n + \frac{1}{n+1}\Delta_y) - T(F_n)] \\ &= (n+1)[T_{n+1}(y_1, \ldots, y_n, y) - T_n(y_1, \ldots, y_n)] \end{aligned}$$

For example, consider the harmonic mean for a set $y_1, \ldots, y_n$ of non-zero numbers. Here

$$T(F_n) = T\left(\frac{1}{n}\sum_{i=1}^{n} I_{[y_i,\infty)}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}\right)^{-1}$$

hence

$$T([1-\epsilon]F_n + \epsilon\Delta_y) = \left(\frac{1}{n}(1-\epsilon)\sum_{i=1}^{n}\frac{1}{y_i} + \epsilon\frac{1}{y}\right)^{-1}$$

$$
\begin{aligned}
T([1-\epsilon]F_n + \epsilon\Delta_y) - T(F_n) &= \frac{1}{\frac{1}{n}(1-\epsilon)\sum_{i=1}^{n}\frac{1}{y_i} + \epsilon\frac{1}{y}} - \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}} \\
&= \frac{\epsilon[\frac{1}{n}\sum_i \frac{1}{y_i} - \frac{1}{y}]}{[\frac{1}{n}(1-\epsilon)\sum_{i=1}^{n}\frac{1}{y_i} + \epsilon\frac{1}{y}][\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}]} \\
\lim_{\epsilon\to 0}\frac{T([1-\epsilon]F_n + \epsilon\Delta_y) - T(F_n)}{\epsilon} &= \lim_{\epsilon\to 0}\frac{\frac{1}{n}\sum_i \frac{1}{y_i} - \frac{1}{y}}{[\frac{1}{n}(1-\epsilon)\sum_{i=1}^{n}\frac{1}{y_i} + \epsilon\frac{1}{y}][\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}]} \\
IF_{T,F_n}(y) &= \frac{\frac{1}{n}\sum_i \frac{1}{y_i} - \frac{1}{y}}{\left(\frac{1}{n}\sum_i \frac{1}{y_i}\right)^2} \\
&= \frac{T^{-1} - y^{-1}}{T^{-2}} \\
&= T - \frac{T^2}{y}
\end{aligned}
$$

In fact, we often have no way of knowing whether new data is bad or not; so we measure, for each point in the sample space, how much the estimate is affected by the introduction of more observations at that point. Figure 2.1 shows the influence function for the harmonic mean of a sample from -3 to 3 with harmonic mean equal to 4. Notice that the harmonic mean is very sensitive to numbers close to zero and unbounded, but the effect of large numbers on the harmonic is small and bounded.

When a statistic is computed via an iterative algorithm, as is often the case in this thesis, the direct method of computing the influence function we used for the harmonic mean is not available.

2.2 Iteratively Defined Statistics

Estimates are frequently computed iteratively when MLE or robust methods are used. In this thesis, we are concerned with statistics that are iteratively defined in the following sense:

1. An initial value $\theta^{(0)}$ is given.;
2. The new estimate, $\theta^{(1)}$ is obtained by a known function $h(\theta^{(0)}, F_n)$. In

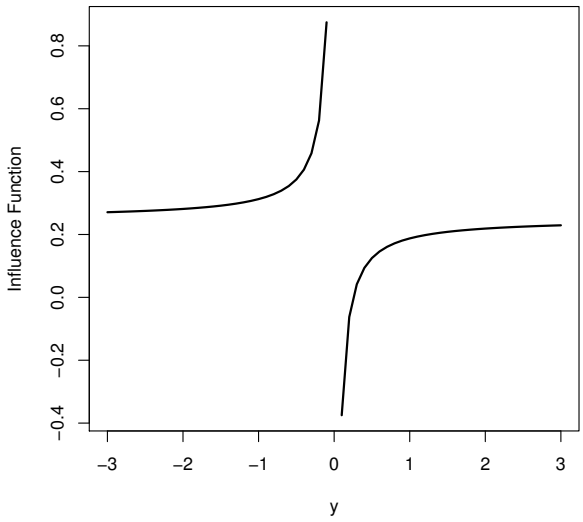

Figure 2.1: Influence function for harmonic-mean

general, we can write

$$\theta^{(k+1)} = h(\theta^{(k)}, F_n) \;\; k = 0, 1, 2, \ldots$$

where k indicates the k^{th} step of the iterative process; and

3. The estimate $\hat{\theta}$ is defined as the limit of the sequence $\left\{\theta^{(k)}\right\}_{k=1}^{\infty}$ if it exists.

Notice that, $\hat{\theta} = h(\hat{\theta}, F_n)$. The function h we call the updating function when the first argument θ is free and the second argument F_n is fixed. In contrast, when the first argument θ is fixed and the second argument F_n is free, it is a statistical functional.

Jorgensen [30] mentioned that it is often easy to compute the empirical influence function for closed formulae like the second part of the algorithm above, but it is computationally expensive to compute the influence function for estimates generated by the iterative process like the third part of the algorithm above. For example, suppose an estimator $\hat{\theta}$ is iteratively obtained after ten steps, then computing the influence function for $\hat{\theta}$ at seven arbitrary

points in the sample space requires seventy steps altogether. In addition, it is desirable to have an analytical result for the influence function for the iteratively defined statistics. An alternative method is required to compute the influence function for this situation. It will be discussed in section 2.3.

2.3 Computing the Influence Function of Iteratively Defined Statistics

Jorgensen [30] introduced the method to compute the influence function for an estimator. It is computed based on the one-step influence function and the derivative of the iteration function of the algorithm evaluated at the fixed point.

2.3.1 Jorgensen's Method

The true influence function for iteratively defined statistics T can be computed by (2.2), using the concept of the one-step influence function

$$IF_{T,F_n}(y) = (I - J)^{-1} IF^1_{T,F_n}(y) \tag{2.2}$$

where

1. I is an identity matrix;
2. J is the derivative of the iteration matrix evaluated at the estimate $\hat{\theta}$, defined as
$$J = \left[\frac{\partial}{\partial\theta} h(\theta, F_n)\right]_{\theta=\hat{\theta}}$$
where h is the updating function; and
3. The statistical functional G may be defined by the updating function h from the m^{th} step to the $(m+1)^{\text{th}}$ step,
$$G(F_n) = h(\hat{\theta}, F_n)$$
where $\hat{\theta}$ is fixed. $IF_{G,F_n}(y)$ is an influence function for G. This measures the change to the parameter estimate caused by a single step of the algorithm with infinitesimally contaminated data. We define $IF_{G,F_n}(y)$ as a one-step influence function of T and we use the notation
$$IF^1_{T,F_n}(y) = IF_{G,F_n}(y)$$

2.3.2 Jacobian Matrix

Definition: *Jacobian Matrix*
Let $f : \mathcal{R}^m \to \mathcal{R}^n$ be a vector function. That means, $y_i = f_i(x_1, \ldots, x_m)$, $i = 1, \ldots, n$. The Jacobian matrix J is the matrix of all first order partial derivatives of $f = (f_1, \ldots, f_n)$ with respect to $x = (x_1, \ldots, x_m)$.

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial f_n}{\partial x_1} & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial f_n}{\partial x_m} \end{pmatrix}$$

In practice often $m = n$, although it is useful not to assume this as it makes the confusion of J with its transpose less likely. We are interested having an analytical expression for the derivative of the iteration function, h, which is required to compute the true influence function for iteratively defined statistics (section 2.3.1). In fact, h is mapping from $\mathcal{R}^p$ to $\mathcal{R}^p$ by $\theta^{(m+1)} = h(\theta^{(m)})$, where p is a number of parameters and m denotes the iteration step. Therefore the derivative of the iteration function, h, is a Jacobian matrix. We use the notation

$$J(\theta) = \frac{\partial h}{\partial \theta}$$

Jorgensen [30] describes how to obtain this matrix in a fairly general form for the case of iteratively reweighted least squares.

2.3.3 Example of the Jorgensen Method

In this section, the influence function will be computed analytically and numerically based on Jorgensen method.

The One Step Influence Function

Consider a statistic defined by the iteration with updating function $h(\theta, F_n)$, where $w()$ satisfies (1.7). For example, the location M-estimates of robust statistics may be specified by (2.3)

$$h(\theta, F_n) = \frac{\sum_{i=1}^{n} w(y_i - \theta) y_i}{\sum_{i=1}^{n} w(y_i - \theta)} \tag{2.3}$$

Note that $\hat{\theta} = h(\hat{\theta}, F_n)$. Now consider the perturbed data set $y_1, \ldots, y_n, y$ with weights $\frac{1-\epsilon}{n}, \ldots, \frac{1-\epsilon}{n}, \epsilon$. The empirical influence function for the statistic $G(F_n) = h(\hat{\theta}, F_n)$, is therefore given below.

$$IF_{G,F_n}(y) = \lim_{\epsilon \to 0} \frac{h(\hat{\theta}, (1-\epsilon)F_n + \epsilon \Delta_y) - h(\hat{\theta}, F_n)}{\epsilon}$$

$$IF_{G,F_n}(y) = \lim_{\epsilon \to 0} \frac{h(\hat{\theta}, (1-\epsilon)F_n + \epsilon \Delta_y) - \hat{\theta}}{\epsilon} \tag{2.4}$$

From (2.3), we can write

$$h(\hat{\theta}, (1-\epsilon)F_n + \epsilon \Delta_y) = \frac{\frac{(1-\epsilon)}{n} \sum_i w_i y_i + \epsilon w\, y}{\frac{(1-\epsilon)}{n} \sum_i w_i + \epsilon w}$$

where $w_i = w(y_i - \hat{\theta})$ and $w = w(y - \hat{\theta})$. Consider

$$\begin{aligned} h(\hat{\theta}, (1-\epsilon)F_n + \epsilon \Delta_y) - h(\hat{\theta}, F_n) &= \frac{\epsilon w \left(y \sum_i w_i - \sum_i w_i y_i\right)}{\sum_i w_i \left(\frac{(1-\epsilon)}{n} \sum_i w_i + \epsilon w\right)} \\ \lim_{\epsilon \to 0} \frac{h(\hat{\theta}, (1-\epsilon)F_n + \epsilon \Delta_y) - h(\hat{\theta}, F_n)}{\epsilon} &= \frac{w \left(y \sum_i w_i - \hat{\theta} \sum_i w_i\right)}{(1/n) \left(\sum_i w_i\right)^2} \end{aligned}$$

$$IF_{G,F_n}(y) = \frac{w(y - \hat{\theta})\,(y - \hat{\theta})}{(1/n) \sum_i w(y_i - \hat{\theta})} \tag{2.5}$$

Derivative of Updating Function

From (2.3), we can write

$$h(\theta, F) \sum_{i=1}^{n} w(y_i - \theta) = \sum_{i=1}^{n} w(y_i - \theta) y_i$$

Differentiate both sides with respect to θ

$$\frac{d}{d\theta} h(\theta, F) \sum_{i=1}^{n} w_i - h(\theta, F) \sum_{i=1}^{n} w_i' = -\sum_{i=1}^{n} w_i' \times y_i$$

where $w_i' = w'(y_i - \theta)$.

$$J = \frac{d}{d\theta} h(\hat{\theta}, F) = -\frac{\sum_{i=1}^{n} (y_i - \hat{\theta})\, w_i'}{\sum_{i=1}^{n} w_i} \tag{2.6}$$

$$1 - J = \frac{\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (y_i - \hat{\theta})\, w_i'}{\sum_{i=1}^{n} w_i}$$

True Influence Function using Jorgensen's Method

True influence function, $IF_{T,F_n}(y)$

$$\begin{aligned} IF_{T,F_n}(y) &= [1 - J]^{-1} IF_{h,F_n}(y) \\ &= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (y_i - \hat{\theta})\, w_i'} \frac{(y - \hat{\theta}) w}{(1/n) \sum_i w_i} \\ &= \frac{(y - \hat{\theta})\, w}{\frac{1}{n} \left[\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (y_i - \hat{\theta})\, w_i' \right]} \end{aligned}$$

The influence function of an M-estimate of location [27] is

$$IF_{\theta,F}(y) = \frac{\psi(y - \theta)}{E_F(\psi'((y - \theta)))}$$

We know $\psi(r) = w(r)r$, where $r = y - \theta$. Hence $\psi'(r) = w'(r)r + w'(r)$

$$IF_{\theta,F_n}(y) = \frac{rw(r)}{\frac{1}{n}(\sum_{i=1}^{n} w'(r)r + \sum_{i=1}^{n} w(r_i))}$$

This result is exactly same as Jorgensen's result.

2.3.4 Numerical Illustration

Based on the series of previous results, we compute the one-step influence function and the Jacobian matrix followed by the true influence function. Later, we will compute the empirical influence function using the definition

of (2.1). We describe this situation numerically by considering the simple example of a location estimate, even though they are theoretically same.

Consider the Example 1.1 to compute the influence function of the Huber location estimate. Before computing the influence function, we need to estimate the location parameter θ. In this situation, we do not know the scale parameter which is required to avoid the scale equivariant problem. Hence, robust statistics (s = mad = 4.45) for this parameter is computed before estimating the location parameter. The location estimate, $\hat{\theta}$, by the Huber method, is 27.38.

We investigate the effect on the estimate throughout the sample space. We generated the data points from -60 to 60 in steps of 0.01. Results are given in Figure 2.2. Figure 2.2(d) shows that the plot of empirical influence values using (2.1) versus influence values obtained by Jorgensen's method gives a straight line with 45^0 degree slope. That means, the influence function obtained by Jorgensen's method tallies with the original definition of the influence function.

2.4 Asymptotic Results for M-estimates

Result: ([27], page 14)

$$\sqrt{n}(T(F_n) - T(F)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF_{T,F}(y) + R_n$$

Since the term R_n, which is the remaining terms of the Taylor expansion, is asymptotically negligible, $\sqrt{n}(T(F_n) - T(F))$ is asymptotically normal with mean zero and variance, $V[T(F)]$.

$$V[T(F)] = E_F(IF_{T,F}^2)$$

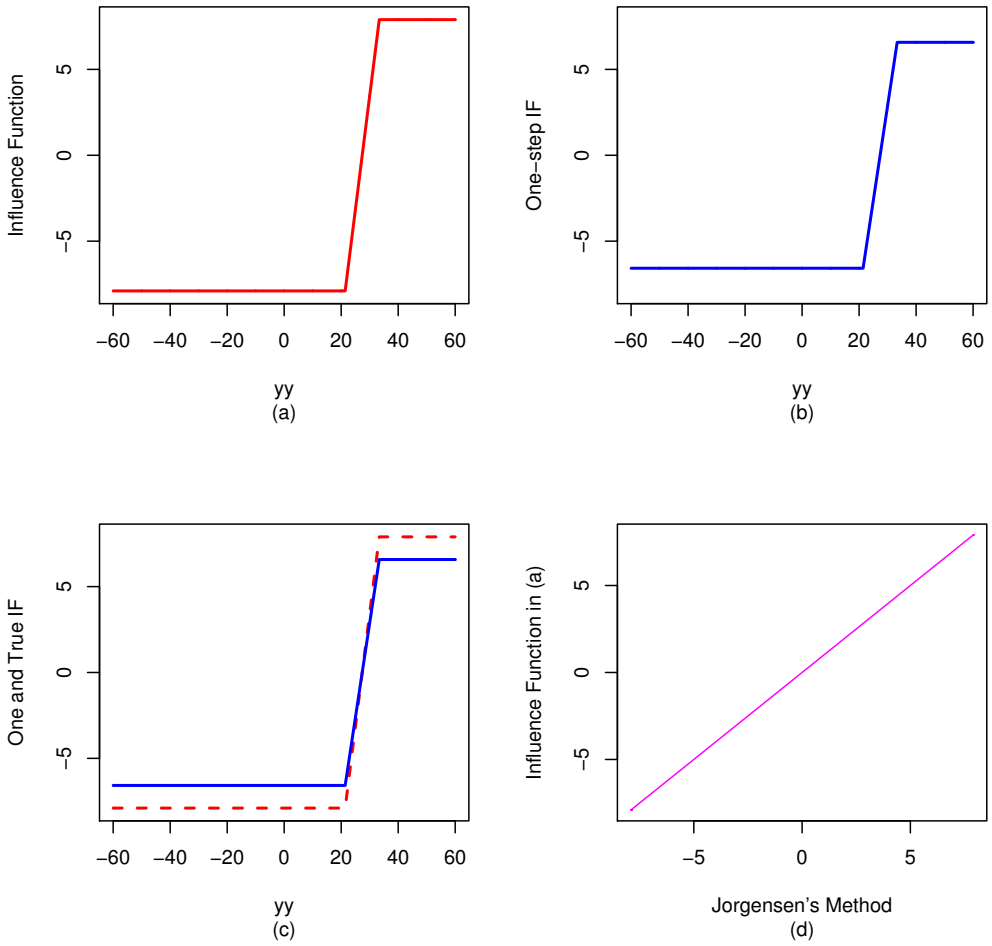

Figure 2.2: Influence functions for Huber location estimator of speed of light data: (a) Influence Function for $\hat{\theta}$ obtained using (2.1); (b) Influence Function for the one-step estimate obtained using (2.1); (c) True and one-step influence function for $\hat{\theta}$ obtained using Jorgensen's method; (d) Comparison between influence functions obtained by (a) and Jorgensen's method.

In the speed of light example, $T(F) = \theta$, $T(F_n) = \hat{\theta} = 27.38$, $s = 4.45$ and $IF_{T,F_n}(y) = \frac{s \times \psi\left(\frac{y-\hat{\theta}}{s}\right)}{\frac{1}{n}\sum_{i=1}^{n} \psi'\left(\frac{y_i-\hat{\theta}}{s}\right)}$ ([35], page 56). Hence

$$\begin{aligned} V(\hat{\theta}) &= \frac{4.45^2}{0.758^2} E_{F_n}\left[\psi^2\left(\frac{y-\hat{\theta}}{s}\right)\right] \\ &= \frac{4.45^2}{0.758^2} \times \frac{1}{n}\sum \psi^2\left(\frac{y_i-\hat{\theta}}{s}\right) \\ &= \frac{4.45^2}{0.758^2} \times 0.778 \\ &= 26.81 \end{aligned}$$

A robust 95% confidence interval for θ based on the previously computed scale parameter is $[26.1, 28.6] (= 27.38 \pm 1.96 \times \sqrt{\frac{26.81}{66}})$.

2.5 Gross Error Sensitivity

The maximum absolute value of the influence function of the statistics T over the sample space, is called *gross error sensitivity*, and denoted as γ^*.

$$\gamma^* = \sup_y |IF_{T,F}(y)|$$

If the influence function of T is bounded ((i.e.) $\gamma^* < \infty$), the estimate T is called *B-robust*. If $\gamma^* = \infty$, then the estimator is completely intolerant of outliers; a single outlier can ruin the estimator.

Chapter 3

Mixture Models and the EM Algorithm

In this chapter, we consider classification of the subjects into groups, followed by the finite mixture model and its applications. Later, we describe how to estimate the mixture model parameters.

3.1 Clustering

The classification of observations into groups, which means observations are similar to each other within a group in some way, is an important part of study in statistics, particularly where there is no prior information about the underlying group structure. Such a classification in statistics is called *cluster analysis*. There are many different methods of cluster analysis. These methods can be accommodated into two classes: (i) Hierarchical techniques; (ii) Non-hierarchical techniques. The main difference between these two classes is that the classification into groups can be optimized in the non-hierarchical techniques, by considering the reallocation of observations to other possible groups based on the statistical criteria; but an observation once assigned to one group, is not allowed to move to another group in the hierarchical techniques. Further information about these techniques is found in ([38], section 1.3), [39] and [28]. The mixture model approach for clustering is a non-hierarchical technique, and is recognised as perhaps the best method for grouping observations.

3.2 Mixture Models

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$ be a $p-$dimensional random variable and let $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{ip})$ be the i^{th} observation, where $i = 1, \ldots, n$. Assume that the probability density function, $f(.)$, of an observation $\mathbf{y}$ is of the form

$$f(\mathbf{y}) = \pi_1 f_1(\mathbf{y}, \theta_1) + \pi_2 f_2(\mathbf{y}, \theta_2) + \cdots + \pi_k f_k(\mathbf{y}, \theta_k) \tag{3.1}$$

where θ_j is a vector of unknown parameters in the j^{th} group;

$$f_j(\mathbf{y}, \theta_j) \geq 0, \quad \int f_j(\mathbf{y}, \theta_j) d\mathbf{y} = 1, \quad j = 1, \ldots, k$$

; and

$$\pi_j > 0 \;\; j = 1, \ldots, k; \quad \sum_{j=1}^{k} \pi_j = 1$$

This is known as a *finite mixture density function.* Let $\boldsymbol{\phi} = (\boldsymbol{\theta^t}, \boldsymbol{\pi^t})^{\boldsymbol{t}}$ be the unknown model parameters in the mixture model to be estimated, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^t$ is associated with the parametric form and $\theta_j = (\theta_{1j}, \ldots, \theta_{l_j j})^t$ is a vector for all $j = 1, \ldots, k$, and $\boldsymbol{\pi}= (\pi_1, \ldots, \pi_k)^t$ is a mixing proportion. The mixture density function above with parameters may then be written as follows

$$f(\mathbf{y}, \boldsymbol{\phi}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y}, \theta_j) \tag{3.2}$$

Finite mixture models are often applied in one of two ways, described as direct or indirect ([50], page 2). In direct applications, we believe that the observations come from one of these k underlying groups. In indirect applications, the mixture model is used as a mathematical tool to obtain a flexible, tractable form of analysis. This is used in robust literature. For example, heavy tailed normal distributions may be defined as a contaminated normal distribution (see section 1.1). Next we describe how to estimate the mixture model parameters.

3.3 Parameter Estimation

The MLE method is often used for estimating the vector parameter $\boldsymbol{\phi}$. The likelihood function of the $\boldsymbol{\phi}$ based on the observed data $\boldsymbol{y}$ is

$$\mathcal{L}_o(\boldsymbol{\phi}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(\mathbf{y_i}, \theta_j)$$

It is not easy to handle $\mathcal{L}_o$ to obtain the estimates, because the set of score equations cannot be solved explicitly, and the likelihood for the mixture model is often unbounded. This means, the maximum likelihood estimator does not exist. For further details about this problem see [16] , [38], [50].

If the data were fully categorized, the MLE would have an explicit form. The completely categorized data $\mathbf{x_i}$ may be expressed as ordered pairs $(\mathbf{y_i}, z_i)$, where $z_i = z_{ij}, i = 1, \ldots n$, and $j = 1, \ldots, k$, where

$$z_{ij} = \begin{cases} 1 & \text{if } y_i \in j^{th} \text{ group} \\ 0 & \text{if otherwise} \end{cases}$$

The density function $f_X(\mathbf{x})$ of the random variable X is a joint probability density function $f_{YZ}(\mathbf{y}, z)$ of the random variables Y and Z.

$$f_X(\mathbf{x_i}) = f_{Y,Z}(\mathbf{y_i}, z_{ij}) = \prod_{j=1}^{k} [\pi_j f_j(\mathbf{y_i})]^{z_{ij}} \text{ and}$$

$$f_Y(\mathbf{y_i}) = \sum_{j} f_{Y,Z}(\mathbf{y_i}, z_{ij}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y_i})$$

The likelihood function corresponding to $\mathbf{x_1}, \ldots, \mathbf{x_n}$ is the complete likelihood function, $\mathcal{L}_c(\boldsymbol{\phi})$, for the model (3.2)

$$\mathcal{L}_c(\boldsymbol{\phi}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{z_{ij}} [f_j(\mathbf{y_i}, \theta_j)]^{z_{ij}}$$

Hence, the complete log-likelihood function, $l_c(\boldsymbol{\phi})$, is

$$l_c(\boldsymbol{\phi}) = \log \mathcal{L}_c(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \{z_{ij} \log \pi_j + z_{ij} \log[f_j(\mathbf{y_i}, \theta_j)]\}$$

The unobservable indicator variables z_{ij} are initially treated as unknown parameters to be estimated along with parameter $\boldsymbol{\phi}$. This MLE is not consistent, because the z_{ij} increases in number with the number observations, which means there are more parameters to be estimated than there are observations. Later the alternative approach, which is described in section 3.4.1,was maintained.

$\mathcal{L}_o(\boldsymbol{\phi})$ corresponding to the marginal density of $y_1 \dots y_n$ is obtained by summing the $\mathcal{L}_c(\boldsymbol{\phi})$ over $z_1, \dots, z_n$. That means, mixture data is considered as incomplete data with missing values of indicator vector $\mathbf{z}$. Therefore, this problem can be seen as a missing value problem. The EM algorithm can be used to compute the parameters when missing values are present. This will be described in the next section.

3.4 The EM Algorithm

It is difficult to use the maximum likelihood estimator when information for part of the data is absent. The EM algorithm is a method for finding roots of a score function when information is missing. In other words, the EM algorithm is a tool to obtain the maximum likelihood estimates of the parameters iteratively when data can be viewed as incomplete.

Dempster, Laird and Rubin [13] introduced the EM algorithm and derived important fundamental properties of the algorithm. This paper is known as 'DLR' paper. The name EM algorithm, given by Dempster, Laird, and Rubin, indicates the Expectation and Maximization to be made for each iteration. The two steps for each iteration are known as E-step and M-step. Dempster, Laird and Rubin [14] show that the EM algorithm under certain distributional assumptions may be considered as an IRLS procedure.

3.4.1 Application of EM Algorithm to the Mixture Models

We describe the application of the EM algorithm for finite mixture models. This procedure may be described in the simple format below.

Step 1: Give initial values to the parameter $\boldsymbol{\phi}$.

Step 2: Estimate the membership probabilities of each observation for each component, z_{ij}.

Step 3: Estimate the parameters $\hat{\boldsymbol{\phi}}$ based on the complete log likelihood function, $l_c(\boldsymbol{\phi})$.

Step 4: Replace the initial parameter values by Step 3.

Step 5: Repeat Steps 2 to 4 until parameter values converge.

Algebraic description of these steps for the finite mixture models are given below.

Step 1: Appropriate initial values are chosen for the parameters, say $\hat{\boldsymbol{\phi}}^{(0)}$.

Step 2: E-step

$$\begin{aligned}
\hat{z}_{ij}^{(0)} &= E\left[Z_{ij}|\,\mathbf{y} \text{ and } \boldsymbol{\phi}^{(0)}\right] \\
&= \sum_{z_{ij}=0}^{1} z_{ij} \frac{f_{Y,Z}(\mathbf{y_i}, z_{ij}|\boldsymbol{\phi}^{(0)})}{f_Y(\mathbf{y_i}|\boldsymbol{\phi}^{(0)})} \\
&= \frac{f_{Y,Z}(\mathbf{y_i}, z_{ij}=1|\boldsymbol{\phi}^{(0)})}{f_Y(\mathbf{y_i}|\boldsymbol{\phi}^{(0)})} \\
&= \frac{\pi_j^{(0)} f_j(\mathbf{y_i}, \boldsymbol{\theta}^{(0)})}{\sum_{j=1}^{k} \pi_j^{(0)} f_j(\mathbf{y_i}, \boldsymbol{\theta}^{(0)})}
\end{aligned}$$

Step 3: Maximize the log-likelihood $l_c(\boldsymbol{\phi})$ replacing z_{ij} by $\hat{z}_{ij}^{(0)}$. This step is the M-step. Maximization can be easily implemented because the data is now considered as complete data.

$$\begin{aligned}
\hat{\boldsymbol{\phi}}^{(1)} &= \arg\max_{\boldsymbol{\phi}} l_c(\boldsymbol{\phi}) \\
&= \arg\max_{\boldsymbol{\pi},\, \boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{k} \{\hat{z}_{ij}^{(0)} \log \pi_j + z_{ij}^{(0)} \log[f_j(\mathbf{y_i}, \theta_j)]\}
\end{aligned}$$

Since the parameter vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are not related, the M-step separates into two maximization problems, one involving the mixing proportions $\pi_1 \ldots \pi_n$ and the other involving the parameters of $\boldsymbol{\theta}$.

Step 4: Repeat steps 2 and 3. The general forms $\hat{z}_{ij}^{(m)}$ and $\hat{\boldsymbol{\phi}}^{(m+1)}$ are defined at the m^{th} iteration.

$$\hat{z}_{ij}^{(m)} = \frac{\pi_j^{(m)} f_j(\mathbf{y_i}, \boldsymbol{\theta}^{(m)})}{\sum_{j=1}^{k} \pi_j^{(m)} f_j(\mathbf{y_i}, \boldsymbol{\theta}^{(m)})}$$

$$\hat{\boldsymbol{\phi}}^{(m+1)} = \arg\max_{\boldsymbol{\pi},\, \boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{k} \{\hat{z}_{ij}^{(m)} \log \pi_j + z_{ij}^{(m)} \log[f_j(\mathbf{y_i}, \theta_j)]\}$$

Step 5: If limit of the sequence $\left\{\hat{\boldsymbol{\phi}}^{(m)}\right\}_{m=1}^{\infty}$ exists, the estimate for $\boldsymbol{\phi}$ is

$$\hat{\boldsymbol{\phi}} = \lim_{m \to \infty} \hat{\boldsymbol{\phi}}^{(m)}$$

3.4.2 Overview of the EM Algorithm

DLR show that the likelihood function, $\mathcal{L}(\boldsymbol{\phi})$, is an non-decreasing sequence,

$$\mathcal{L}(\boldsymbol{\phi}^{(\boldsymbol{m+1})}) \geq \mathcal{L}(\boldsymbol{\phi}^{(\boldsymbol{m})})\ m = 1, 2, \ldots$$

and it will converge. Note that the convergence of the sequence may depend on the starting values, especially if the likelihood function has more than one maximum value. Moreover, DLR show that convergence is linear, with a rate of convergence proportional to the maximal fraction of missing information. This implies that the convergence of the EM algorithm may be a slow process. However, the EM algorithm is simple and numerically stable compared with existing numerical methods such as Newton-Raphson and Fisher's method of scoring. The methods are compared in ([50], page 88).

Unlike the Newton-Raphson and Fisher's scoring method, the EM algorithm does not need a storage space for the matrix, of second derivatives of the likelihood or inverse Fisher's information matrix. But the observed information matrix, which is a useful result to compute the standard errors of the parameters, may not be easily obtained in the EM algorithm. However, Louis [34] presents a procedure for extracting the observed information matrix when using the EM algorithm. Later, Oakes [43] discusses how to obtain the observed information matrix directly for the case of EM algorithm. McLachlan and Basford [38] discuss the observed information matrix for mixture models.

3.4.3 Remark on Starting Values

In general, no optimization algorithm, including the EM algorithm, is guaranteed to converge to a local maximum. However, the EM algorithm may converge to a local maximum under fairly general conditions, and good choice of starting values from the parameter space. To ensure that the local maximum is achieved, we have to try a number starting values for the EM algorithm and look for the most common result. Such a unique result is considered as the local maximum.

Everitt and Hand [16] found that two sets of starting values may give two different sets of final values. This may happen due to the existence of multiple roots. It is common in mixture models, where the likelihood equation has multiple roots. An obvious choice is the one with the largest maximum. Although many starting values may need to be explored to increase the chance of obtaining the global optimum.

Peters and Walker ([44], [45]) find that consistent estimates $\hat{\boldsymbol{\phi}}$ for $\boldsymbol{\phi}$, may be obtained by choosing initial estimates close enough to $\hat{\boldsymbol{\phi}}$. McLachlan and Basford [38] give detail about choosing the initial values for mixture models.

3.4.4 Information Matrices

In the EM algorithm, there are three sets of data: incomplete data, complete data and missing data. Three information matrices, the observed information matrix, $\boldsymbol{I_o(\phi, y)}$, the complete information matrix, $\boldsymbol{I_c(\phi, x)}$, and the missing information matrix, $\boldsymbol{\mathcal{I}_m(\phi, y)}$, for the EM estimates can be defined and given below.

$$\text{Observed Information Matrix: } \boldsymbol{I_o(\phi, y)} = -\frac{\partial^2[l_o(\boldsymbol{\phi})]}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}$$

$$\text{Complete Information Matrix: } \boldsymbol{I_c(\phi, x)} = -\frac{\partial^2[l_c(\boldsymbol{\phi})]}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}$$

$$\text{Missing Information Matrix: } \mathcal{I}_m(\boldsymbol{\phi, y}) = -E_{\boldsymbol{\phi}}\left[\frac{\partial^2[\log f_{X|Y}(\boldsymbol{x}|\, \boldsymbol{y}, \boldsymbol{\phi})]}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}|\, y\right]$$

where $f_{X|Y}$ is the conditional probability density function. Note that analytical evaluation of the right hand side of the $\boldsymbol{I_o(\phi, y)}$ is not easy, particularly

for mixture models. For this reason, the observed information matrix is extracted from the complete information matrix and from the expected missing information matrix. The link for these three quantities is

$$\boldsymbol{I_o}(\boldsymbol{\phi}, \boldsymbol{y}) = \mathcal{I}_c(\boldsymbol{\phi}, \boldsymbol{y}) - \mathcal{I}_m(\boldsymbol{\phi}, \boldsymbol{y}) \tag{3.3}$$

where

$$\mathcal{I}_c(\boldsymbol{\phi}, \boldsymbol{y}) = E_{\boldsymbol{\phi}} \left[\mathbf{I}_c(\boldsymbol{\phi}, X) | \, \mathbf{y} \right]$$

For further detail about information matrices of the EM estimates, see McLachlan and Krishnan ([40], Chapter 4). The rate of convergence of the EM algorithm is given by the largest eigenvalue of $\mathcal{I}_c^{-1}(\boldsymbol{\phi}, \boldsymbol{y}) \mathcal{I}_m(\boldsymbol{\phi}, \boldsymbol{y})$.

DLR show that if $\boldsymbol{\phi}^{(\boldsymbol{k})} \rightarrow \hat{\boldsymbol{\phi}}$ as $k \rightarrow \infty$ then

$$J(\hat{\boldsymbol{\phi}}) = \mathcal{I}_c^{-1}(\hat{\boldsymbol{\phi}}, \boldsymbol{y}) \mathcal{I}_m(\hat{\boldsymbol{\phi}}, \boldsymbol{y})$$

where J is a Jacobian matrix, which is easy to compute numerically (see, McLachlan and Krishnan ([40], section 4.5.2). Hence (3.3) can be written as

$$\boldsymbol{I_o}(\hat{\boldsymbol{\phi}}, \boldsymbol{y}) = \mathcal{I}_c(\hat{\boldsymbol{\phi}}, \boldsymbol{y})[I - J(\hat{\boldsymbol{\phi}})] \tag{3.4}$$

where I is an identity matrix with a dimension equal to the number of parameters in the model.

3.4.5 Examples using the EM Algorithm

In this section, examples such as contaminated models and mixtures of multivariate normal distribution, are considered to explain the usage of the EM algorithm. These examples help us to understand our proposed robust statistics.

Contaminated Normal Model

Huber [25] uses a contaminated normal model with known parameters to generate contaminated data. In contrast, Little and Rubin ([33], section 10.5), and Meng and Rubin [41] considered the estimation of the unknown parameters for the contaminated data. Later McLachlan and Krishnan [40] use this example to compute the observed information $\boldsymbol{I_o}$. It is useful to discuss this in detail for motivating our proposed robust estimates in the

following chapters. Here all parameters are treated as unknown and in the original scale.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a random sample of size n from the univariate contaminated normal model, which is obtained from (3.2). In the expression (3.2), we have $k = 2$, $\pi_2 = 1 - \pi_1$, $\theta_1 = (\mu, \sigma)^t$, $\theta_2 = (\mu, \alpha * \sigma)^t$ and $f_1 = f_2 = \Phi$, where α is a constant. For convenient notation, we use $\pi = \pi_1$ and $z_i = z_{i1}$.

$$\begin{aligned} l_c(\boldsymbol{\phi}) &= \text{constant} - \frac{n}{2}\log\sigma^2 + \sum_{i=1}^{n}[-\frac{z_i}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2 - \frac{1-z_i}{2c}\left(\frac{y_i-\mu}{\sigma}\right)^2 \\ &\quad + z_i\log\pi + (1-z_i)\log(1-\pi)] \end{aligned}$$

where $c = \alpha^2$. z_i at the m^{th} iteration can be computed at the E-step of the EM algorithm.

$$\hat{z}_i^{(m)} = \frac{\pi\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp(-\frac{1}{2}[d_i^{(m)}]^2)}{\pi\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp(-\frac{1}{2}[d_i^{(m)}]^2) + (1-\pi)\frac{1}{\sqrt{2\pi c}\sigma^{(m)}}\exp(-\frac{1}{2c}[d_i^{(m)}]^2)}$$

where $d_i^{(m)} = \frac{y_i - \mu^{(m)}}{\sigma^{(m)}}$. At the M-step, π, μ and σ can be obtained by solving the following score functions

$$\begin{aligned} \mathbf{U}_{c,\pi} &= \frac{\partial l_c(\boldsymbol{\phi})}{\partial\pi} = 0 \\ \mathbf{U}_{c,\pi} &= \sum_{i=1}^{n}\frac{z_i - n\pi}{\pi(1-\pi)} = 0 \\ \Rightarrow \pi^{(m+1)} &= \frac{\sum_{i=1}^{n}\hat{z}_i^{(m)}}{n} \end{aligned}$$

$$\begin{aligned} \mathbf{U}_{c,\mu} &= \frac{\partial l_c(\boldsymbol{\phi})}{\partial\mu} = 0 \\ \mathbf{U}_{c,\mu} &= \frac{1}{c\sigma^2}\sum_{i=1}^{n}(y_i-\mu)(cz_i - z_i + 1) = 0 \\ \Rightarrow \mu^{(m+1)} &= \frac{\sum_{i=1}^{n}[\hat{z}_i^{(m)}c - \hat{z}_i^{(m)} + 1]y_i}{\sum_{i=1}^{n}[\hat{z}_i^{(m)}c - \hat{z}_i^{(m)} + 1]} \end{aligned}$$

$$
\begin{aligned}
\mathbf{U}_{c,\sigma^2} &= \frac{\partial l_c(\boldsymbol{\phi})}{\partial \sigma^2} = 0 \\
\mathbf{U}_{c,\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2c\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2(cz_i - z_i + 1) = 0 \\
\Rightarrow \sigma^{(m+1)^2} &= \frac{\sum_{i=1}^{n}[\hat{z}_i^{(m)}c - \hat{z}_i^{(m)} + 1](y_i - \mu^{(m+1)})^2}{nc}
\end{aligned}
$$

Therefore, the estimates may be defined: $\hat{\pi} = \lim_{m\to\infty}\pi^{(m)}$, $\hat{\mu} = \lim_{m\to\infty}\mu^{(m)}$, and $\hat{\sigma}^2 = \lim_{m\to\infty}\sigma^{(m)^2}$.

Mixtures of Multivariate Normal Distributions

This is an extension of the contaminated model in terms of the number of variables and the number of mixtures. We assume that the density of the components for the mixture is a normal density with different parameters. The dimension of the data matrix is $(n \times p)$. The standard multivariate normal density function is

$$
f_j(\mathbf{y}, \theta_j) = \frac{1}{(2\pi)^{p/2}|\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu_j)^t\sum\nolimits_j^{-1}(\mathbf{y}-\mu_j)\right\} \quad j = 1, 2, ...k
$$

where μ_j and $\sum_j$ are the $p \times 1$ column vector of mean and $(p \times p)$ covariance matrix of j^{th} component of the mixture respectively.

$$
\begin{aligned}
\mu_j &= (\mu_{1j}, \mu_{2j}, \ldots, \mu_{pj})^t \\
\\
\boldsymbol{\phi} &= (\theta_1, \theta_2, \ldots, \theta_k, \pi_1, \pi_2, \ldots, \pi_{k-1})
\end{aligned}
$$

where $\theta_j = (\mu_j, \sum_j)$ is an ordered pair. The log likelihood functions for observed and completed data respectively are given below

$$
\begin{aligned}
l_o(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\log\left\{\sum_{j=1}^{k}\pi_j f_j(\mathbf{y_i}, \theta_j)\right\} \\
l_c(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\{-\frac{z_{ij}}{2}\left[(\mathbf{y_i}-\mu_j)^t\sum\nolimits_j^{-1}(\mathbf{y_i}-\mu_j) + \log(|\sum\nolimits_j|) + \log(2\pi)\right] \\
&\quad + z_{ij}\log(\pi_j)\}
\end{aligned}
$$

The EM algorithm can be applied to $l_c(\boldsymbol{\phi})$ where z_{ij} are treated as missing observations. Therefore, the E-step gives estimated values for the missing observations z_{ij}, given below at m^{th} iteration.

$$\hat{z}_{ij}^{(m)} = \frac{\pi_j^{(m)} f_j(\mathbf{y}_i, \theta_j^{(m)})}{\sum_{j=1}^{k} \pi_j^{(m)} f_j(\mathbf{y}_i, \theta_j^{(m)})}$$

Since our intention is to estimate the parameter $(\boldsymbol{\phi})$, it can be obtained at the M-step and is given below.

$$\begin{aligned}
\hat{\pi}_j &= \frac{\sum_{i=1}^{n} \hat{z}_{ij}}{n} \\
\hat{\mu}_j &= \frac{\sum_{i=1}^{n} \hat{z}_{ij} \mathbf{y_i}}{\sum_{i=1}^{n} \hat{z}_{ij}} \\
\hat{\sum}_j &= \frac{\sum_{i=1}^{n} \hat{z}_{ij} (\mathbf{y_i} - \hat{\mu}_j)(\mathbf{y_i} - \hat{\mu}_j)^t}{\sum_{i=1}^{n} \hat{z}_{ij}}
\end{aligned}$$

More specifically, the r^{th} element of the vector $\hat{\mu}_j$ is given by:

$$\hat{\mu}_{rj} = \frac{\sum_{i=1}^{n} \hat{z}_{ij} y_{ir}}{\sum_{i=1}^{n} \hat{z}_{ij}} \quad r = 1, \ldots p \text{ and } j = 1, \ldots, k$$

These results are obtained after the M-step converges to a fixed number, known as the local maxima. McLachlan and Basford [38] consider various situations for the mixtures of multivariate normal distributions, such as common covariance ($\sum = \sum_1 = \sum_2 = , \ldots, = \sum_k$) and the robust estimation for mixture models.

Chapter 4

Robust Location Estimator

A location estimator can be viewed as a summarized result or as an estimator for central tendency of a set of observations $y_1, \ldots, y_n$. One class of these is the L_p estimator, which may be defined as the value of the parameter θ which minimizes the expression E

$$E = \sum_{i=1}^{n} |y_i - \theta|^p \tag{4.1}$$

where $p \geq 0$. The cases $p = 1, 2$ yield the median and mean respectively. More generally, we may replace the function $|y-\theta|^p$ by the function $\rho(y-\theta)$, where ρ is a function (often termed a loss function) that is symmetric about 0 and increasing over the positive real numbers. Because the focus of this thesis is on the maximum likelihood estimators in statistical models, we introduce a model whose maximum likelihood estimators are location estimators of this kind.

4.1 Model

Consider the probability density function over the real numbers

$$f(y, \theta) = f(y - \theta) = k(\theta) \exp\{-\rho(y - \theta)\} \tag{4.2}$$

where $k(\theta)$ is a normalizing constant. We will consider location estimators for a data set $y_1, \ldots, y_n$ that are maximum likelihood estimates of θ with respect to this density.

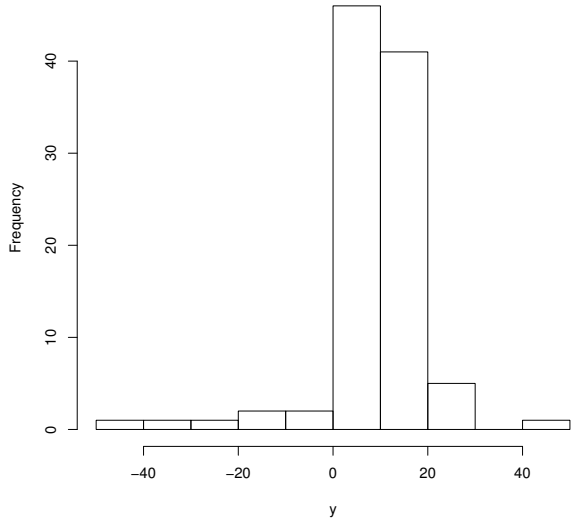

Figure 4.1: A Histogram of the generated data

For numerical illustration, a set of 80 observations randomly generated from $\mathcal{N}(\theta = 10, \sigma^2 = 9)$, which we consider to be correct, contaminated by 20 observations from $\mathcal{N}(\theta = 10, \sigma_1^2 = 3\sigma^2)$. The advantage of using generated data is that we know the true value of the location parameter, $\theta = 10$, which can be compared with the estimated value. We also know other parameters such as the scale parameter, $\sigma = 3$ and the proportion of the non-contaminated observations (0.8). A histogram for the generated data is given in Figure 4.1.

4.2 Calculating the Estimator

The parameter estimate $\hat{\theta}$ is defined by

$$\hat{\theta} = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \rho(y_i - \theta) \right\} \tag{4.3}$$

where "arg min" stands for "the value which minimizes". Note that if α_i is a weight associated with the observation y_i for all $i = 1 \ldots n$, then we will minimise (4.4) to obtain a weighted version of the estimator.

$$\sum_{i=1}^{n} \alpha_i \, \rho(y_i - \theta) \tag{4.4}$$

A necessary condition for $\hat{\theta}$ in (4.3) to exist is

$$\sum_{i=1}^{n} \psi(y_i - \hat{\theta}) = 0 \tag{4.5}$$

where $\psi(r) = \rho'(r)$. In general, ψ is non-linear except for the case $\rho(r) = \frac{1}{2}r^2$. One of the methods to solve the non-linear equation (4.5) is to use iteratively re-weighted least squares [5]. Let

$$w(r) = \begin{cases} \frac{\psi(r)}{r} & \text{if} \quad r \neq 0 \\ \\ \psi'(r) & \text{if} \quad r = 0 \end{cases} \tag{4.6}$$

Then (4.5) can be written as

$$\sum_{i=1}^{n} w(y_i - \hat{\theta}) \, (y_i - \hat{\theta}) = 0 \tag{4.7}$$

So that $\hat{\theta}$ must satisfy

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{4.8}$$

where $w_i = w(y_i - \hat{\theta})$. This is not an explicit expression for $\hat{\theta}$, because the weights w_i depends on $\hat{\theta}$. Hence it leads to an iterative computation for $\hat{\theta}$. This means, for a given initial value $\hat{\theta}^{(0)}$, we can get the new estimate of $\hat{\theta}^{(1)}$ using the equation (4.9)

$$\hat{\theta}^{(1)} = \frac{\sum_{i=1}^{n} w(y_i - \hat{\theta}^{(0)}) y_i}{\sum_{i=1}^{n} w(y_i - \hat{\theta}^{(0)})} \tag{4.9}$$

The limit of the sequence $\{\hat{\theta}^{(m)}\}_{m=1}^{\infty}$, if it converges, is the estimate for θ. Holland and Welsch [24] explain this method for various ρ functions.

However, it is an explicit equation for some situations. For example, if $\rho(r) = \frac{1}{2}r^2$ then $\psi(r) = r$ and $w(r) = 1$, which does not depend on θ. In this case, we do not need an iterative process and the estimate is the mean, $\bar{y}$.

4.3 Influence Function of the Estimator

The influence function of a location estimator, $\hat{\theta}$ defined in (4.3), is given by [27] as

$$IF_{\hat{\theta}}(y, F) = \frac{\psi(y - \hat{\theta})}{E_F[\psi'(y - \hat{\theta})]} \tag{4.10}$$

Notice that the influence function is proportional to the ψ function, so if ψ is unbounded then the influence function is unbounded.

Let F_n be the empirical cumulative distribution function of $y_1, \ldots, y_n$ then

$$E_{F_n}[\psi'(y - \theta)] = \frac{1}{n}\sum_{i=1}^{n} \psi'(y - \theta) \tag{4.11}$$

In the case of MLE, where $\rho(r) = \frac{1}{2}r^2$, which leads to $\psi(r) = r$ and $\hat{\theta} = \bar{y}$, so that we have from (4.10)

$$IF_{\bar{y}}(y, F_n) = \frac{y - \bar{y}}{1} = y - \bar{y}$$

Since $IF_{\hat{\theta}}(y, F_n) \to \pm\infty$ as $y \to \pm\infty$, the influence function for MLE is unbounded. Therefore, we can say that the estimate will be heavily affected by a single large observation.

4.4 Mixture Model

Because many maximum likelihood estimates, like the mean, are sensitive to extreme observations. We seek in this thesis to find alternatives that are less sensitive to extreme observations. Our strategy is to consider a two component mixture between the nominal model for the data and a dispersed distribution over the data space. The hope is that on fitting the mixture

model, the extreme data will tend to be assigned to the dispersed component and so have their influence on the estimated parameters of the nominal model reduced. Formally, we consider the mixture model

$$p(y, \theta) = \lambda f(y, \theta) + (1 - \lambda) g(y) \tag{4.12}$$

where g is the dispersed parameter free function over the sample space and $1 - \lambda$ is a fixed small positive number which may be thought of as the proportion of contaminated data. We will often choose λ to be 0.95 or similar. Remember f is defined in (4.2).

For fixed choices of g and λ we will consider the robustness properties of

$$\tilde{\theta} = \arg\max_{\theta} L_o(\theta) \tag{4.13}$$

where $L_o(\theta)$ is the observed likelihood function for θ.

$$L_o(\theta) = \Pi_{i=1}^{n} p(y_i, \theta) \tag{4.14}$$

It is expected that this will give a robustified estimator.

4.5 Calculating the Robustified Estimator

In this section, we would like to explain how to obtain the estimator, $\tilde{\theta}$. It is complicated to maximize the observed likelihood, $L_o(\theta)$ in (4.13), so we consider the complete likelihood $L_c(\theta)$ to achieve (4.13)

$$\tilde{\theta} = \arg\max_{\theta} L_c(\theta) \tag{4.15}$$

where

$$L_c(\theta) = \Pi_{i=1}^{n} \left[[\lambda f(y_i, \theta)]^{z_i} \times [(1 - \lambda) g(y_i)]^{1 - z_i} \right] \tag{4.16}$$

and

$$z_i = \begin{cases} 1 & \text{if } y_i \in f \\ 0 & \text{if } y_i \in g \end{cases} \tag{4.17}$$

In general, z_i's are treated as unobserved random variables. They can be estimated using the E-step of EM algorithm and then θ is estimated at the M-step.

4.5.1 E-Step

Let $q(y, z)$ be a joint probability density function of random variables Y and Z, where θ is fixed

$$q(y, z) = [\lambda f(y, \theta)]^z \times [(1 - \lambda)g(y)]^{1-z}$$

The marginal distribution of y is defined as

$$p(y) = \sum_{z=0}^{1} q(y, z) = \lambda f(y, \theta) + (1 - \lambda)g(y)$$

We use the expectation of z given y to derive an estimator for z

$$\begin{aligned} \tilde{z} = E[z|y] &= \sum_{z=0}^{1} zp(z|y) \\ &= 0 \times p(0|y) + 1 \times p(1|y) \\ &= p(1|y) \\ &= \frac{q(y, 1)}{p(y)} \quad \text{[Bayes Theorem]} \end{aligned}$$

$$\tilde{z} = \frac{\lambda f(y, \theta)}{\lambda f(y, \theta) + (1 - \lambda)g(y)} \tag{4.18}$$

In simple language, we can say that $\tilde{z}$ is a function of y and θ and denote $\tilde{z} = z(y, \theta)$.

4.5.2 M-Step

The M-step is to maximize the complete likelihood function (4.16) with z_i replaced by $\tilde{z}_i$ for $i = 1, \ldots, n$

$$\begin{aligned} l_c(\theta) &= \log L_c(\theta) \\ &= \sum_{i=1}^{n} \tilde{z}_i \log f(y_i - \theta) + \text{constant} \end{aligned}$$

For MLE, therefore

$$l_c(\theta) = \sum_{i=1}^{n} \tilde{z}_i (y_i - \theta)^2 + \text{constant} \tag{4.19}$$

In order to maximize the $l_c(\theta)$ with respect to θ, $\left[\frac{dl_c(\theta)}{d\theta}\right]_{\theta=\hat{\theta}} = 0$, which gives (4.20)

$$\sum_{i=1}^{n} \tilde{z}_i(y_i - \theta) = 0 \tag{4.20}$$

This leads to the simple form of θ given in (4.21)

$$\theta = \frac{\sum_{i=1}^{n} \tilde{z}_i y_i}{\sum_{i=1}^{n} \tilde{z}_i} \tag{4.21}$$

Since $\tilde{z}_i$ in the right hand side of (4.21) does depend on θ, we iterate the E-step of (4.18) and the M-step of (4.21) until the updates no longer change the parameter estimates. For a given initial value for θ, say $\tilde{\theta}^{(0)}$, the new estimate $\tilde{\theta}^{(1)}$ is obtained by

$$\tilde{\theta}^{(1)} = \frac{\sum_{i=1}^{n} \tilde{z}_i^{(0)} y_i}{\sum_{i=1}^{n} \tilde{z}_i^{(0)}}$$

where $\tilde{z}_i^{(0)} = z(y_i, \tilde{\theta}^{(0)})$ for $i = 1, \ldots, n$. At the m^{th} process of the iteration, we can write

$$\tilde{\theta}^{(m+1)} = \frac{\sum_{i=1}^{n} \tilde{z}_i^{(m)} y_i}{\sum_{i=1}^{n} \tilde{z}_i^{(m)}} \tag{4.22}$$

where $\tilde{z}_i^{(m)} = z(y_i, \tilde{\theta}^{(m)})$ for $m = 0, 1 \ldots$. If the limit of sequence of $\left\{\tilde{\theta}^{(m)}\right\}_{m=0}^{\infty}$ exists, the estimate, $\tilde{\theta}$, for θ is taken as the limit

$$\tilde{\theta} = \lim_{m \to \infty} \tilde{\theta}^{(m)}$$

4.6 Influence Function for $\tilde{\theta}$

The $\tilde{\theta}$ is computed iteratively so that Jorgensen's [30] method is employed to compute the influence function for $\tilde{\theta}$. Before applying this method, we need to define the updating function.

Let $h(\theta, F_n)$ be the updating function defined by (4.23)

$$h(\theta, F_n) = \frac{\sum_{i=1}^{n} z(y_i, \theta) y_i}{\sum_{i=1}^{n} z(y_i, \theta)} \tag{4.23}$$

In simple language, we can write $\theta^{(m+1)} = h(\theta^{(m)}, F_n),\ \ m = 0, 1, 2, \ldots,$ and the statistical functional $\breve{\theta}(F_n)$ is a statistic obtained from one step to the next step when $\theta = \tilde{\theta}$. This means

$$\breve{\theta}(F_n) = h(\tilde{\theta}, F_n) = \tilde{\theta} \tag{4.24}$$

In order to compute the true influence function of $\tilde{\theta}$ using the Jorgensen's method, we need to compute two important components, the one-step influence function and a Jacobian matrix, which are discussed in the sections 4.6.1 and 4.6.2.

4.6.1 One-Step Influence Function for $\tilde{\theta}$

Consider the perturbed data set $y_1, \ldots, y_n, y$ with weights $\frac{1-\epsilon}{n}, \ldots, \frac{1-\epsilon}{n}, \epsilon$. The new estimate can be derived from the likelihood concept for this case. However, the new estimate can be easily defined by (4.25) based on (4.23) and (4.24)

$$\breve{\theta}((1-\epsilon)F_n + \epsilon\Delta_y) = h(\tilde{\theta}, (1-\epsilon)F_n + \epsilon\Delta_y) = \frac{\frac{1-\epsilon}{n}\sum_{i=1}^{n} z_i y_i + \epsilon\, z\, y}{\frac{1-\epsilon}{n}\sum_{i=1}^{n} z_i + \epsilon y} \tag{4.25}$$

Note that the true influence function of $\breve{\theta}$ is an one-step influence function of $\tilde{\theta}$, which is denoted by $IF^1_{\tilde{\theta}}(y, F_n)$.

$$\begin{aligned}
IF^1_{\tilde{\theta}}(y, F_n) &= IF_{\breve{\theta}}(y, F_n) \\
&= \lim_{\epsilon\to 0} \frac{\breve{\theta}((1-\epsilon)F_n + \epsilon\Delta_y) - \breve{\theta}(F_n)}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\epsilon\left[zy\sum_{i=1}^{n} z_i - z\sum_{i=1}^{n} z_i y_i\right]}{\epsilon\left[\sum_{i=1}^{n} z_i\right]\left[\frac{1-\epsilon}{n}\sum_{i=1}^{n} z_i + \epsilon z\right]} \\
&= \frac{\left[zy\sum_{i=1}^{n} z_i - z\sum_{i=1}^{n} z_i y_i\right]}{\frac{1}{n}\left[\sum_{i=1}^{n} z_i\right]^2} \\
&= \frac{n\, z\, (y - \tilde{\theta})}{\sum_{i=1}^{n} z_i}
\end{aligned}$$

where $z_i = z(y_i, \tilde{\theta})$ and $z = z(y, \tilde{\theta})$.

4.6.2 Jacobian Matrix

We focus only on the parameter θ here so that the Jacobian matrix is a single term matrix. In this section, the Jacobian matrix will be derived based on the formula given in (4.26)

$$J = \left[\frac{dh(\theta, F_n)}{d\theta}\right]_{\theta=\tilde{\theta}} \tag{4.26}$$

$$\begin{aligned} h(\theta, F_n) &= \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i} \\ \left[\sum_{i=1}^{n} z_i\right] h(\theta, F_n) &= \sum_{i=1}^{n} z_i y_i \end{aligned}$$

Differentiate both sides of the above equation with respect to θ, we will get

$$\begin{aligned} \left[\sum_{i=1}^{n} z_i'\right] h(\theta, F_n) + \left[\sum_{i=1}^{n} z_i\right] \frac{dh(\theta, F_n)}{d\theta} &= \sum_{i=1}^{n} z_i' y_i \\ \left[\sum_{i=1}^{n} z_i\right] \frac{dh(\theta, F_n)}{d\theta} &= \sum_{i=1}^{n} z_i'(y_i - h(\theta, F_n)) \\ \left[\frac{dh(\theta, F_n)}{d\theta}\right]_{\theta=\tilde{\theta}} &= \frac{\sum_{i=1}^{n} z_i'(y_i - \tilde{\theta})}{\sum_{i=1}^{n} z_i} \\ J &= \frac{\sum_{i=1}^{n} z_i'(y_i - \tilde{\theta})}{\sum_{i=1}^{n} z_i} \end{aligned}$$

where $z_i' = \frac{dz_i}{d\theta}$ and $z_i = \frac{\lambda f(y_i,\theta)}{\lambda f(y_i,\theta)+(1-\lambda)g(y_i)}$.

$$\begin{aligned} z_i' = \frac{dz_i}{d\theta} &= \frac{\lambda(1-\lambda)g(y_i)f'(y_i,\theta)}{[\lambda f(y_i,\theta) + (1-\lambda)g(y_i)]^2} \\ &= \frac{\lambda(1-z_i)f'(y_i,\theta)}{\lambda f(y_i,\theta) + (1-\lambda)g(y_i)} \end{aligned}$$

4.6.3 True Influence Function for $\tilde{\theta}$

The true influence function for $\tilde{\theta}$ is denoted by $IF_{\tilde{\theta}}(y, F_n)$.

$$\begin{aligned} IF_{\tilde{\theta}}(y, F_n) &= (1-J)^{-1} IF^1_{\tilde{\theta}}(y, F_n) \\ &= \left(\frac{\sum_{i=1}^n z_i}{\sum_{i=1}^n z_i - \sum_{i=1}^n z_i'(y_i - \tilde{\theta})} \right) \left(\frac{n\, z\, (y - \tilde{\theta})}{\sum_{i=1}^n z_i} \right) \\ &= \frac{n\, z\, (y - \tilde{\theta})}{\sum_{i=1}^n z_i - \sum_{i=1}^n z_i'(y_i - \tilde{\theta})} \end{aligned}$$

4.7 Numerical Results

We begin to illustrate the method for the generated data described in section 4.1. Our main interest here is to compute $\hat{\theta}$ and $\tilde{\theta}$ numerically when $\sigma = 3$. In addition, the influence functions for $\tilde{\theta}$ are computed.

For the mixture approach, the function g needs to be defined as a parameter free function. We know the smallest and largest values for the data set, which are -42.0743 and 46.2700 respectively, so that we have chosen a very dispersed uniform distribution with parameters $[a = -80, b = 80]$. Note that, if y goes beyond the support of g then weights z's become one and the robustness will be lost. Therefore g must be chosen after inspecting the data. This is a reason we chose the domain g beyond the minimum and maximum value of the data.

The estimates for θ based on the various λ, including true $\lambda = 0.8$, and MLE are given in Table 4.1. These results are obtained after considering various initial values of the parameter θ in order to obtain a unique result. However, the mean of the data was finally chosen as an initial value for the θ. The estimates $\tilde{\theta}$ for various λ are almost same and all values of $\tilde{\theta}$ are very close to the true value, 10. It is evident that $\tilde{\theta}$ is a better estimate than the maximum likelihood estimate, $\hat{\theta} = 8.7313$.

For the numerical distribution of the estimate $\tilde{\theta}$, this procedure is repeated a thousand times (that means a thousand new sets of data were generated) for the particular case of $\lambda = 0.95$. A summary of the $\tilde{\theta}$ is given in Table 4.2. A histogram of the $\tilde{\theta}$ is given in Figure 4.2; 96.1% (= 961/1000) of the

Table 4.1: Location estimates (MLE and Mixture) for the generated data

	$\hat{\boldsymbol{\theta}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.8}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.9}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95}}$
Estimate for $\boldsymbol{\theta}$	8.7313	9.7637	9.7736	9.7854

$\tilde{\theta}_{\lambda^*}$ is an estimate for θ by mixture approach when $\lambda = \lambda^*$

Table 4.2: Summary statistics of $\tilde{\theta}$

Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
9.182	9.819	9.982	9.985	10.151	10.758

estimated values lie between 9.5 and 10.5.

We like to determine whether $\tilde{\theta}$ is affected by a large observation or not. The true and one-step empirical influence functions for the $\tilde{\theta} = 9.7854$ are given in Figure 4.3, where $J = 0.1133$. Notice from Figure 4.3 that the influence functions are bounded. That means, the estimate $\tilde{\theta}$ is not heavily affected by extreme values.

4.7.1 Comparison of $\tilde{\theta}$ with Standard Robust Estimates

In this section, when $\lambda = 0.95$, the estimate $\tilde{\theta}$ is compared with the estimates generated by the Huber and the Tukey ρ functions, defined in section 1.4. This investigation is made using the generated data set explained in section 4.1. The standard turning constants such as $c = 1.345$ for Huber, and $c = 4.685$ for Tukey were chosen. The scale parameter σ is here chosen as the true value of 3, which was used in the mixture model too. This was used

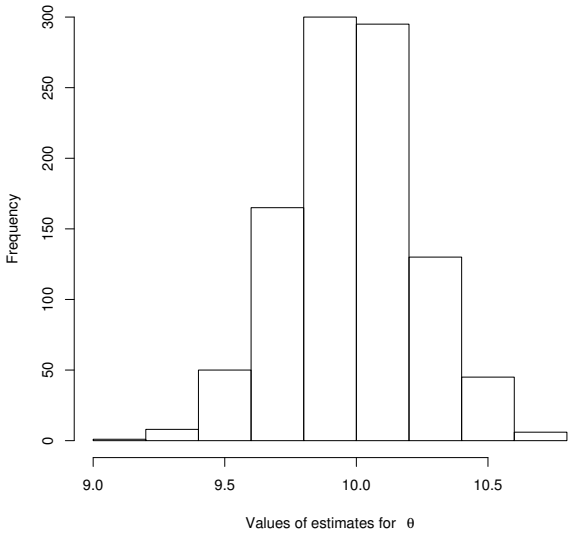

Figure 4.2: Distribution of $\tilde{\theta}$

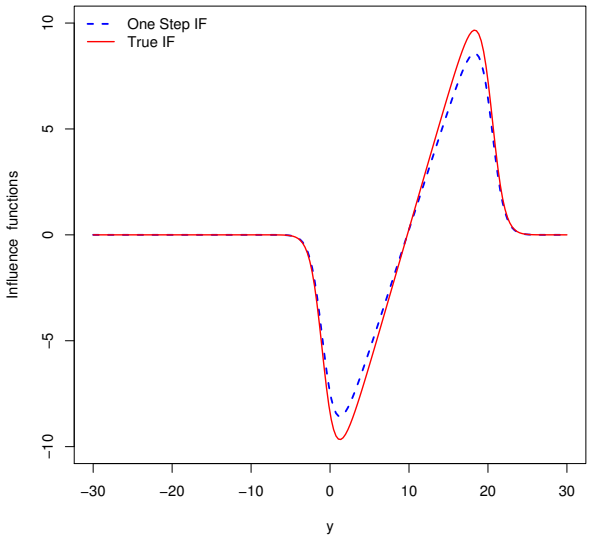

Figure 4.3: True and one-step influence functions for $\tilde{\theta}$

Table 4.3: Location estimates for the generated data

MLE	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95}}$	**Tukey estimator**	**Huber estimator**
8.7313	9.7854	9.7518	9.6620

to standardize the data to compute the Huber estimate and Tukey estimate.

Again, the mean is allocated as the starting value of the parameter θ to run the iterative algorithm. The estimates for θ are based on the various methods given in Table 4.3. The estimate $\tilde{\theta}_{0.95}$ is very similar to the Huber and Tukey estimates. We know that the influence functions for Huber and Tukey estimates are always bounded, so that attention was not given to derive the influence function for those estimates. However, we found that the influence function for $\tilde{\theta}$ has a similar pattern to the Tukey estimate. That means, the weights allocated to the very extreme observations are almost zero for both methods.

4.8 Efficiency

In this section, we are comparing the mixture estimate with the M estimates such as Huber and Tukey based on efficiency. That means, the main intention is to compute the variance of the estimates via a simulation study.

Before computing the variance of the estimate, we need to fix the λ in the mixture model, and tuning constants c in the loss functions of Huber and Tukey in order to ensure that these estimates have similar efficiency at normal. We know from the literature that the efficiency of the M-estimates such as Huber and Tukey at the normal are 5% less efficient than the mean when $c = 1.345$ and $c = 4.685$ respectively. We would like to choose the λ for computing the mixture estimate, which has 95% efficiency when sampling from a normal distribution.

4.8.1 Choosing λ

In order to fix the λ, we generate one thousand random observations from the standard normal distribution. Then we compute mean and mixture estimates for various lambda using this data. This procedure is repeated 1000 times. Now, we can able to compute the variance of these estimates, listed in Table 4.4. When $\lambda = 1$, the mixture estimate is considered as the maximum likelihood estimate. We can see from Table 4.4 that when $\lambda = 0.900$, the variance of the mixture estimate at the normal is 5% larger than the mean. We will use this λ value for comparing the efficiency of estimates.

4.8.2 Comparing Estimates Based on the Efficiency

For the comparison, a set of 100 observations is randomly generated, in which 80 are considered as "good" observations generated from the standard normal distribution and 20 are considered as "bad" observations generated from the $\mathcal{N}(0, 10)$. Estimates based on methods such as the mixture, Huber and Tukey can be computed. This procedure is repeated 1000 times to evaluate the variance of these estimates.

Table 4.5 gives the average value of the estimates and variance of the estimates. All four estimates for location are closer to zero, which is the true value of the model. The variance of the mixture estimates is smaller than other estimates, even though all three variance of the robustified statistics are quite similar.

Up to now we have investigated the efficiency of the estimate when observations are contaminated by the symmetric distribution. Next we try to examine the non-symmetric situation. Here we randomly generate "bad" data from the gamma distribution with the shape parameter = 4 and the scale parameter = 1.

Table 4.6 gives estimates of the parameter and their variance for the non-symmetric contamination situation. The variance of all four estimates are very similar; however, the mixture estimate is a much better estimate than other estimates, because the mixture estimate is close to the true value, which is zero.

Table 4.4: Variance of mixture estimates for various λ when sampling from a standard normal distribution

$\boldsymbol{\lambda}$	**Variance**	**Efficiency** $= \frac{\textbf{Variance of Mixture estimates}}{\textbf{Variance of mean}}$
0.895	0.0010340410	0.9488459
0.900	0.0010320769	0.9506516
0.905	0.0010300905	0.9524848
0.910	0.0010280799	0.9543476
0.915	0.0010260431	0.9562421
0.920	0.0010239779	0.9581706
0.925	0.0010218819	0.9601360
0.930	0.0010197522	0.9621411
0.935	0.0010175860	0.9641893
0.940	0.0010153796	0.9662845
0.945	0.0010131291	0.9684309
0.950	0.0010108301	0.9706335
0.955	0.0010084773	0.9728980
0.960	0.0010060647	0.9752310
0.965	0.0010035853	0.9776404
0.970	0.0010010305	0.9801355
0.975	0.0009983893	0.9827284
0.980	0.0009956472	0.9854349
0.985	0.0009927813	0.9882797
0.990	0.0009897449	0.9913115
0.995	0.0009863897	0.9946834
1.000	0.0009811455	1.0000000

Table 4.5: Location estimates by various methods and their variance in the case of symmetric contamination

	Mean	**Mixture**	**Huber**	**Tukey**
$\boldsymbol{\theta}$	0.00660239	0.0004006751	0.0009648057	0.001350660
$\mathbf{Var}(\boldsymbol{\theta})$	0.03137133	0.002697769	0.003463336	0.00625405

Table 4.6: Location estimates by various methods and their variance in the case of non-symmetric contamination

	Mean	**Mixture**	**Huber**	**Tukey**
$\boldsymbol{\theta}$	0.7966905	0.2730741	0.3981594	0.665913
$\mathbf{Var}(\boldsymbol{\theta})$	0.01642922	0.01618420	0.01220916	0.01201733

Chapter 5

Robust Estimation for Linear Models

In this chapter, we begin the discussion on the robust estimation of the parameters of linear regression models. We consider the standard regression model

$$y = \mu + \epsilon \tag{5.1}$$

where $\mu = X\beta$ is a linear, the response variable y is an $(n \times 1)$ vector, ϵ is an $(n \times 1)$ error vector, β is a $(p \times 1)$ unknown parameter vector and X is an $(n \times p)$ design matrix. The structure of this chapter is similar to Chapter 4.

5.1 Model

In some cases, (5.1) may be written as

$$y = X\beta + \sigma\epsilon \tag{5.2}$$

where σ is a known or previously estimated scale parameter.

We assume the random variable ϵ is distributed as the standard normal distribution. The probability density function of the random variable ϵ is given in (5.3)

$$f_{\epsilon_i}(\epsilon_i) = k \exp\{-\frac{1}{2}\epsilon_i^2\} \tag{5.3}$$

where $k = \frac{1}{\sqrt{2\pi}}$. Hence, the probability density function of Y_i is

$$\begin{aligned} F_{Y_i}(y_i) &= P[Y_i \leq y_i] \\ &= P\left[\epsilon_i \leq \frac{y_i - x_i\beta}{\sigma}\right] \\ F_{Y_i}(y_i) &= F_{\epsilon_i}\left(\frac{y_i - x_i\beta}{\sigma}\right) \\ f_{Y_i}(y_i) &= \frac{1}{\sigma} f_{\epsilon_i}\left(\frac{y_i - x_i\beta}{\sigma}\right) \end{aligned}$$

where x_i is a i^{th} row vector of the matrix X. Therefore, we can define the probability density function of Y_i as

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma}\right)^2\right\} \tag{5.4}$$

We consider regression coefficient estimates for a data set $(y_1, x_1), \ldots (y_n, x_n)$, where $x_i = (x_{i1} \ldots x_{ip})$ $i = 1 \ldots n$ is the i^{th} row vector of the matrix X, that are maximum likelihood estimates of β with respect to the density in (5.4). We start with a real example to show the ineffectiveness of the least-square estimates and proceed to find alternative robust estimates.

Example 5.1 - Belgium Phone Call Data: The source for this data is P. J. Rousseeuw and A. M. Leroy ([48], page 25), where they explain the data in detail. The total number of international phone calls (in tens of million) per year from Belgium were recorded from 1950 to 1973. It was found that another recording system (the total number of minutes of these calls) was used from 1964 to 1969. As far as total number of international phone calls point of view, the given information is invalid. This data set is available in the **R** library MASS under the name of "phones". In phones, the total number of phone calls were recorded in millions per year, which is considered as the response variable, and year is considered as the predictor variable for this analysis. Figure 5.1 shows the data and the straight line fitted by least squares to the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1 \ldots n$$

where y is the number of phone calls in millions per year and x is the year.

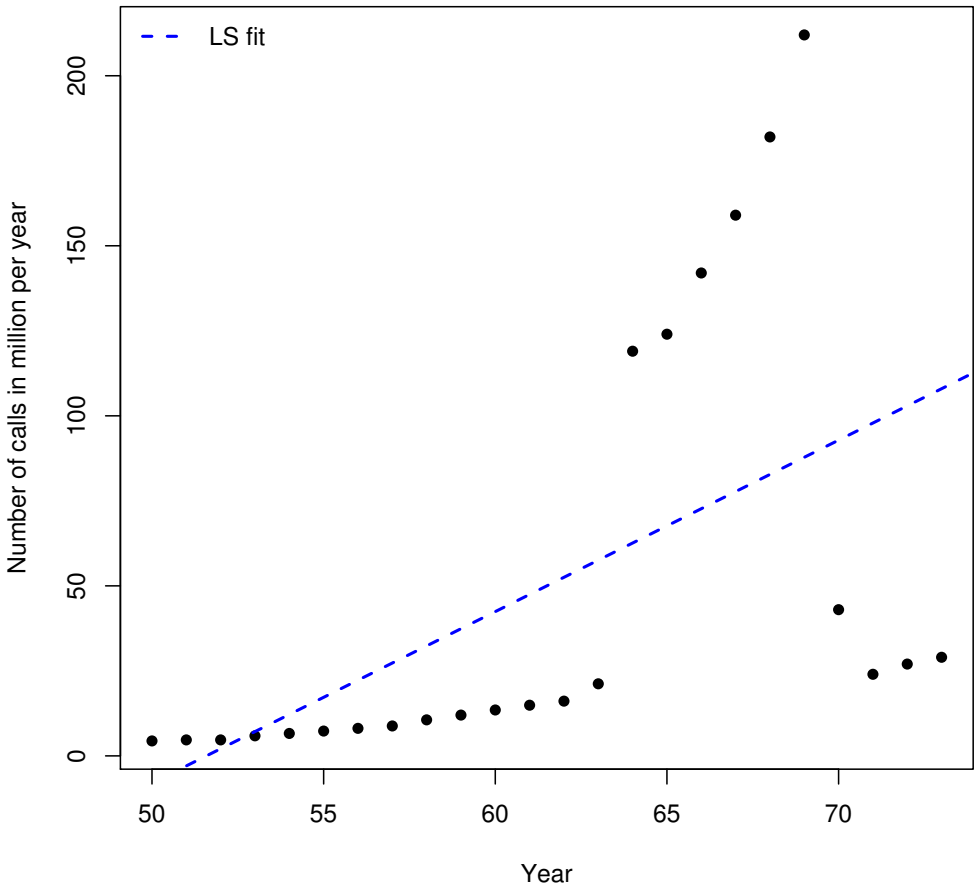

Figure 5.1: Number of international phone calls from Belgium in the years $1950 - 1973$ with least squares (LS) fit

5.2 Calculating the Estimator

The parameter estimate β is defined by

$$\hat{\beta} = \arg\max_{\beta} \{L(\beta)\} \tag{5.5}$$

where $L(\beta)$ is the likelihood function for β and defined below

$$L(\beta) = \Pi_{i=1}^{n} f_{Y_i}(y_i)$$

A necessary condition for $\hat{\beta}$ in (5.5) is

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0$$

$$\sum_{i=1}^{n} \left(\frac{y_i - x_i \hat{\beta}}{\sigma} \right) x_{ij} = 0 \ \forall j = 1 \ldots p \tag{5.6}$$

This leads to the following result in matrix form

$$X^t y - (X^t X)\hat{\beta} = 0$$

where "t" indicates the transpose of a matrix. Therefore

$$\hat{\beta} = (X^t X)^{-1} X^t y \tag{5.7}$$

In the case of Example 5.1, $\hat{\beta}_0 = -260.059, \hat{\beta}_1 = 5.041$, and the LS fitted model is $\hat{y} = -260.059 + 5.041\, x$. Figure 5.1 shows that the fitted straight line does not fit the bulk of the data.

Note that if α_i is a weight associated with the observations (y_i, x_i) for all $i = 1 \ldots n$, then the parameter estimate β is defined by (5.8) to obtain a weighted version of the estimator

$$\hat{\beta} = \arg\max_{\beta} \{\Pi_{i=1}^{n} [f_{Y_i}(y_i)]^{\alpha_i}\} \tag{5.8}$$

This leads to

$$X^t \Lambda y - (X^t \Lambda X)\beta = 0 \tag{5.9}$$

where Λ is an $(n \times n)$ diagonal matrix, whose elements are $\Lambda_i = \alpha_i^{0.5}$ $i = 1 \ldots n$. Hence

$$\hat{\beta} = (X^t \Lambda X)^{-1} X^t \Lambda y$$

5.3 Influence Function for $\hat{\beta}$

As there is more than one parameter, (4.10) needs to be modified to (5.10) ([35], page 71)

$$IF_{\hat{\beta}}((x_0, y_0), F) = -B^{-1}\Psi((x_0, y_0), \hat{\beta}) \tag{5.10}$$

where $\hat{\beta}$ is a $(p \times 1)$ vector, x_0 is a $(1 \times p)$ vector, $\Psi = (\psi_1, \ldots, \psi_p)^t$ is the first derivative of log likelihood with respect to the parameter β, and the matrix B has elements

$$b_{jk} = E\left\{ \left[\frac{\partial \psi_j}{\partial \beta_k} \right]_{\beta=\hat{\beta}} \right\}$$

where $\Psi = (\psi_1, \ldots, \psi_p)^t$ and $\psi_{(i+1)} = \frac{\partial}{\partial \beta_i} \frac{1}{2} \left(\frac{y - \beta_0 - \beta_1 x_1 - \cdots - \beta_{(p-1)} x_{(p-1)}}{\sigma} \right)^2$; $i = 0, \ldots, (p-1)$. Hinkley [23] derived the influence function for ordinary least squares, which is also given by Cook and Weisberg ([10], section 3.3)

$$IF_{\hat{\beta}}((x_0, y_0), F) = [E_F(X^t X)]^{-1} x_0^t (y_0 - x_0 \hat{\beta}) \tag{5.11}$$

In the case of Example 5.1, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^t$ and

$$\Psi = \begin{pmatrix} \psi_1 \\ \\ \psi_2 \end{pmatrix} = \begin{pmatrix} -\frac{y - \beta_0 - \beta_1 x}{\sigma^2} \\ \\ -(\frac{y - \beta_0 - \beta_1 x}{\sigma^2}) x \end{pmatrix} = - \begin{pmatrix} 1 \\ \\ x \end{pmatrix} (\frac{y - \beta_0 - \beta_1 x}{\sigma^2})$$

and

$$B = \frac{1}{n\sigma^2} \begin{pmatrix} n & \sum x_i \\ \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \frac{1}{n\sigma^2} (X^t X)$$

Therefore, the influence function for $\hat{\beta}$ using (5.10) is

$$IF_{\hat{\beta}}((x_0, y_0), F_n) = n \left(X^t X\right)^{-1} x_0^t \left(y_0 - x_0 \hat{\beta}\right)$$

Note that this result is similar to (5.11). The influence function for the LS estimate (MLE) tends to infinity for any fixed x_0 if y_0 tends to infinity. Hence, we can say that these estimates are heavily affected by extreme values.

5.4 Mixture Model

In the following sections, we seek to develop a procedure that gives a good fit to the bulk of the data without being perturbed by a small proportion of outliers, and that does not require us to decide which observations are outliers.

Our strategy here is similar to that in section 4.4. Formally, we consider the mixture model

$$p(y, X, \beta) = \lambda f(y) + (1 - \lambda)g(y) \tag{5.12}$$

where g is a dispersed parameter free function over the sample space and $1-\lambda$ is a fixed small positive number which may be thought of as the proportion of contaminated data. Often we choose λ to be 0.95 or similar. Remember f is defined in (5.4).

For fixed choices of g and λ we will consider the robustness properties of

$$\tilde{\beta} = \arg\max_{\beta} L_o(\beta) \tag{5.13}$$

where $L_o(\beta)$ is the observed likelihood function for β.

$$L_o(\beta) = \Pi_{i=1}^{n} \left[\lambda f(y_i) + (1 - \lambda)g(y_i)\right]$$

This will be referred to as the robustified estimator.

5.5 Calculating the Robustified Estimator

It is hard to maximize the observed likelihood function, $L_o(\beta)$, so that the complete likelihood function, $L_c(\beta)$, is needed to estimate β. That is,

$$\tilde{\beta} = \arg\max_{\beta} L_c(\beta) \tag{5.14}$$

where

$$L_c(\beta) = \Pi_{i=1}^{n} \left[[\lambda f(y_i)]^{z_i} \left[(1 - \lambda)g(y_i)\right]^{1-z_i}\right]$$

$$z_i = \begin{cases} 1 & \text{if } y_i \in f \\ 0 & \text{if } y_i \in g \end{cases}$$

Since z_i's are unknown, the z_i's are treated as missing observations. They can be estimated using the E-step of the EM algorithm; followed by an estimation of β at the M-step.

5.5.1 E-Step

Following section 4.5.1, the z_i's are estimated by (5.15)

$$\tilde{z}_i = E[z_i|y_i, x_i, \beta] = \frac{\lambda f(y_i)}{\lambda f(y_i) + (1-\lambda)g(y_i)} \tag{5.15}$$

Sometimes we use the notation $\tilde{z}_i = z(y_i, x_i, \beta) = z(y_i - x_i\beta)$.

5.5.2 M-Step

The M-step is to maximize the complete likelihood function $L_c(\beta)$, where z_i is replaced by $\tilde{z}_i$ $i = 1, \ldots, n$

$$\begin{aligned} l_c &= \log L_c(\beta) \\ &= \sum_{i=1}^{n} \tilde{z}_i \log f(y_i) + \text{constant} \\ &= -\frac{1}{2} \sum_{i=1}^{n} \tilde{z}_i \left(\frac{y_i - x_i\beta}{\sigma} \right)^2 + \text{constant} \end{aligned}$$

A necessary condition for $\tilde{\beta}$ is

$$\sum_{i=1}^{n} \tilde{z}_i \left(\frac{y_i - x_i\beta}{\sigma} \right) x_{ij} = 0 \ \forall j \tag{5.16}$$

In matrix form, we can write

$$X^t Z y - X^t Z X \beta = 0 \tag{5.17}$$

where Z is a diagonal matrix whose elements are $\tilde{z}_i$ $i = 1 \ldots n$. It is similar to (5.9), but with the matrix Λ replaced by Z.

It is an iterative algorithm. For a given starting value for β, say $\tilde{\beta}^{(0)}$, we can compute the new estimate $\tilde{\beta}^{(1)}$ using the fixed point equation (5.17).

$$X^t Z^{(0)} y - (X^t Z^{(0)} X)\tilde{\beta}^{(1)} = 0$$

The elements of the matrix $Z^{(0)}$ are $\tilde{z}_i^{(0)} = z(y_i, x_i, \tilde{\beta}^{(0)})$, $i = 1 \ldots n$. In general, we can write

$$X^t Z^{(m)} y - (X^t Z^{(m)} X)\tilde{\beta}^{(m+1)} = 0 \quad m = 0, 1, \ldots$$

The limit of the sequence $\{\tilde{\beta}^{(m)}\}_{m=1}^{\infty}$, if it converges, is the estimate for β. That is,

$$\tilde{\beta} = \lim_{m \to \infty} \tilde{\beta}^{(m)}$$

5.6 Influence Function for $\tilde{\beta}$

Since the estimate $\tilde{\beta}$ is obtained using iteration, we use the Jorgensen's [30] method to derive the influence function for $\tilde{\beta}$. Let $h(\beta, F_n)$ be the updating function defined by

$$h(\beta, F_n) = (X^t Z X)^{-1} X^t Z y \tag{5.18}$$

where F_n is the empirical distribution which places mass $\frac{1}{n}$ at the n points $\{(x_1, y_1), \ldots (x_n, y_n)\}$ in $\mathcal{R}^{(p+1)}$. Note that h is a $(p \times 1)$ column vector. The one step statistical functional $\breve{\beta}$ is defined in (5.19)

$$\breve{\beta}(F_n) = (X^t Z X)^{-1} X^t Z y \tag{5.19}$$

where the matrix Z is considered as fixed at $Z(\tilde{\beta})$. The link between $\breve{\beta}$ and $\tilde{\beta}$ is given in (5.20).

$$\breve{\beta}(F_n) = h(\tilde{\beta}, F_n) = (X^t Z X)^{-1} X^t Z y \tag{5.20}$$

5.6.1 One-Step Influence Function for $\tilde{\beta}$

Jorgensen derived the influence function for weighted ordinary least squares, which is presented in [30]. This result was derived from the influence function for ordinary least squares done by Hinkley [23].

The empirical influence function for $\breve{\beta}$, $IF_{\breve{\beta}}((x_0, y_0), F_n)$, is derived using the definition of influence function and given in (5.21). The derivation for this result is given in Appendix 5.10.

$$IF_{\breve{\beta}}((x_0, y_0), F_n) = n\tilde{z}_0 (X^t Z X)^{-1} x_0^t (y_0 - x_0 \tilde{\beta}) \tag{5.21}$$

We call $IF_{\breve{\beta}}((x_0, y_0), F_n)$ the one-step influence function for $\tilde{\beta}$ denoted as $IF^1_{\tilde{\beta}}((x_0, y_0), F_n)$. If $\tilde{z}_0 = 1$, $\tilde{\beta} = \hat{\beta}$ and Z is an $n \times n$ identity matrix, the result in (5.21) is exactly similar to the influence function of maximum likelihood estimate $\hat{\beta}$, described in section 5.3.

5.6.2 Jacobian Matrix

We derive the Jacobian matrix J, which is the derivative of the updating function h. From (5.18), we can write

$$(X^t Z X) h(\beta, F_n) = X^t Z y$$

Differentiate both sides with respect to β.

$$\begin{aligned}
(X^t \frac{\partial Z}{\partial \beta} X) h(\beta, F_n) + (X^t Z X) \frac{\partial h}{\partial \beta} &= X^t \frac{\partial Z}{\partial \beta} y \\
(X^t Z X) \frac{\partial h}{\partial \beta} &= X^t \frac{\partial Z}{\partial \beta} (y - X h(\beta, F_n)) \\
J = \left[\frac{\partial h}{\partial \beta} \right]_{\beta = \tilde{\beta}} &= (X^t Z X)^{-1} X^t \frac{\partial Z}{\partial \beta} (y - X\tilde{\beta})
\end{aligned}$$

Let $r_i = y_i - x_i \tilde{\beta}$ and consider

$$\frac{\partial Z}{\partial \beta_j}(y - X\tilde{\beta}) = \begin{pmatrix} r_1 \frac{\partial z_1}{\partial \beta_j} \\ . \\ . \\ . \\ r_n \frac{\partial z_n}{\partial \beta_j} \end{pmatrix} = \begin{pmatrix} r_1 x_{1j} z_1' \\ . \\ . \\ . \\ r_n x_{nj} z_n' \end{pmatrix} = \begin{pmatrix} r_1 z_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & r_n z_n' \end{pmatrix} \begin{pmatrix} x_{1j} \\ . \\ . \\ . \\ x_{nj} \end{pmatrix}$$

Therefore

$$\begin{aligned}
\frac{\partial Z}{\partial \beta}(y - X\tilde{\beta}) &= \left(\frac{\partial Z}{\partial \beta_1}(y - X\tilde{\beta}), \ldots, \frac{\partial Z}{\partial \beta_p}(y - X\tilde{\beta}) \right) \\
&= \begin{pmatrix} r_1 z_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & r_n z_n' \end{pmatrix} \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix} \\
&= VX
\end{aligned}$$

where V is $(n \times n)$ diagonal matrix whose elements are $v_{ii} = r_i z_i'$. Hence

$$J = (X^t Z X)^{-1} X^t V X \tag{5.22}$$

$$
\begin{aligned}
z_i &= \frac{\lambda f(y_i, x_i, \beta)}{\lambda f(y_i, x_i, \beta) + (1-\lambda) g(y_i)} \\
z_i' = \frac{\partial z_i}{\partial \beta} &= \frac{(1-\lambda)\lambda g(y_i) \frac{\partial f(y_i, x_i, \beta)}{\partial \beta}}{\lambda f(y_i, x_i, \beta) + (1-\lambda) g(y_i)} \\
&= \frac{(1-z_i)\lambda \frac{\partial f(y_i, x_i, \beta)}{\partial \beta}}{\lambda f(y_i, x_i, \beta) + (1-\lambda) g(y_i)}
\end{aligned}
$$

$$
\begin{aligned}
f(y_i, x_i, \beta) &= K \exp\left\{ -\frac{1}{2} \left(\frac{y_i - x_i\beta}{\sigma} \right)^2 \right\} \\
\log f(y_i, x_i, \beta) &= \text{constant} - \frac{1}{2} \left(\frac{y_i - x_i\beta}{\sigma} \right)^2 \\
\frac{\partial f(y_i, x_i, \beta)}{\partial \beta_j} &= \frac{1}{\sigma^2} f(y_i, x_i, \beta) \left(y_i - x_i\beta \right) \; \forall j = 1 \ldots p.
\end{aligned}
$$

5.6.3 True Influence Function for $\tilde{\beta}$

Let I be the $p \times p$ identity matrix.

$$
\begin{aligned}
I - J &= I - (X^t Z X)^{-1} (X^t V X) \\
&= (X^t Z X)^{-1} [(X^t Z X) - (X^t V X)] \\
&= (X^t Z X)^{-1} [X^t (Z - V) X] \\
(I - J)^{-1} &= [X^t (Z - V) X]^{-1} (X^t Z X)
\end{aligned}
$$

The true influence function for $\tilde{\beta}$ is denoted by $IF_{\tilde{\beta}}((x_0, y_0), F_n)$

$$\begin{aligned} IF_{\tilde{\beta}}((x_0, y_0), F_n) &= (I - J)^{-1} IF^1_{\tilde{\beta}}((x_0, y_0), F_n) \\ &= n\tilde{z}_0[X^t(Z - V)X]^{-1}x_0^t(y_0 - x_0\tilde{\beta}) \\ &= n\tilde{z}_0[X^tUX]^{-1}x_0^t(y_0 - x_0\tilde{\beta}) \end{aligned}$$

where $U = Z - V$ is a diagonal matrix whose elements are $u_i = z_i - r_i z_i'$.

5.7 Comparison of Methods Based on the Influence Function of β

The influence function for $\hat{\beta}$ obtained by the maximum likelihood method is discussed in section 5.3 and given below

$$IF_{\hat{\beta}}((x_0, y_0), F_n) = n\,(X^tX)^{-1}\,x_0^t\,(y_0 - x_0\hat{\beta})$$

Obviously, the maximum likelihood estimator $\hat{\beta}$ is not robust against both outliers and high leverage points.

The M-estimator $\hat{\beta}_M$ is defined in (1.22) and given below.

$$\hat{\beta}_M = (X^tWX)^{-1}X^tWy$$

where W is a diagonal matrix, whose elements are $w(r) = \frac{\psi(r)}{r}$, and $\psi(r)$ is redescending function. The influence function for the M-estimator $\hat{\beta}_M$ can be obtained using the result in section 5.6.3. In order to get this, we have to replace $\hat{z}_0$ by w_0, $\tilde{\beta}$ by $\hat{\beta}_M$, and $U = W - V$ is a diagonal matrix whose elements are $u_i = w_i - v_i = w_i - (y_i - \hat{\beta}_M x_i)w_i'$, where $w' = \frac{\partial w}{\partial \beta}$. Therefore, the influence function for $\hat{\beta}_M$ is given below

$$\begin{aligned} IF_{\hat{\beta}_M}((x_0, y_0), F_n) &= n\,w_0\,(X^tUX)^{-1}\,x_0^t\,(y_0 - x_0\hat{\beta}_M) \\ &= n\,\sigma\,\psi(r_0)(X^tUX)^{-1}x_0^t \end{aligned}$$

where $r_0 = \frac{y_0 - x_0\hat{\beta}_M}{\sigma}$. Since $\psi(r)$ is redescending function, $\psi(r)$ tends to 0 if either $x_0 \to \pm\infty$ or $y_0 \to \pm\infty$ or both $x_0, y_0 \to \pm\infty$ so that $IF_{\hat{\beta}_M}((x_0, y_0), F_n)$ diverges only when $x_0 \to \pm\infty$. That means, M-estimator is robust against outliers but not robust against high leverage points. It gives better a estimate than MLE.

The influence function for the mixture estimate $\tilde{\beta}$ is given in section 5.6.3, where

$$\begin{aligned}
\tilde{z}_0 &= \frac{K\exp(-\frac{1}{2\sigma^2}(y_0 - x_0\tilde{\beta})^2)}{K\exp(-\frac{1}{2\sigma^2}(y_0 - x_0\tilde{\beta})^2) + \text{constant}} \\
&= \frac{\exp(-\frac{1}{2\sigma^2}\eta_0^2)}{\exp(-\frac{1}{2\sigma^2}\eta_0^2) + K^*} \\
&= \frac{1}{1 + K^*\exp(\frac{1}{2\sigma^2}\eta_0^2)}
\end{aligned}$$

where $\eta_0 = y_0 - x_0\tilde{\beta}_0$.

If $\eta_0 = 0$, then $\tilde{z}_0$ is positive constant and the influence function $IF_{\tilde{\beta}}((x_0, y_0), F_n)$ tends to 0 as $x_0 \to \infty$.

Consider the case of $\eta_0 \neq 0$. If $x_0 \to \pm\infty$, $y_0 \to \pm\infty$, then $\eta_0^2 \to \infty$, so that $\tilde{z}_0$ quickly converges to 0. Hence $IF_{\tilde{\beta}}((x_0, y_0), F_n)$ converges to 0 if $x_0 \to \pm\infty$, $y_0 \to \pm\infty$. That means, the mixture estimator is robust against outliers as well as high leverage points.

5.8 Numerical Results

In this section, we illustrate our method numerically using the Example 5.1. In fact we do not know the value of scale parameter for this example. According to standard practice, the scale parameter is estimated before starting the estimation of the regression parameters. We take the scale parameter as a MAD of the response variable. That is, $\sigma = MAD(y) = 15.1225$. Another option for σ is standardized MAD, known as MADN.

For the mixture model, the parameter free density function g needs to be defined. We know the total number of calls will not be negative and the

Table 5.1: Regression estimates for the Belgium phone call data

	MLE	$\boldsymbol{\lambda = 0.75}$	$\boldsymbol{\lambda = 0.90}$	$\boldsymbol{\lambda = 0.95}$
Estimate for $\boldsymbol{\beta_0}$	- 260.059	- 63.256	- 63.404	- 63.444
Estimate for $\boldsymbol{\beta_1}$	5.041	1.300	1.303	1.304

largest response value is 212, so we have chosen g as a very dispersed uniform distribution with parameters $[a = 0, b = 300]$. It is an iterative process which requires starting values for the parameters β. In this problem, we have chosen least squares estimates as starting values.

We compute the estimates for β for the various λ values such as true value $0.75(18/24)$, 0.90 and 0.95. The results are given in Table 5.1. The estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are almost equal for the various λ. This means λ has little impact on the estimates. However, our estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are very different from the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. When $\lambda = 0.95$, the fitted line is $\tilde{y} = -63.444 + 1.304x$ (plotted as a solid line in Figure 5.2), which avoids the outliers and clearly this line fits the majority of the data.

Since this is a two-dimensional parameter problem, the Jacobian becomes a 2×2 matrix. The estimated Jacobian matrix is

$$J = \begin{bmatrix} -0.00726 & -0.51432 \\ \\ 0.00013 & 0.00945 \end{bmatrix}$$

The influence functions for $\tilde{\beta}_0 = -63.670$ and $\tilde{\beta}_1 = -1.308$ were computed to determine whether $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are affected by a large observation in either x or y or both. The true and one-step influence functions for the $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are given in Figure 5.3 and in Figure 5.4 respectively, where we can see that the observations far away from the bulk of the data do not have any impact on the estimates. Figure 5.3 and Figure 5.4 show that the influence functions for $\tilde{\beta}_0$ and $\tilde{\beta}_0$ are bounded at different levels, which depend on the x values.

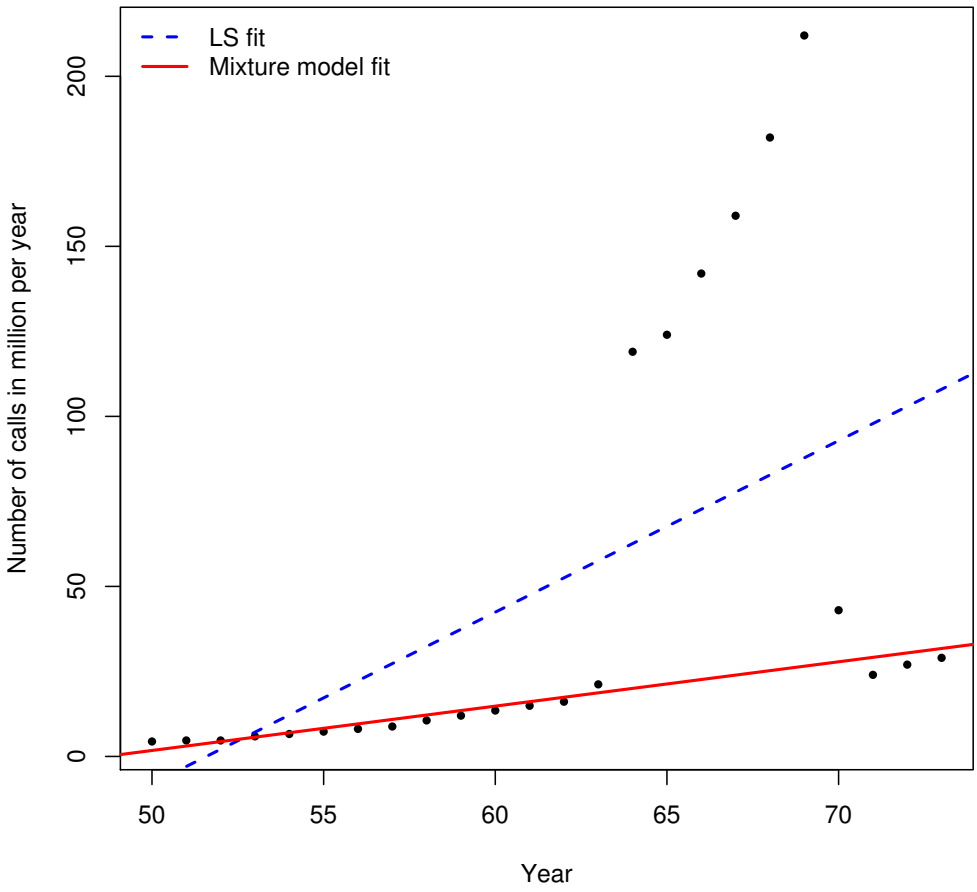

Figure 5.2: Number of international phone calls from Belgium in the years $1950 - 1973$ with LS fit and mixture model fit

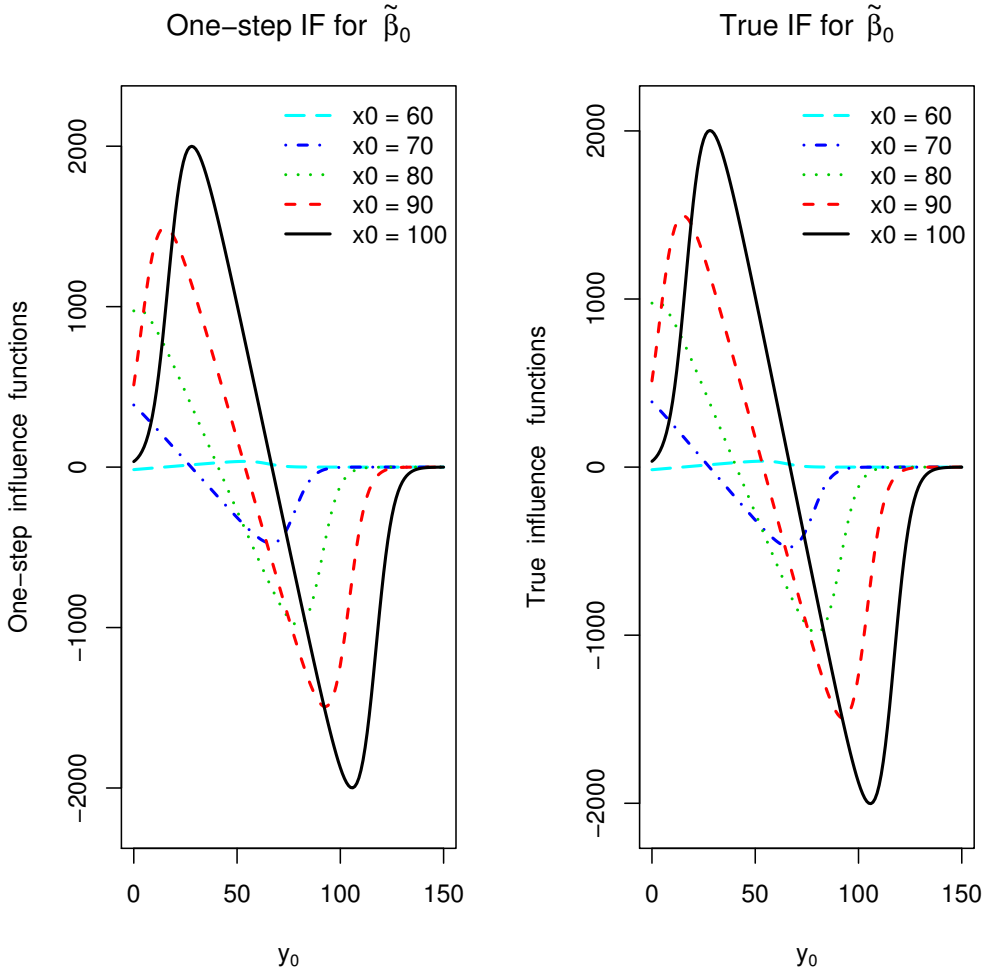

Figure 5.3: True and one-step influence functions for $\tilde{\beta}_0$

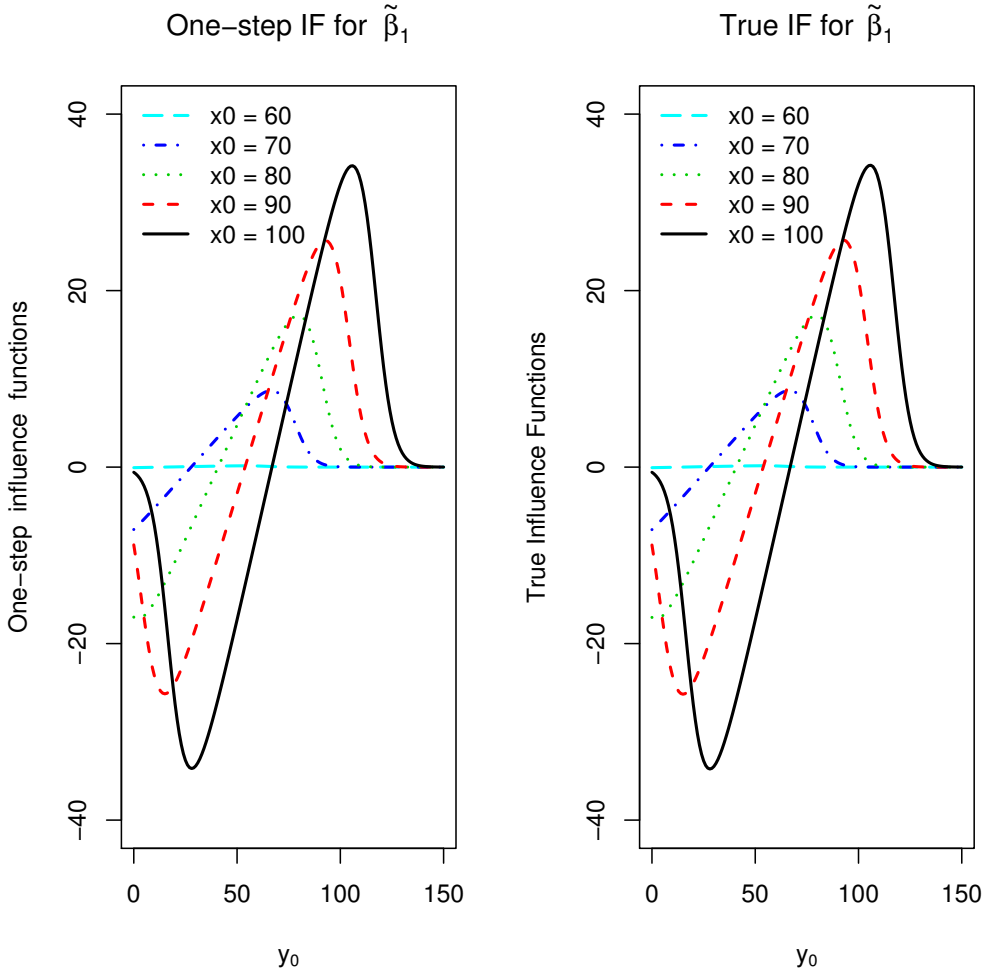

Figure 5.4: True and one-step influence functions for $\tilde{\beta}_1$

Table 5.2: Regression estimates given by various methods for the Belgium phone call data

	Estimates for $\boldsymbol{\beta_0}$	**Estimates for $\boldsymbol{\beta_1}$**
MLE	- 260.059	5.041
Mixture	- 63.444	1.304
Tukey	- 62.677	1.289
Huber	- 113.715	2.263

5.8.1 Comparison of $\tilde{\beta}$ with Standard Robust Estimates

Next, the estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are compared with traditional estimates obtained by Huber and Tukey methods (refer sections 1.4 and 1.4.3). The standard turning constants such as $c = 1.345$ for Huber, and $c = 4.685$ for Tukey, and scale parameter $\sigma = 15.1225$ were used to estimate for β_0 and β_1. Results are given in the Table 5.2. Results based on our method and the Tukey method are very similar, but Huber estimates differ from the estimates $\tilde{\beta}$ and Tukey estimates.

Figure 5.5 shows the LS, mixture, Tukey and Huber fits computed by the relevant methods. The fitted lines made by the mixture method and by the Tukey method overlap. In addition, the Huber line does not fit the majority of the data. These results may be improved if we compute the scale parameter simultaneously with estimating the regression parameters. This will be discussed in Chapter 7.

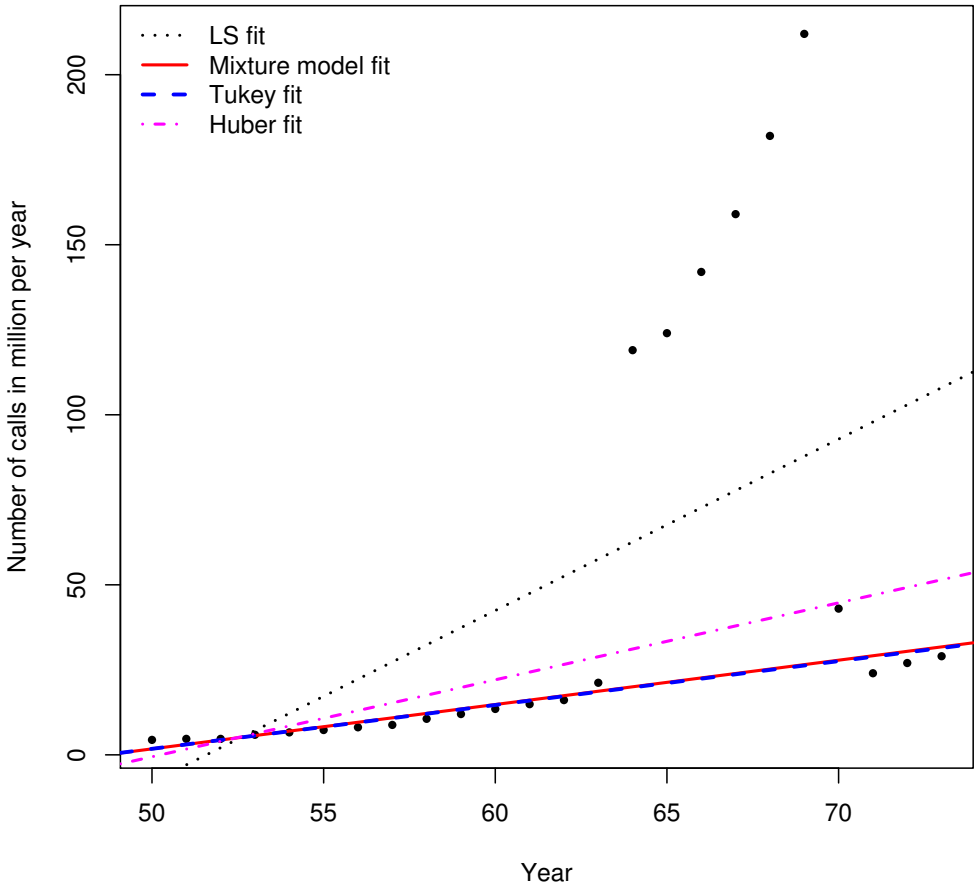

Figure 5.5: Number of international phone calls from Belgium in the years $1950 - 1973$ with fit made by the various methods

Table 5.3: Regression estimates for phone data by various methods and their variance in brackets

	MLE	**Mixture**	**Huber**	**Tukey**
$\boldsymbol{\beta_0}$	-252.229 (228.1466)	-57.82234 (452.892)	-130.3087 (780.6379)	-60.72171 (591.4358)
$\boldsymbol{\beta_1}$	4.925469 (0.05947817)	1.23211 (0.1416882)	2.613586 (0.2527835)	1.289169 (0.1923757)

5.9 Efficiency

This investigation is very similar to section 4.8, but the variance of estimates for β_0 and β_1 will be discussed here. Again this is entirely a simulation study.

Initially, a set of 20 "good" observations are selected from the phone data. This data set is used to obtain maximum likelihood estimates for β_0 and β_1, which gives $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 * \text{year}$. Then a set of 20 observations are randomly generated from $\mathcal{N}(\hat{\mu}, s)$, where s is the standard deviation of calls in the good 20 observations of the phone data. The full data set is a combined data set of generated data and a set of 6 "bad" observations in the phone data. Now, we are able to compute β_0 and β_1 by various methods such as maximum likelihood estimation, mixture method, and M-estimation of Huber and Tukey. This procedure is repeated 1000 times to obtain the variance of these estimates.

Table 5.3 gives estimates of β_0 and β_1 and their variance. The variance of the mixture estimate is smaller than M-estimates of Huber and Tukey. However, the mixture estimates are similar to the Tukey estimates.

5.10 Appendix: The Empirical Influence Function for One Step Estimator $\breve{\beta}$

Consider the one step statistical functional $\breve{\beta} = (X^t Z X)^{-1} X^t Z y$, and perturbed data set $(x_1, y_1), \ldots (x_n, y_n), (x_0, y_0)$ with weights $\frac{1-\epsilon}{n}, \ldots, \frac{1-\epsilon}{n}, \epsilon$. From the result (5.19), we can define

$$\breve{\beta}((1-\epsilon)F_n + \epsilon\Delta_{(x_0,y_0)}) = (X^t_{new} Z_{new} X_{new})^{-1} X^t_{new} Z_{new} y_{new}$$

where

$$y_{new} = \begin{bmatrix} y_1 \\ . \\ \\ . \\ y_n \\ y_0 \end{bmatrix} = \begin{bmatrix} y \\ y_0 \end{bmatrix}_{(n+1)\times 1} \quad X_{new} = \begin{bmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \\ x_{01} & . & . & x_{0p} \end{bmatrix} = \begin{bmatrix} X \\ x_0 \end{bmatrix}_{(n+1)\times p}$$

$$\begin{aligned} Z_{new} &= \begin{bmatrix} \frac{1-\epsilon}{n}\hat{z}_1 & 0 & . & . & . & 0 \\ 0 & \frac{1-\epsilon}{n}\hat{z}_2 & . & . & . & 0 \\ . & . & . & . & . & . \\ \\ 0 & 0 & . & . & \frac{1-\epsilon}{n}\hat{z}_n & 0 \\ 0 & 0 & . & . & 0 & \epsilon\hat{z}_0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1-\epsilon}{n} Z & \underline{0} \\ \\ \underline{0}^t & \epsilon\hat{z}_0 \end{bmatrix} = \begin{bmatrix} Z^* & \underline{0} \\ \underline{0}^t & \epsilon\hat{z}_0 \end{bmatrix} \end{aligned}$$

where $\underline{0}^t = (0, \ldots, 0)_{1\times n}$ and $Z^* = \frac{1-\epsilon}{n} Z$. We use the notation $\breve{\beta}_{new} = \breve{\beta}((1-\epsilon)F_n + \epsilon\Delta_y)$

$$\begin{aligned}
\breve{\beta}_{new} &= \left(\begin{bmatrix} X^t & x_0^t \end{bmatrix} \begin{bmatrix} Z^* & \underline{0} \\ \underline{0}^t & \epsilon\hat{z}_0 \end{bmatrix} \begin{bmatrix} X \\ x_0 \end{bmatrix} \right)^{-1} \\
&\quad \left(\begin{bmatrix} X^t & x_0^t \end{bmatrix} \begin{bmatrix} Z^* & \underline{0} \\ \underline{0}^t & \epsilon\hat{z}_0 \end{bmatrix} \begin{bmatrix} y \\ y_0 \end{bmatrix} \right) \\
&= [X^t Z^* X + x_0^t(\epsilon\hat{z}_0)x_0]^{-1}[X^t Z^* y + x_0^t(\epsilon\hat{z}_0)y_0] \\
&= \left[(X^t Z^* X)^{-1} - \frac{(X^t Z^* X)^{-1} x_0^t (\epsilon\hat{z}_0 x_0)(X^t Z^* X)^{-1}}{1 + (\epsilon\hat{z}_0 x_0)(X^t Z^* X)^{-1} x_0^t} \right] [X^t Z^* y + x_0^t(\epsilon\hat{z}_0)y_0] \\
&= \tilde{\beta} + \frac{\epsilon\hat{z}_0 y_0 (X^t Z^* X)^{-1} x_0^t - \epsilon\hat{z}_0 (X^t Z^* X)^{-1} x_0^t x_0 \tilde{\beta}}{1 + \epsilon\hat{z}_0 x_0 (X^t Z^* X)^{-1} x_0^t}
\end{aligned}$$

$$\begin{aligned}
\frac{\breve{\beta}_{new} - \breve{\beta}(F_n)}{\epsilon} &= \frac{\hat{z}_0 (X^t Z^* X)^{-1} x_0^t (y_0 - x_0\tilde{\beta})}{1 + \epsilon\hat{z}_0 x_0 (X^t Z^* X)^{-1} x_0^t} \\
&= \frac{\frac{n}{1-\epsilon}\hat{z}_0 (X^t Z X)^{-1} x_0^t (y_0 - x_0\tilde{\beta})}{1 + \frac{n\epsilon}{1-\epsilon}\hat{z}_0 x_0 (X^t Z^* X)^{-1} x_0^t} \\
\lim_{\epsilon\to 0} \frac{\breve{\beta}_{new} - \breve{\beta}(F_n)}{\epsilon} &= n\hat{z}_0 (X^t Z X)^{-1} x_0^t (y_0 - x_0\tilde{\beta})
\end{aligned}$$

That is

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{\breve{\beta}((1-\epsilon)F_n + \epsilon\Delta_y) - \breve{\beta}(F_n)}{\epsilon} &= n\hat{z}_0 (X^t Z X)^{-1} x_0^t (y_0 - x_0\tilde{\beta}) \\
IF_{\breve{\beta}}((x_0, y_0)F_n) &= n\hat{z}_0 (X^t Z X)^{-1} x_0^t (y_0 - x_0\tilde{\beta})
\end{aligned}$$

Chapter 6

Robust Estimation for Generalized Linear Models

Generalized linear modeling (GLM) is a framework for statistical analysis that is able to tackle a wide range of data with different types of response variables. Nowadays, it is commonly applied in the fields of science, medicine, business, etc. The non-robustness of the MLE for β has been investigated widely in the statistical modelling literature [42].

Methods for the robust estimation of GLMs have developed much more slowly than robust methods for linear models. At present we have some methods for the robust estimation of GLM, described in [1], [8], [35], and those methods are usually limited to the binomial model with logit link and to the Poisson model. Our robust approach, which is an alternative to the classical approach, may be useful for the whole class of generalized linear models.

This chapter starts with a brief description of the estimation procedure of GLM. It then shows how to compute the influence functions of the estimate. It ends with an introduction of how to obtain the robustified estimates for the GLM parameters and their influence functions. In addition, we provide empirical examples for binomial and Poisson regression models.

6.1 Introduction to GLM

Nelder and Wedderburn [42] introduced the class of generalized linear models. In a large measure, the success of this class of statistical models is due to the ability of IRLS algorithms to reliably fit models of this kind.

GLMs are defined as follows

1. A response variable y, continuous or discrete, with a distribution from exponential family, defined in (6.1).

$$f(y, \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]; \tag{6.1}$$

2. A set of independent variables $X_1, \ldots, X_n$; and

3. Let k be a monotone differentiable function, defined in (6.2)

$$k(\mu) = X\beta \tag{6.2}$$

where $\mu = E[Y]$, X is a $(n \times p)$ design matrix, β is a $(p \times 1)$ vector being the model parameter of interest, and k is known as a link function. Indeed, we consider more general situations in which the regressors affect the distribution of y, which is assumed to depend on the linear combination $X\beta$ only. In the GLM literature, $\eta = X\beta$ is called linear predictor. For further reference, see [15].

6.2 Model

Consider the exponential family defined in (6.1). The term $b(\theta)$ is considered to be the natural parameter, and if $a(y) = y$, f said to be in the canonical form. In this chapter, we assume that f is in the canonical form of the exponential family. The natural parameters for the binomial and the Poisson distributions are given in Table 6.1.

The mean and variance for the exponential family are defined below

$$\mu = E(y) = -\frac{c'(\theta)}{b'(\theta)}$$

Table 6.1: Example of binomial and Poisson distributions to explain the exponential family

	$\boldsymbol{a(y)}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{c(\theta)}$	$\boldsymbol{d(y)}$
Binomial	y	$\log(\frac{\theta}{1-\theta})$	$n\log(1-\theta)$	$\log\begin{pmatrix} n \\ y \end{pmatrix}$
Poisson	y	$\log\theta$	$-\theta$	$-\log y!$

$$var(y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

where $b'(\theta) = \frac{\partial b(\theta)}{\partial\theta}$, $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial\theta^2}$ $c'(\theta) = \frac{\partial c(\theta)}{\partial\theta}$ and $c''(\theta) = \frac{\partial^2 c(\theta)}{\partial\theta^2}$. For example

1. If f is binomial probability density

$$\mu = -\frac{n\frac{-1}{1-\theta}}{\frac{1}{\theta} - \frac{-1}{1-\theta}} = n\theta$$

$$var(y) = \frac{(\frac{-1}{\theta^2} + \frac{1}{(1-\theta)^2})(\frac{-n}{1-\theta}) + (\frac{n}{(1-\theta)^2})(\frac{1}{\theta} + \frac{1}{1-\theta})}{(\frac{1}{\theta} + \frac{1}{1-\theta})^3} = n\theta(1-\theta)$$

2. If f is Poisson probability density

$$\mu = -\frac{-1}{\frac{1}{\theta}} = \theta$$

$$var(y) = \frac{-\frac{-1}{\theta^2} - 0}{(\frac{1}{\theta})^3} = \theta$$

6.3 Calculating the Estimator

Consider random variables $Y_1, \ldots, Y_n$ satisfying the properties of a generalized linear model. We wish to estimate the parameters β, which are related to the Y_i's through $E[Y_i] = \mu_i$ and $k(\mu_i) = x_i\beta$. where x_i is the i^{th} row vector of the design matrix X. The parameter estimate β is defined by

$$\hat{\beta} = \arg\max_{\beta} \{L(\theta)\} \tag{6.3}$$

where θ depends on β,and the likelihood function, $L(\theta)$, is defined below.

$$L(\theta) = \Pi_{i=1}^{n} f_{Y_i}(y_i)$$

For simplicity, $\log L(\theta)$ is often considered instead of $L(\theta)$

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{n} \log f_Y(y_i) \\ &= \sum_{i=1}^{n} (y_i b(\theta_i) + c(\theta_i) + d(y_i)) \\ &= \sum_{i=1}^{n} l_i(\theta) \end{aligned}$$

where $l_i(\theta) = y_i b(\theta_i) + c(\theta_i) + d(y_i)$. A necessary condition for $\beta = \hat{\beta}$ in (6.3) is

$$U(\hat{\beta}) = \left[\frac{\partial l(\theta)}{\partial \beta}\right]_{\beta=\hat{\beta}} = 0$$

where

$$U(\beta) = \begin{pmatrix} U_1(\beta) \\ . \\ \\ . \\ U_p(\beta) \end{pmatrix} = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \beta_1} \\ . \\ \\ . \\ \frac{\partial l(\theta)}{\partial \beta_p} \end{pmatrix}$$

$$
\begin{aligned}
U_j(\beta) &= \frac{\partial l(\theta)}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\theta)}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\theta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} [y_i b'(\theta_i) + c'(\theta_i)] \left(\frac{1}{b'(\theta_i) var(y_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{var(y_i)} \right) x_{ij} \frac{\partial \mu_i}{\partial \eta_i}
\end{aligned}
$$

$$
U(\beta) = X^t W \Gamma
$$

where Γ is a column vector whose elements are $\gamma_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$ and W is a diagonal matrix whose elements are w_i.

$$
w_i = \frac{[\frac{\partial \mu_i}{\partial \eta_i}]^2}{var(y_i)}
$$

Since w_i is a function of y_i and η_i, it may be written as $w_i = w(y_i, \eta_i)$. The $U(\beta)$ is known as the score function.

The well known Fisher scoring iterative algorithm is used for finding the roots of the equation $U(\beta) = 0$. This gives an iterative procedure

$$
\beta^{(m)} = \beta^{(m-1)} + \left(\Sigma^{(m-1)}\right)^{-1} U^{(m-1)} \tag{6.4}
$$

where $\Sigma = Var(U) = E(UU^t)$. The $(p\,q)^{\text{th}}$ element of Σ is

$$\begin{aligned}
\Sigma_{pq} &= E(U_p U_q^t) \\
&= E\left[\left(\sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{var(y_i)}\right)x_{ip}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(\sum_{j=1}^{n}\left(\frac{y_j-\mu_j}{var(y_j)}\right)x_{jq}\frac{\partial\mu_j}{\partial\eta_j}\right)\right] \\
&= E\left[\sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{var(y_i)}\right)^2\left[\frac{\partial\mu_i}{\partial\eta_i}\right]^2 x_{ip}\,x_{iq}\right] \\
&= \sum_{i=1}^{n}\left(\frac{1}{var(y_i)}\right)\left[\frac{\partial\mu_i}{\partial\eta_i}\right]^2 x_{ip}\,x_{iq} \\
\Sigma &= X^tWX
\end{aligned}$$

From (6.4)

$$\Sigma^{(m-1)}\beta^{(m)} = \Sigma^{(m-1)}\beta^{(m-1)} + U^{(m-1)} \tag{6.5}$$

Consider the right hand side of (6.5)

$$\begin{aligned}
\Sigma\beta + U &= X^tWX\beta + X^tW\Gamma \\
&= X^tW(X\beta + \Gamma) \\
&= X^tWR
\end{aligned}$$

where R is a column vector whose elements are

$$r_i = x_i\beta + (y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i} = \eta_i + (y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i}$$

Since r_i is a function of y_i and η_i, it may be written as $r_i = r(y_i,\eta_i)$. The expression (6.5) becomes

$$X^tW^{(m-1)}X\hat{\beta}^{(m)} = X^tW^{(m-1)}R^{(m-1)} \tag{6.6}$$

The (6.6) takes the same form as weighted least squares, but it has to be applied iteratively because R and W depend on β. Hence, the estimate $\hat{\beta}$ can be obtained by IRLS with weights $w(y,\eta)$ and working response variable $r(y,\eta)$ provided we choose good initial estimates.

The limit of the sequence $\{\hat{\beta}^{(m)}\}_{m=1}^{\infty}$, if it converges, is the estimate for β. That is

$$\hat{\beta} = \lim_{m\to\infty}\hat{\beta}^{(m)}$$

6.4 Influence function for $\hat{\beta}$

The estimate $\hat{\beta}$ is computed here as iteratively re-weighted least squares, so the Jorgensen method is employed to compute the influence function of $\hat{\beta}$.

6.4.1 One-Step Influence function for $\hat{\beta}$

The one-step influence function of $\hat{\beta}$ can be directly derived from (5.21), where the matrix Z is replaced by the matrix W, and y_0 is replaced by r_0,

$$r_0 = x_0\hat{\beta} + (y_0 - \mu_0)\frac{\partial \eta_0}{\partial \mu_0}$$

Hence, (5.21) becomes

$$IF^1_{\hat{\beta}}((x_0, y_0), F_n) = nw_0(X^tWX)^{-1}x_0^t(r_0 - x_0\hat{\beta}) \tag{6.7}$$

$$IF^1_{\hat{\beta}}((x_0, y_0), F_n) = nw_0(X^tWX)^{-1}x_0^t(y_0 - \mu_0)\frac{\partial \eta_0}{\partial \mu_0}$$

where $\mu_0 = k^{-1}(\eta_0)$ and $\eta_0 = x_0\hat{\beta}$.

6.4.2 Jacobian Matrix

Next, we want to compute the Jacobian matrix J. The updating function $h(\beta, F_n)$ can be defined as follows

$$(X^tWX)h(\beta, F_n) = X^tWR \tag{6.8}$$

The computation of the Jacobian matrix is very similar to the section 5.6.2, but instead of y, we have R in the right hand side (6.8). It depends on β, so that computation of J is more complicated than previously.

Differentiate (6.8) both sides with respect to β.

$$X^t\frac{\partial W}{\partial \beta}Xh(\beta, F_n) + (X^tWX)\frac{\partial h(\beta, F_n)}{\partial \beta} = X^t\frac{\partial W}{\partial \beta}R + X^tW\frac{\partial R}{\partial \beta}$$

$$(X^tWX)\frac{\partial h(\beta, F_n)}{\partial \beta} = X^t\frac{\partial W}{\partial \beta}(R - Xh(\beta, F_n)) + X^tW\frac{\partial R}{\partial \beta}$$

$$J = \left[\frac{\partial h(\beta, F_n)}{\partial \beta}\right]_{\beta=\hat{\beta}} = (X^t W X)^{-1} X^t \left\{\frac{\partial W}{\partial \beta}(R - X\hat{\beta}) + W\frac{\partial R}{\partial \beta}\right\} \qquad (6.9)$$

Let $s_i = r_i - x_i\hat{\beta}$, and consider

$$\frac{\partial W}{\partial \beta_j}(R - X\hat{\beta}) = \begin{pmatrix} s_1\frac{\partial w_1}{\partial \beta_j} \\ . \\ . \\ . \\ s_n\frac{\partial w_n}{\partial \beta_j} \end{pmatrix} = \begin{pmatrix} s_1 x_{1j} w_1' \\ . \\ . \\ . \\ s_n x_{nj} w_n' \end{pmatrix}$$

$$= \begin{pmatrix} s_1 w_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & s_n w_n' \end{pmatrix} \begin{pmatrix} x_{1j} \\ . \\ . \\ . \\ x_{nj} \end{pmatrix}$$

$$\frac{\partial W}{\partial \beta}(R - X\hat{\beta}) = \left(\frac{\partial W}{\partial \beta_1}(R - X\hat{\beta}) \ldots \frac{\partial W}{\partial \beta_p}(R - X\hat{\beta})\right)$$

$$= \begin{pmatrix} s_1 w_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & s_n w_n' \end{pmatrix} \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix}$$

$$= \begin{pmatrix} s_1 w_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & s_n w_n' \end{pmatrix} X$$

$$\frac{\partial R}{\partial \beta_j} = \begin{pmatrix} \frac{\partial r_1}{\partial \beta_j} \\ . \\ . \\ . \\ \frac{\partial r_n}{\partial \beta_j} \end{pmatrix} = \begin{pmatrix} r_1' x_{1j} \\ . \\ . \\ . \\ r_n' x_{nj} \end{pmatrix}$$

$$\frac{\partial R}{\partial \beta} = \begin{pmatrix} r_1' x_{11} & . & . & r_1' x_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ r_n' x_{n1} & . & . & r_n' x_{np} \end{pmatrix} = \begin{pmatrix} r_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & . & . & r_n' \end{pmatrix} X$$

$$W\frac{\partial R}{\partial \beta} = \begin{pmatrix} w_1 r_1' & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & . & . & w_n r_n' \end{pmatrix} X$$

Therefore

$$\frac{\partial W}{\partial \beta}(R - X\hat{\beta}) + W\frac{\partial R}{\partial \beta} = \begin{pmatrix} s_1 w_i' + w_1 r_1' & 0 & . & 0 \\ . & & . & . & . \\ . & & . & . & . \\ . & & . & . & . \\ 0 & & . & . & s_n w_n' + w_n r_n' \end{pmatrix} X = VX$$

where V is a diagonal matrix, whose entries are v_i defined below

$$v_i = w_i'(r_i - x_i\hat{\beta}) + w_i r_i'$$

Hence, (6.9) can be written as

$$J = (X^t W X)^{-1} X^t V X \tag{6.10}$$

6.4.3 True Influence Function for $\hat{\beta}$

From section 5.6.3, we can write

$$(I - J)^{-1} = [X^t(W - V)X]^{-1} X^t W X$$

The true influence function for β is denoted by $IF_{\hat{\beta}}((x_0, y_0), F_n)$.

$$IF_{\hat{\beta}}((x_0, y_0), F_n) = n w_0 [X^t(W - V)X]^{-1} x_0^t (y_0 - \mu_0) \frac{\partial \eta_0}{\partial \mu_0}$$

6.5 Examples

In this section, we will discuss two models such as Poisson and binomial models with canonical link function.

6.5.1 Poisson Model

Consider the Poisson model with log link

$$\begin{aligned} var(y_i) &= \mu_i \quad \text{and} \\ \eta_i &= \log \mu_i \\ \frac{\partial \eta_i}{\partial \mu_i} &= \frac{1}{\mu_i} = \frac{1}{var(y_i)} \end{aligned}$$

Hence

$$w_i = \frac{\mu_i^2}{\mu_i} = \mu_i = var(y_i)$$

$$r_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}$$

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \exp(x_i \beta)) x_{ij} \end{aligned}$$

From (6.7), the one-step influence function for $\hat{\beta}$ under this model is given below

$$\begin{aligned} IF^1_{\hat{\beta}}((x_0, y_0), F_n) &= n var(y_0)(X^t W X)^{-1} x_0^t (y_0 - \mu_0) \frac{1}{var(y_0)} \\ &= n(X^t W X)^{-1} x_0^t (y_0 - \exp(\eta_0)) \end{aligned}$$

The details of the Jacobian matrix for this case follow

$$w_i = \mu_i = e^{\eta_i} \qquad r_i = \eta_i + \tfrac{y_i - e^{\eta_i}}{e^{\eta_i}}$$

$$w_i' = e^{\eta_i} = w_i \qquad r_i' = 1 - \tfrac{y_i}{e^{\eta_i}}$$

$$r_i' = 1 - \tfrac{y_i}{w_i}$$

$$\begin{aligned} v_i &= w_i' r_i + w_i r_i' - w_i' x_i \hat{\beta} \\ &= w_i \eta_i + w_i \frac{y_i - e^{\eta_i}}{e^{\eta_i}} + w_i - y_i - w_i \eta_i \\ &= 0 \end{aligned}$$

That is, $v_i = 0 \;\; i = 1 \ldots n$, which leads to give $J = 0$. Therefore, the true influence function and the one-step influence functions are similar. In general, the Jacobian matrix will vanish for the model with a canonical link. But this is not true for non-canonical links. For example, consider the Poisson model with identity link function.

$$\eta_i = \mu_i \text{ and } \frac{d\eta_i}{d\mu_i} = 1$$

$$w_i = \tfrac{1}{\mu_i} = \eta_i^{-1} \qquad r_i = \eta_i + (y_i - \eta_i) = y_i$$

$$w_i' = -\eta_i^{-2} \qquad r_i' = 0$$

$$\begin{aligned} v_i &= w_i'(r_i - \eta_i) + r_i' w_i \\ &= -\frac{r_i - \eta_i}{\eta_i^2} \\ &= -\frac{y_i - \mu_i}{\mu_i^2} \neq 0 \end{aligned}$$

6.5.2 Binomial

$$
\begin{aligned}
var(y_i) &= \frac{\mu_i(n_i - \mu_i)}{n} \quad \text{and} \\
\eta_i &= \log\left(\frac{\theta_i}{1-\theta_i}\right) = \log\mu_i - \log(n_i - \mu_i) \\
\frac{\partial \eta_i}{\partial \mu_i} &= \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{var(y_i)}
\end{aligned}
$$

Hence

$$
\begin{aligned}
w_i &= \frac{\mu_i(n_i - \mu_i)}{n_i} = var(y_i) \\
&= \left(\frac{n_i}{1+e^{-\eta_i}}\right)\left(\frac{1}{1+e^{\eta_i}}\right) \\
w_i' &= -\mu_i\left(\frac{1}{1+e^{-\eta_i}}\right)\left(\frac{1}{1+e^{\eta_i}}\right) + w_i\left(\frac{1}{1+e^{-\eta_i}}\right)\left(\frac{1}{1+e^{\eta_i}}\right) \\
&= \frac{\mu_i w_i^2}{n_i} - \frac{\mu_i^2 w_i}{n_i} \\
r_i &= \eta_i + \frac{y_i - \mu_i}{var(y_i)} \\
&= \eta_i + \frac{y_i - \mu_i}{w_i} \\
r_i' &= 1 + \frac{w_i\left(-\frac{d\mu_i}{d\eta_i}\right) - (y_i - \mu_i)w_i'}{w_i^2} \\
&= -\frac{(y_i - \mu_i)w_i'}{w_i^2}
\end{aligned}
$$

$$
\begin{aligned}
U_j(\beta) &= \sum_{i=1}^{n}(y_i - \mu_i)x_{ij} \\
&= \sum_{i=1}^{n}\left(y_i - \left[\frac{n_i}{1+\exp(-\eta_i)}\right]\right)x_{ij}
\end{aligned}
$$

For binomial model with logit link, the one-step influence function for $\hat{\beta}$ form (6.7) is

$$
IF^1_{\hat{\beta}}((x_0, y_0), F_n) = n(X^tWX)^{-1}x_0^t\left(y_0 - \frac{1}{1+\exp(-\eta_0)}\right)
$$

Due to the canonical link, elements of the Jacobian matrix J is zero. Therefore $IF^1_{\hat{\beta}}((x_0, y_0), F_n)$ is considered as the true influence function for $\hat{\beta}$.

6.6 Mixture Model

We estimate β using the mixture model

$$
p(y, X, \beta) = \lambda f(y) + (1-\lambda)g(y) \tag{6.11}
$$

where g is a dispersed parameter free function over the sample space and $1-\lambda$ is a fixed small positive number which may be thought of as the proportion of contaminated data. We will often choose λ to be 0.95 or similar. Remember f is defined in (6.1).

For fixed choices of g and λ, we will consider the robustness properties of

$$
\tilde{\beta} = \arg\max_{\beta} L_o(\beta) \tag{6.12}
$$

where $L_o(\beta) = \Pi_{i=1}^n p(y_i, x_i, \beta)$ is the observed likelihood function for β. The $\tilde{\beta}$ is considered as our robustified estimator.

6.7 Calculating the Robustified Estimator

It is algebraically difficult to maximize the observed likelihood function, $L_o(\beta)$, so that the complete likelihood function, $L_c(\beta)$, is considered to achieve (6.12) and is defined below.

$$
L_c(\beta) = \Pi_{i=1}^n \left[[\lambda f(y_i)]^{z_i}\left[(1-\lambda)g(y_i)\right]^{1-z_i}\right] \tag{6.13}
$$

where

$$z_i = \begin{cases} 1 & \text{if } y_i \in f \\ 0 & \text{if } y_i \in g \end{cases}$$

$$\begin{aligned} l_c(\beta) &= \log L_c(\beta) \\ &= \sum_{i=1}^{n} [z_i \log \lambda + z_i[y_i b(\theta_i) + c(\theta_i) + d(y_i)] + \\ &\quad (1 - z_i) \log(1 - \lambda) + (1 - z_i) \log g(y_i)] \\ &= \sum_{i=1}^{n} l_i \end{aligned}$$

where

$$l_i = z_i \log \lambda + z_i[y_i b(\theta_i) + c(\theta_i) + d(y_i)] + (1 - z_i) \log(1 - \lambda) + (1 - z_i) \log g(y_i)$$

6.7.1 E-Step

In order to maximize $l_c(\beta)$ with respect to β, we need to know the values of the unknown z_i. This unknowns can be replaced by expected values. In other words, z values are computed (as explained in section 4.5.1) at the E-step of the EM algorithm

$$\tilde{z}_i = E[z_i | y_i, x_i, \beta] = \frac{\lambda f(y_i)}{\lambda f(y_i) + (1 - \lambda) g(y_i)} \tag{6.14}$$

Sometimes we denote $\tilde{z}_i = z(y_i, \eta_i) = z(y_i, x_i, \beta) = z(y_i - x_i\beta)$.

6.7.2 M-Step

Now we can replace the z_i by $\tilde{z}_i \; \forall i$ in the expression of $l_c(\beta)$ and we can apply standard MLE procedures, as explained in section 6.3, to compute the statistics for β. That is, we need to solve the following simultaneous equations

$$U(\beta) = \frac{\partial l_c(\beta)}{\partial \beta} = 0 \tag{6.15}$$

where

$$U(\beta) = \begin{pmatrix} U_1(\beta) \\ . \\ \\ . \\ U_p(\beta) \end{pmatrix} = \begin{pmatrix} \frac{\partial l_c(\beta)}{\partial \beta_1} \\ . \\ \\ . \\ \frac{\partial l_c(\beta)}{\partial \beta_p} \end{pmatrix}$$

Consider

$$\begin{aligned} U_j(\beta) &= \frac{\partial l_c(\beta)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \end{aligned}$$

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= \tilde{z}_i \left(y_i \, b'(\theta_i) + c'(\theta_i) \right) \\ &= \tilde{z}_i \, b'(\theta_i) \left(y_i - \mu_i \right) \end{aligned}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b'(\theta_i) \, var(y_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Hence

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^{n} \tilde{z}_i \, b'(\theta_i) \left(y_i - \mu_i \right) \left(\frac{1}{b'(\theta_i) \, var(y_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \sum_{i=1}^{n} \tilde{z}_i \left(\frac{y_i - \mu_i}{var(y_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ U(\beta) &= X^t Z \, W \, \Gamma \end{aligned}$$

where Γ is a column vector, whose elements $\gamma_i = (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$, Z is a diagonal matrix whose elements are $\tilde{z}_i$, and W is a diagonal matrix, whose elements are w_i

$$w_i = \frac{[\frac{\partial \mu_i}{\partial \eta_i}]^2}{var(y_i)}$$

The $(p,q)^{\text{th}}$ element of Σ is

$$\begin{aligned}
\Sigma_{pq} &= E(U_p^t U_q) \\
&= E[\sum_{i=1}^{n} \tilde{z}_i \left(\frac{y_i-\mu_i}{var(y_i)}\right)\frac{\partial\mu_i}{\partial\eta_i}x_{ip}\sum_{j=1}^{n}\tilde{z}_j\left(\frac{y_j-\mu_j}{var(y_j)}\right)\frac{\partial\mu_j}{\partial\eta_j}x_{jq}] \\
&= \sum_{i=1}^{n}\tilde{z}_i^2\left(\frac{\frac{\partial\mu_i}{\partial\eta_i}}{var(y_i)}\right)^2 E[(y_i-\mu_i)^2]x_{ip}x_{iq} \\
&= \sum_{i=1}^{n}\tilde{z}_i^2\left(\frac{[\frac{\partial\mu_i}{\partial\eta_i}]^2}{var(y_i)}\right)x_{ip}x_{iq} \\
\Sigma &= X^t Z W Z X
\end{aligned}$$

Consider the right hand side of (6.5)

$$\begin{aligned}
\Sigma\beta + U &= X^t Z W Z X\beta + X^t Z W \Gamma \\
&= X^t Z W Z (X\beta + Z^{-1}\Gamma) \\
&= X^t Z W Z R
\end{aligned}$$

where R is a column vector whose elements are

$$r_i = x_i\beta + \left(\frac{y_i-\mu_i}{z_i}\right)\frac{\partial\eta_i}{\partial\mu_i} = \eta_i + \left(\frac{y_i-\mu_i}{z_i}\right)\frac{\partial\eta_i}{\partial\mu_i}$$

Hence, (6.5) becomes

$$X^t Z^{(m-1)} W^{(m-1)} Z^{(m-1)} X\tilde{\beta}^{(m)} = X^t Z^{(m-1)} W^{(m-1)} Z^{(m-1)} R^{(m-1)}$$

$$X^t A^{(m-1)} X\tilde{\beta}^{(m)} = X^t A^{(m-1)} R^{(m-1)} \tag{6.16}$$

where A is a diagonal matrix, whose elements a_i are

$$a_i = \tilde{z}_i^2\ w_i$$

Hence, the estimate for β can be obtained by an IRLS method, because (6.16) takes the same form of weighted least squares with weights a_i and it is an iterative process because R and A depend on β.

The limit of the sequence $\{\tilde{\beta}^{(m)}\}_{m=1}^{\infty}$, if it converges, is the estimate for β. That is,

$$\tilde{\beta} = \lim_{m \to \infty} \tilde{\beta}^{(m)}$$

Note that if Z is an identity matrix, $\tilde{\beta} = \hat{\beta}$.

6.8 Influence Function for $\tilde{\beta}$

Computation of the true influence function for $\tilde{\beta}$ is very complicated, because each observation is associated with two weights such as z_i and w_i. Since $\tilde{\beta}$ is computed by IRLS method, the Jorgensen method is employed to compute the true influence function for $\tilde{\beta}$.

6.8.1 One-Step Influence Function for $\tilde{\beta}$

The expressions (6.6) and (6.16) are similar in form for estimating the β, but they have different weights and different adjusted dependent variables. Therefore the one-step influence function for $\tilde{\beta}$ can be predicted from (6.7) by replacing W by A and by replacing r_0 by $r_0 = \eta_0 + \left(\frac{y_0 - \mu_0}{z_0}\right)\frac{\partial \eta_0}{\partial \mu_0}$. That is

$$IF^1_{\tilde{\beta}}((x_0, y_0), F_n) = na_0(X^t A X)^{-1} x_0^t (r_0 - x_0\hat{\beta})$$

where $a_0 = \tilde{z}_0^2 w_0$

$$IF^1_{\tilde{\beta}}((x_0, y_0), F_n) = na_0(X^t A X)^{-1} x_0^t \left(\frac{y_0 - \mu_0}{z_0}\right)\frac{\partial \eta_0}{\partial \mu_0} \tag{6.17}$$

6.8.2 Jacobian Matrix

Next we compute the Jacobian matrix J. The updating function $h(\beta, F_n)$ can be defined in (6.18).

$$X^t Z W Z X h(\beta, F_n) = X^t Z W Z R \tag{6.18}$$

Differentiate (6.18) both sides with respect to β.

$$\begin{aligned} X^t\frac{\partial Z}{\partial \beta}WZXh(\beta,F_n) &+ X^tZ\frac{\partial W}{\partial \beta}ZXh(\beta,F_n) + \\ X^tZW\frac{\partial Z}{\partial \beta}Xh(\beta,F_n) &+ X^tZWZX\frac{\partial h(\beta,F_n)}{\partial \beta} \\ &= \\ X^t\frac{\partial Z}{\partial \beta}WZR &+ X^tZ\frac{\partial W}{\partial \beta}ZR + \\ X^tZW\frac{\partial Z}{\partial \beta}R &+ X^tZWZ\frac{\partial R}{\partial \beta} \end{aligned}$$

$$\begin{aligned} X^tZWZX\frac{\partial h(\beta,F_n)}{\partial \beta} &= X^t\frac{\partial Z}{\partial \beta}WZ(R-Xh(\beta,F_n)) \\ &+ X^tZ\frac{\partial W}{\partial \beta}Z(R-Xh(\beta,F_n)) \\ &+ X^tZW\frac{\partial Z}{\partial \beta}(R-Xh(\beta,F_n)) + X^tZWZ\frac{\partial R}{\partial \beta} \end{aligned}$$

Let $S = R - Xh(\beta, F_n)$ be a n by 1 column vector. Then

$$\begin{aligned} X^tZWZX\frac{\partial h(\beta,F_n)}{\partial \beta} &= X^t\frac{\partial Z}{\partial \beta}WZS + X^tZ\frac{\partial W}{\partial \beta}ZS \\ &+ X^tZW\frac{\partial Z}{\partial \beta}S + X^tZWZ\frac{\partial R}{\partial \beta} \end{aligned}$$

Consider

$$\frac{\partial Z}{\partial \beta_j}WZ = \begin{pmatrix} z_1'x_{1j}w_1z_1 & 0 & . & 0 \\ 0 & z_2'x_{2j}w_2z_2 & . & 0 \\ . & . & . & . \\ \\ . & . & . & . \\ 0 & 0 & . & z_n'x_{nj}w_nz_n \end{pmatrix}$$

$$
\begin{aligned}
\frac{\partial Z}{\partial \beta_j} WZS &= \begin{pmatrix} z_1' x_{1j} w_1 z_1 s_1 \\ z_2' x_{2j} w_2 z_2 s_2 \\ . \\ \\ . \\ z_n' x_{nj} w_n z_n s_n \end{pmatrix} \\
&= \begin{pmatrix} z_1' w_1 z_1 s_1 & 0 & . & 0 \\ 0 & z_2' w_2 z_2 s_2 & . & 0 \\ . & . & . & . \\ \\ . & . & . & . \\ 0 & 0 & . & z_n' w_n z_n s_n \end{pmatrix} \begin{pmatrix} x_{1j} \\ x_{2j} \\ . \\ \\ . \\ x_{nj} \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Z}{\partial \beta} WZS &= \begin{pmatrix} z_1' w_1 z_1 s_1 & 0 & . & 0 \\ 0 & z_2' w_2 z_2 s_2 & . & 0 \\ . & . & . & . \\ \\ . & . & . & . \\ 0 & 0 & . & z_n' w_n z_n s_n \end{pmatrix} X \\
&= \begin{pmatrix} z_1' s_1 & 0 & . & 0 \\ 0 & z_2' s_2 & . & 0 \\ . & . & . & . \\ \\ . & . & . & . \\ 0 & 0 & . & z_n' s_n \end{pmatrix} WZX \\
&= \Lambda_1 WZX
\end{aligned}
$$

Similarly, we can write

$$Z \frac{\partial W}{\partial \beta} ZS = Z \Lambda_2 ZX$$

and

$$ZW \frac{\partial Z}{\partial \beta} S = ZW \Lambda_1 X$$

where

$$\Lambda_1 = \begin{pmatrix} z_1' s_1 & 0 & . & 0 \\ 0 & z_2' s_2 & . & 0 \\ . & . & . & . \\ & & & \\ . & . & . & . \\ 0 & 0 & . & z_n' s_n \end{pmatrix} \quad \text{and} \quad \Lambda_2 = \begin{pmatrix} w_1' s_1 & 0 & . & 0 \\ 0 & w_2' s_2 & . & 0 \\ . & . & . & . \\ & & & \\ . & . & . & . \\ 0 & 0 & . & w_n' s_n \end{pmatrix}$$

$$\frac{\partial R}{\partial \beta_j} = \begin{pmatrix} r_1' x_{1j} \\ r_2' x_{2j} \\ . \\ \\ . \\ r_n' x_{nj} \end{pmatrix}$$

$$\frac{\partial R}{\partial \beta} = \Lambda_3 X$$

where

$$\Lambda_3 = \begin{pmatrix} r_1' & 0 & . & 0 \\ 0 & r_2' & . & 0 \\ . & . & . & . \\ & & & \\ . & . & . & . \\ 0 & 0 & . & r_n' \end{pmatrix}$$

$$\begin{aligned} X^t Z W Z X \frac{\partial h(\beta, F_n)}{\partial \beta} &= X^t [\Lambda_1 W Z + Z \Lambda_2 Z + Z W \Lambda_1 + Z W Z \Lambda_3] X \\ &= X^t [2 \Lambda_1 W Z + Z \Lambda_2 Z + Z W Z \Lambda_3] X \\ &= X^t V X \end{aligned}$$

where $V = 2\Lambda_1 W Z + Z \Lambda_2 Z + Z W Z \Lambda_3$ is a diagonal matrix, whose elements are v_i

$$v_i = 2 z_i' s_i w_i z_i + z_i^2 w_i' s_i + z_i^2 w_i r_i' \; \forall i = 1 \ldots n$$

Hence

$$J = \left[\frac{\partial h(\beta, F_n)}{\partial \beta} \right]_{\beta = \tilde{\beta}} = (X^t A X)^{-1} X^t V X \tag{6.19}$$

Note that if Z is an identity matrix, (6.19) and (6.10) are exactly same.

6.9 Applications

In this section, we analyse two real examples. We investigate examples of the binomial model and the Poisson model.

6.9.1 Binomial Models

In this section, we analyse the *leukemia* data set, which was considered by Cook and Weisberg ([10], Chapter 5, p.193). Later a number of authors used this dataset for their analysis. The data set consist of 33 leukemia patients. The response variable y is defined as follows

$$y = \begin{cases} 1 & \text{if the patient survives at least 52 weeks} \\ 0 & \text{otherwise} \end{cases}$$

Two covariates of white blood cell count (WBC), and presence or absence of certain morphological characteristics in the white cells (AG) were considered.

Let $Y_i = \sum_{j=1}^{n_i} y_j$ be the number of survivors in the group i, where n_i is the sample size of group i. We assume a binomial model with logit link

$$\log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \beta_0 + \beta_1 WBC_i + \beta_2 AG_i \; i = 1 \ldots 30 \tag{6.20}$$

where $\mu_i = E[Y_i]$.

The maximum likelihood estimate for $\beta = (\beta_0, \beta_1, \beta_2)^t$ is given in Table 6.2. Cook and Weisberg found that observation 15 is unusual, because the patient 15 survived for a long time period when WBC = 100000. The MLE estimates are heavily affected by this observation (See Figure 6.1). The result MLE_{-15}, MLE is obtained after removing observation 15, is given in Table 6.2. It was noticed by the authors that this fit is much better than MLE fit.

We like to apply the mixture method to this data. We choose $g = \frac{1}{2}$ as a equal chance of alive and death. Estimates for β are computed for the various values of $\lambda = 0.9, 0.95$, and true value of $\lambda = 0.97(32/33)$. The results are given in Table 6.2. We can observe that coefficients fitted with the mixture model for various λ are very similar. In addition, these estimates are very similar to the MLE_{-15}. That means our method automatically takes

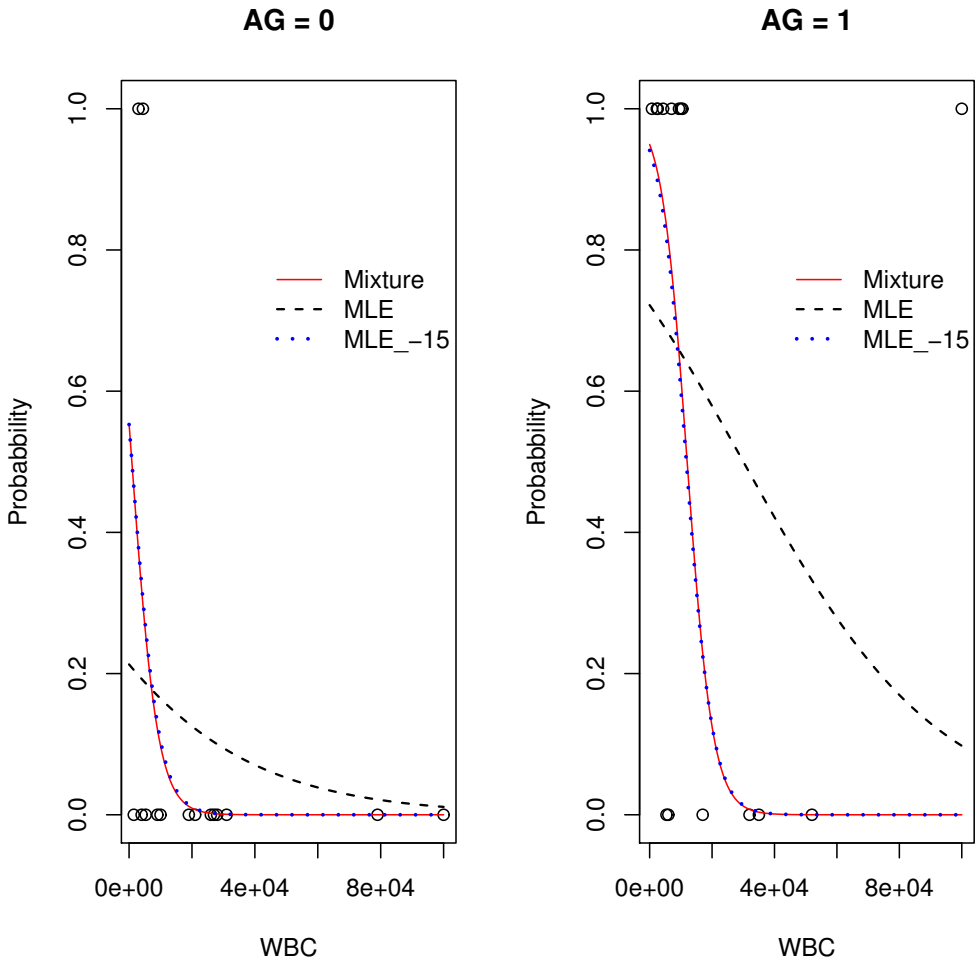

Figure 6.1: Leukemia data: GLM fit with all data and omitting case 15, and mixture fit

Table 6.2: Estimates for leukemia data

	MLE	$\mathbf{MLE_{-15}}$	$\boldsymbol{\lambda = 0.90}$	$\boldsymbol{\lambda = 0.95}$	$\boldsymbol{\lambda = 0.97}$
Estimate for $\boldsymbol{\beta_0}$	-1.307	0.212	0.228	0.216	0.214
Estimate for $\boldsymbol{\beta_1}$	-0.000032	0.000235	-0.000255	-0.000243	-0.000240
Estimate for $\boldsymbol{\beta_2}$	2.261	2.558	2.911	2.717	2.650

into account the particular observation of 15.

The true and one-step influence functions of the mixture estimates are numerically computed when $\lambda = 0.95$. The Jacobian matrix is also computed and given below.

$$\begin{bmatrix} 3.50642489760 & -7.490251e+03 & -1.6366660035 \\ 0.00038329830 & 2.3796280e-01 & 0.00034136500 \\ -3.7618873008 & -1.806380e+04 & -5.2759043718 \end{bmatrix}$$

The response variable y has two outcomes and one of the predictor variables (AG) is also categorical, so that influence functions of $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ are computed for the four cases $y = 1$ and $AG = 1$; $y = 1$ and $AG = 0$; $y = 0$ and $AG = 1$; $y = 0$ and $AG = 0$. These are shown in the Figures 6.1, 6.3, 6.4, 6.5 in order. These plots show that the true influence functions are bounded for all estimates. Hence, we may say that mixture estimates are B-robust.

Next, results obtained by our method are compared with existing methods mentioned in sections 1.4.4 and 1.6. These are given in Table 6.3. The mixture model estimates are better than other estimates, because mixture model estimates are very similar to the MLE_{-15}. The estimate obtained by the quasi-likelihood approach is also reasonably close to the MLE_{-15}, where

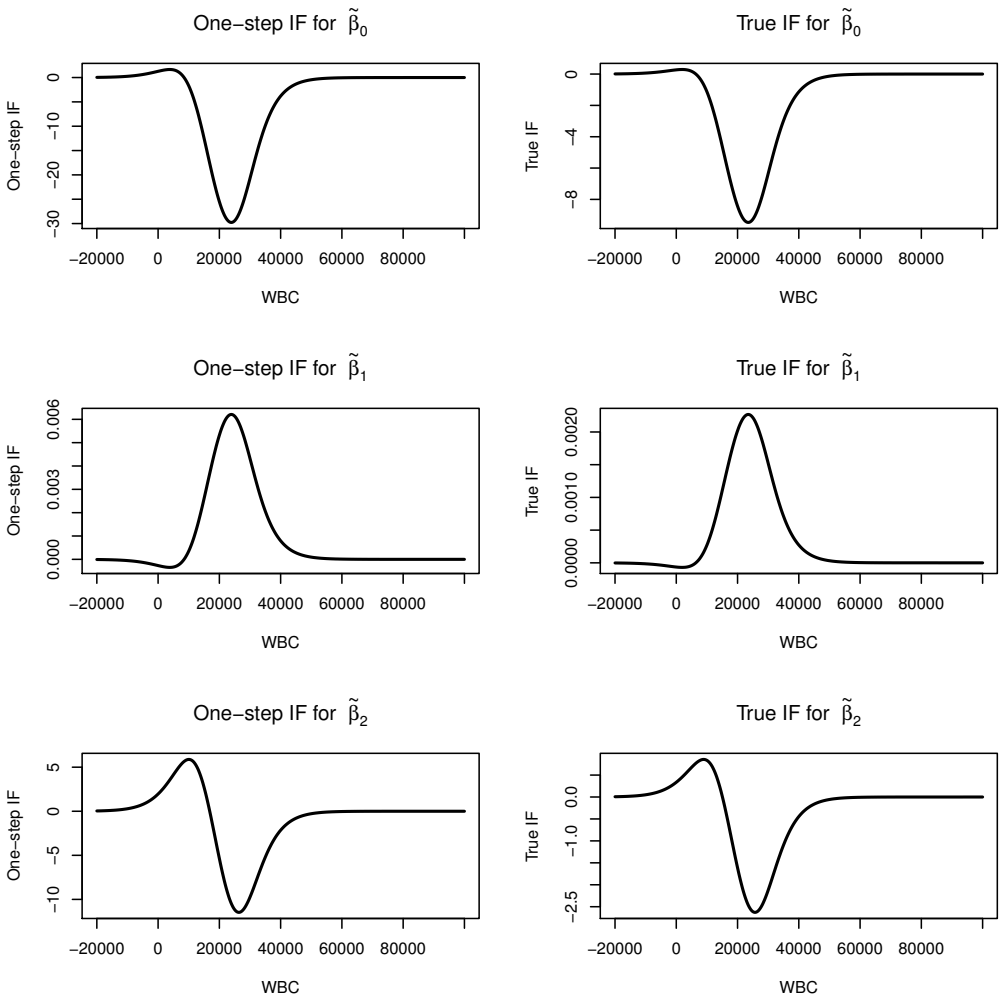

Figure 6.2: The one-step influence function and true influence function for the mixture estimates when $y = 1$ and $AG = 1$

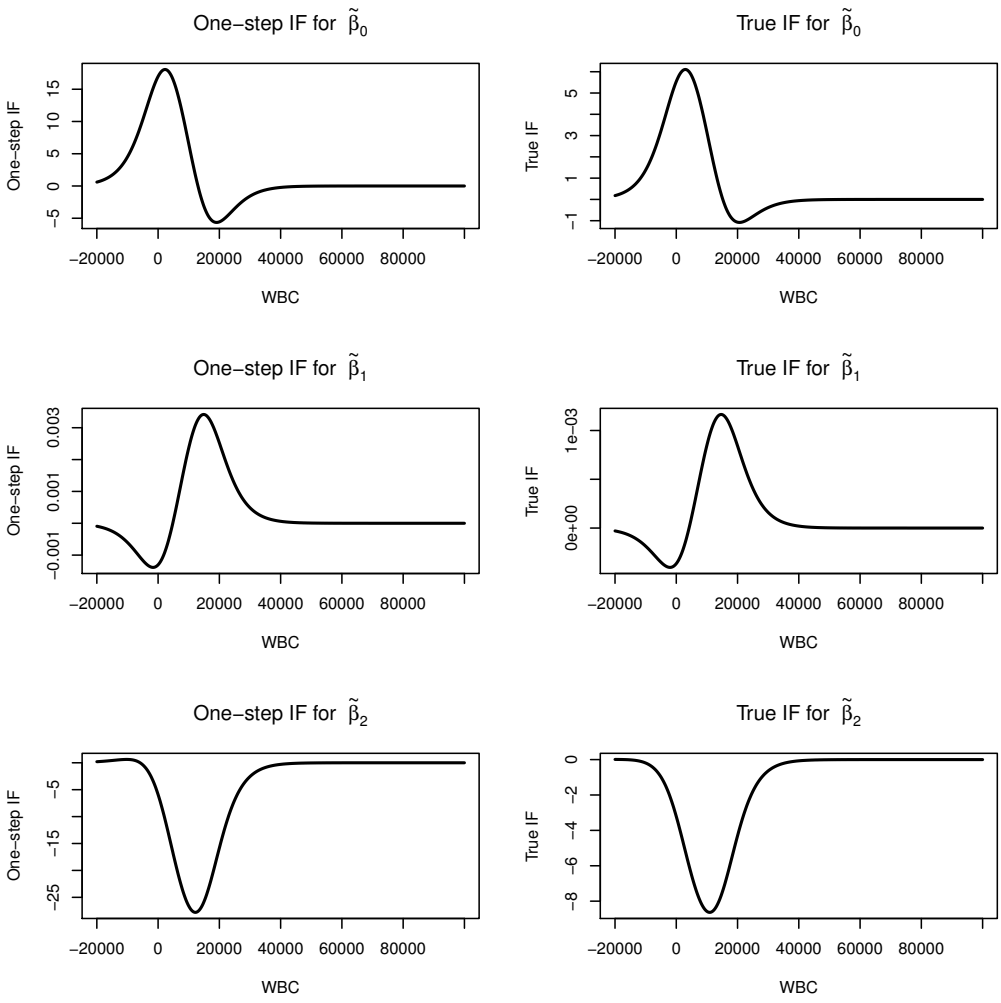

Figure 6.3: The one-step influence function and true influence function for the mixture estimates when $y = 1$ and $AG = 0$

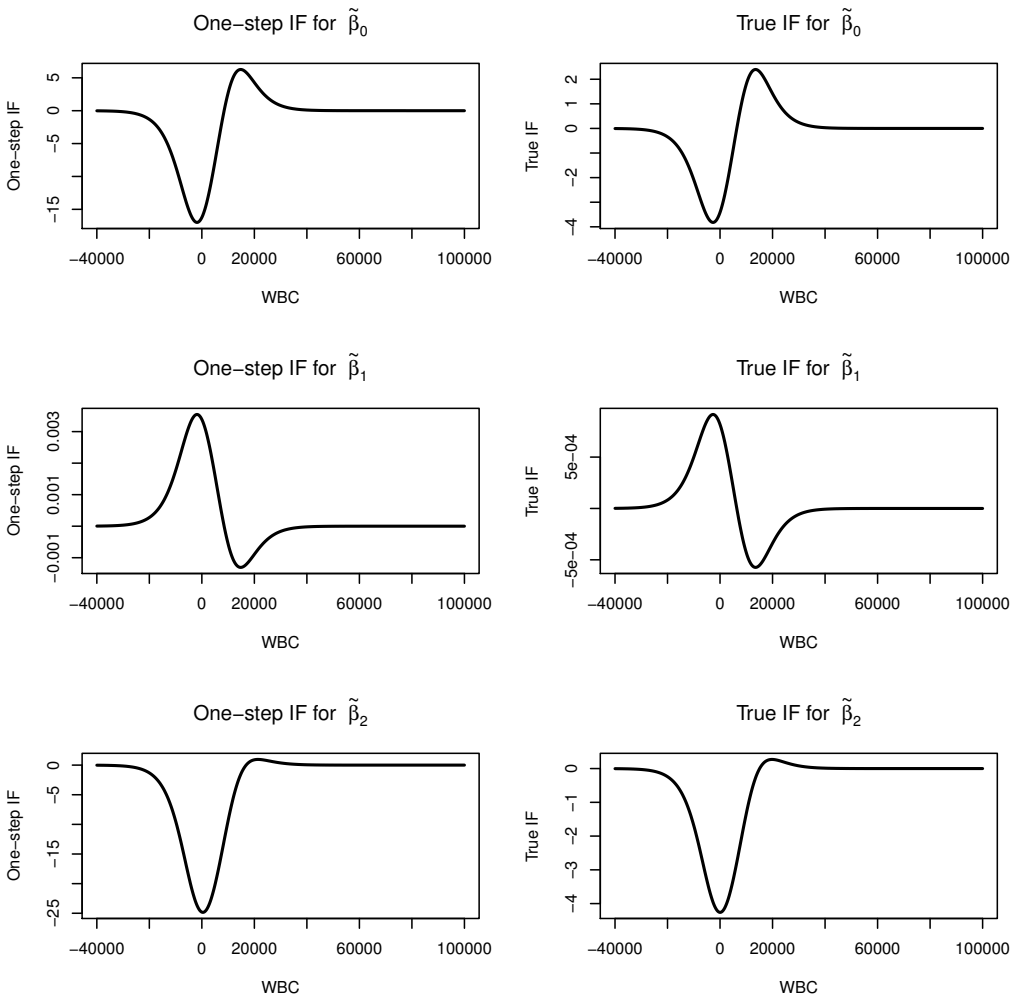

Figure 6.4: The one-step influence function and true influence function for the mixture estimates when $y = 0$ and $AG = 1$

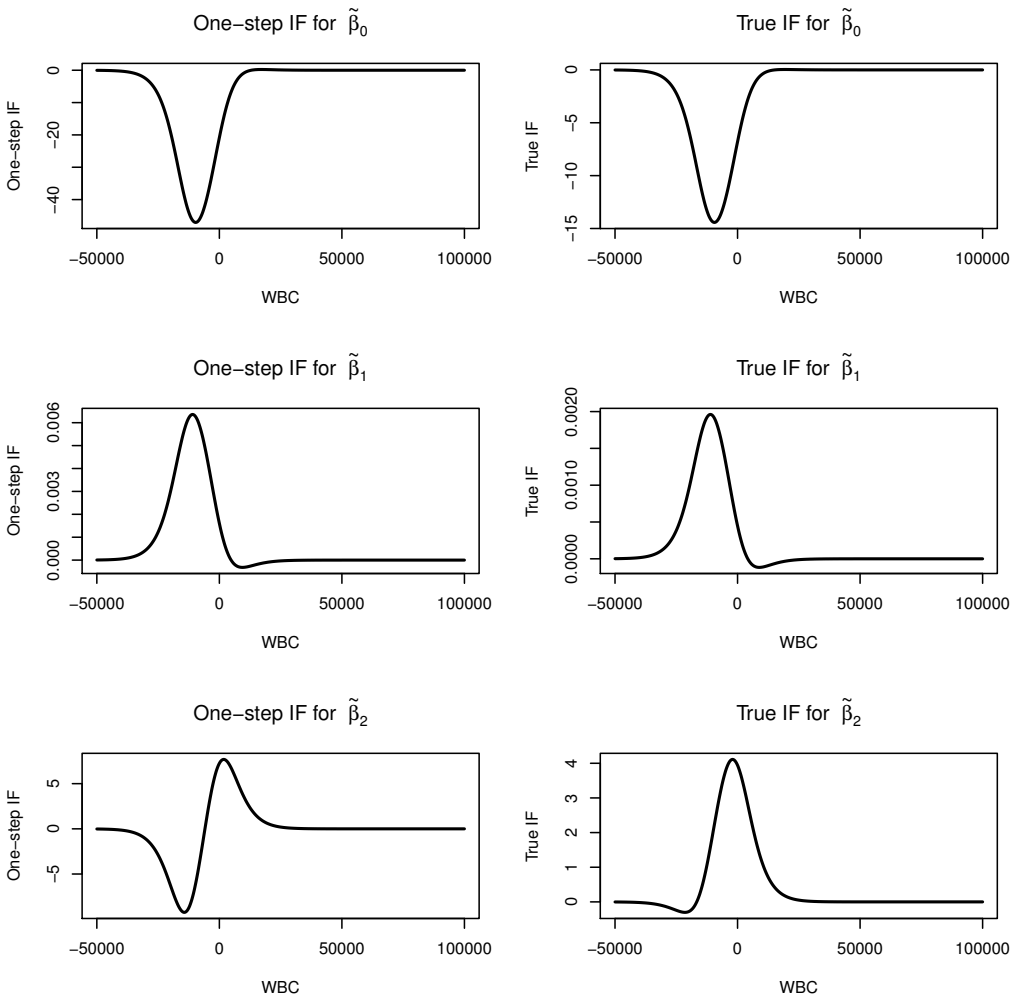

Figure 6.5: The one-step influence function and true influence function for the mixture estimates when $y = 0$ and $AG = 0$

Table 6.3: Estimates for leukemia data based on the various methods

	Estimate for β_0	**Estimate for β_1**	**Estimate for β_2**
Mixture with $\lambda = 0.95$	0.216	-0.000243	2.717
CUBIF ([35], Page 242)	-0.678	-0.0000909	2.249
Cantoni and Ronchetti ([8], 2001)	0.171	-0.000204	2.487

$\sqrt{1-h_{ii}}$ is used as weight for the i^{th} row of the design matrix X and h_{ii} is a i^{th} diagonal element of the hat matrix.

6.9.2 Poisson Models

In this section, we use data from the Canadian Equality, Security, and Community Survey of 2000. Altogether, 4594 observations were collected across Canada. Andersen [1] analyzed this data for Quebec province only, which had 949 respondents. We also use same data set for our investigation.

The response variable is the number of voluntary association to which respondents belonged. The predictors are not continuous variables such as gender (2 levels: Women, Men); Canadian born (2 levels: No, Yes); language spoken in the home (3 levels: French, English, Other). The first level in each variable is considered as the reference category. Figure 6.6 shows the distribution of the response variable based on the predictor variables. Because

of the count response, we fit the Poisson regression model using IRLS under the frame-work of GLM. Table 6.4 displays the results from the GLM model.

Figure 6.7 shows the plot for the Cook's distance with a red dashed line indicating the cut-off defined by Fox ([17], 281). We may say, from this plot, that a large number of observations are influential for these estimates (MLE). However, if we apply the Cook and Weisberg ([11], 358) criteria for the cut-out point, none of the observations are influential for these estimates. These cut-off rules are basically rule of thumb, so that problematic cases are not easily distinguished in the practical situations.

Before applying our method, we use the quasi-likelihood [8] method to obtain robustified estimates for the Poisson model parameters. This method is used with and without applying the weights to predictors. The results are very similar (see Table 6.4). This indicates that values in the design matrix do not influence the estimates.

For the mixture model, the parameter free density function g is defined as a uniform distribution with parameters $[a = 0, b = 25]$, because the minimum and the maximum of the response variable are 0 and 13 respectively, and the response variable cannot be a negative value. The results are given in Table 6.4, where we can observe that the coefficients fitted with the mixture are different to the MLE, but reasonably similar to the quasi-likelihood estimates. Figures 6.8 and 6.9 show the fit made by mixture is slightly better than the quasi-likelihood fit, because the mixture fit gives a better representation of the majority of the data. Note that these graphs are in different y-scale.

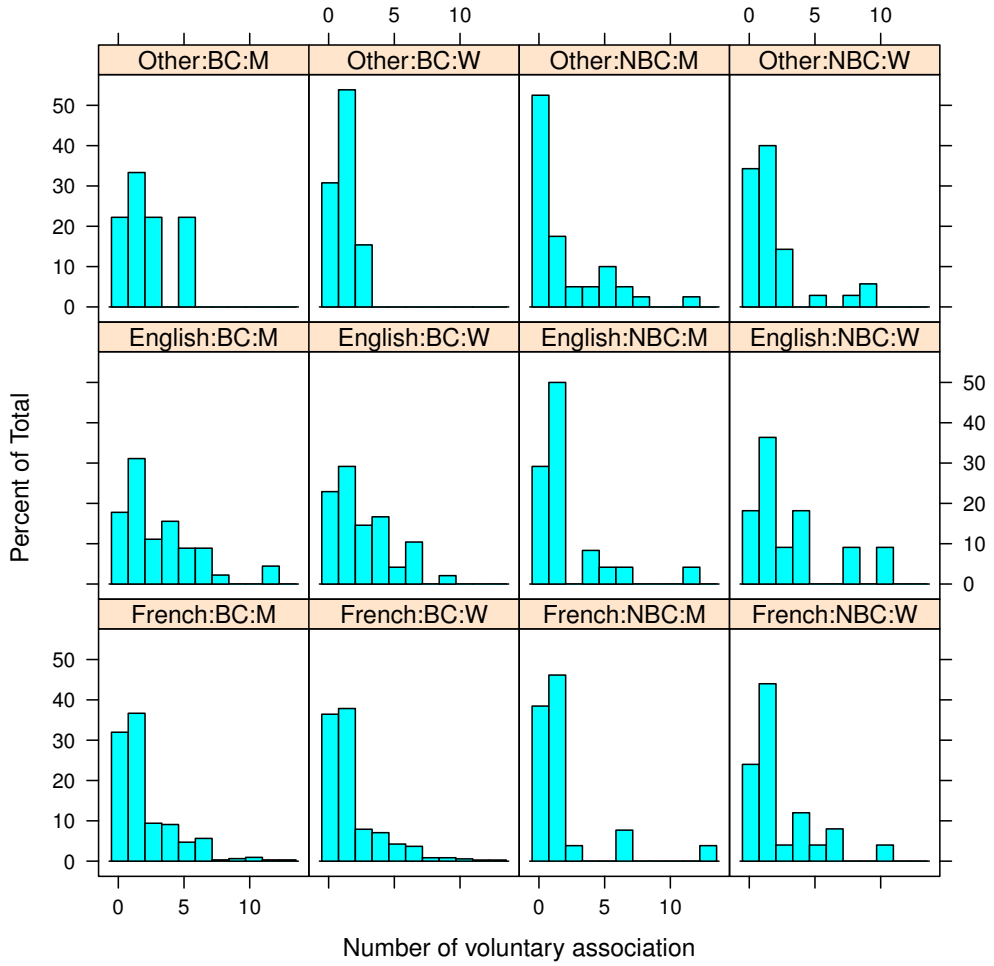

Figure 6.6: Histogram of the Quebec data for the various groups. Notation: BC - Born in Canada; NBC - Not born in Canada; M - Men; W - Women

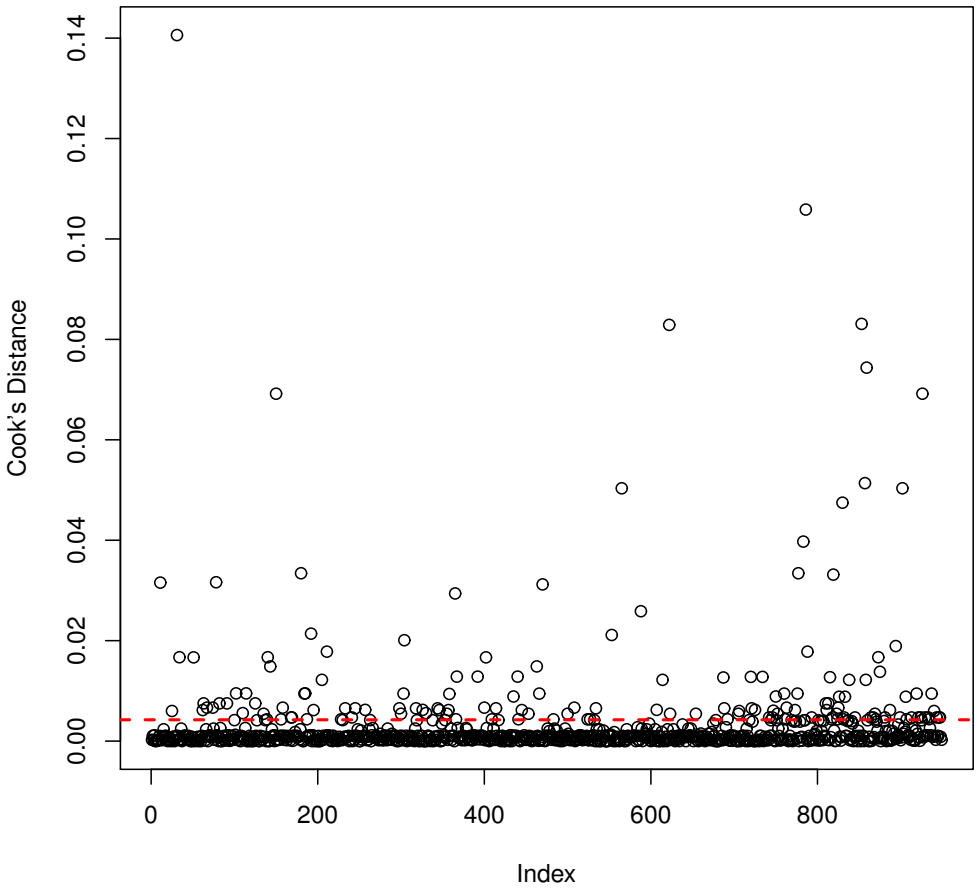

Figure 6.7: Quebec Data: Cook's distance based on the MLE fit

Table 6.4: Estimates for Quebec data based on the various methods

Estimate	**MLE**	**CR without weights on X**	**CR with weights on X**	**Mixture $\lambda = 0.95$**
β_0 **(Intercept)**	0.586	0.120	0.120	0.030
β_1 **(Men)**	0.079	0.084	0.085	0.110
β_2 **(Canadian Born)**	0.027	0.258	0.258	0.356
β_3 **(English)**	0.357	0.537	0.538	0.495
β_4 (Other)	-0.014	0.079	0.079	0.102

CR-Cantoni and Ronchetti 2001 method [8]

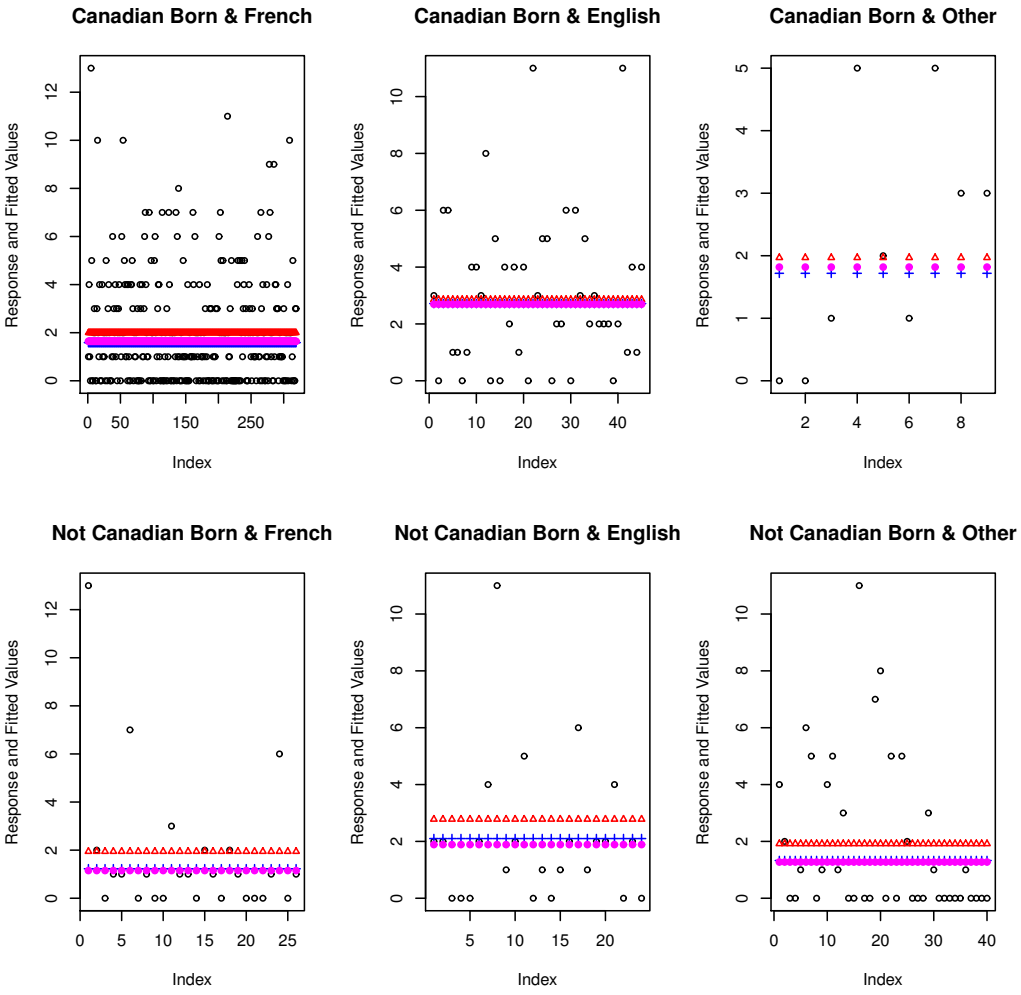

Figure 6.8: Quebec Data: GLM fit, quasi-likelihood fit and mixture fit for the men group: circle with black colour indicate observed data, triangle with red colour indicate GLM fit, plus sign with blue colour indicate quasi-likelihood fit and dark circle with magenta colour indicate mixture fit.

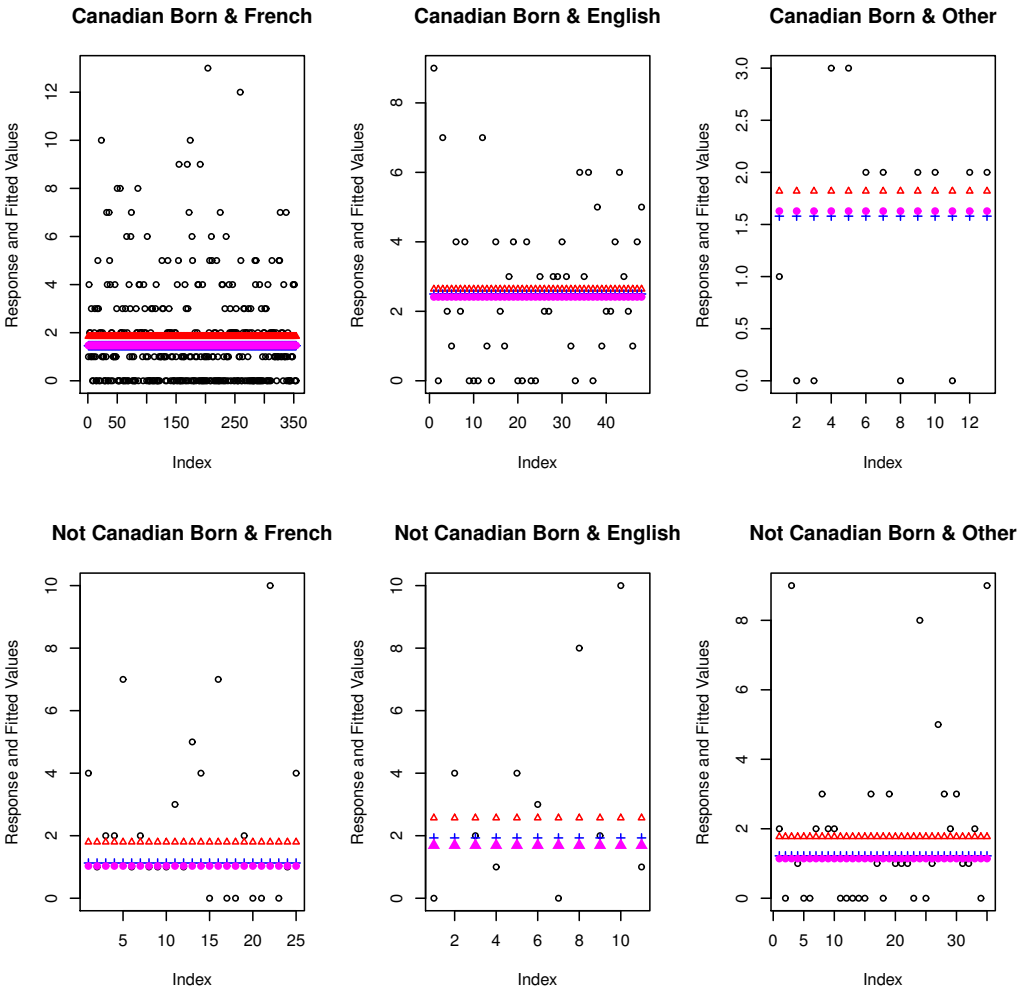

Figure 6.9: Quebec Data: GLM fit, quasi-likelihood fit and mixture fit for the women group: circle with black colour indicate observed data, triangle with red colour indicate GLM fit, plus sign with blue colour indicate quasi-likelihood fit and dark circle with magenta colour indicate mixture fit.

Chapter 7

Extension of Robust Estimation for Linear Models

In Chapter 5, the linear regression parameters were computed based on the assumption of the scale parameter σ being known or previously estimated. The goal of this chapter is to deal with estimation of regression parameter, β, and scale parameter σ simultaneously. In other words, this chapter is similar to the location and scale estimation of the robust literature.

We use the same notation as Chapter 5 and make links with the chapter and this chapter structure is very similar to it. We use Example 5.1 to show the estimation of regression parameters and the scale parameter together. At the end of this chapter, we investigate the speed of light data.

7.1 Model

Consider the model

$$y = X\beta + \sigma\epsilon \tag{7.1}$$

where σ is unknown and the probability density function of random variable Y is defined in (5.4). We consider regression coefficient estimates and scale estimate that are maximum likelihood estimates of β and σ with respect to the density in (5.4).

7.2 Calculating the Estimator

Let $\phi = (\beta, \sigma^2)^t$ be the model parameter. The parameter estimate ϕ is defined by

$$\hat{\phi} = \arg\max_{\phi} \{L(\phi)\} \tag{7.2}$$

where $L(\phi)$ is the likelihood function for ϕ

$$\begin{aligned} L(\phi) &= \Pi_{i=1}^{n} f(y_i) \\ l(\phi) &= \log L(\phi) \\ &= \sum_{i=1}^{n} \text{constant} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma}\right)^2 \end{aligned}$$

A necessary condition for $\hat{\phi}$ in (7.2) is

$$\frac{\partial l(\phi)}{\partial \beta} = 0$$

$$\sum_{i=1}^{n} \left(\frac{y_i - x_i\hat{\beta}}{\sigma}\right) x_{ij} = 0 \ \forall j = 1 \ldots p \tag{7.3}$$

and

$$\frac{\partial l(\phi)}{\partial \sigma^2} = 0$$

$$\sum_{i=1}^{n} \left[-\frac{1}{2\hat{\sigma}^2} + \frac{(y_i - x_i\hat{\beta})^2}{2\hat{\sigma}^4}\right] = 0 \tag{7.4}$$

These give the following results

$$\hat{\beta} = (X^t X)^{-1} X^t y \tag{7.5}$$

and

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^t (y - X\hat{\beta})}{n} \tag{7.6}$$

If α_i is a weight associated with the observations (x_i, y_i) for all $i = 1 \ldots n$, then the estimates are

$$\hat{\beta} = (X^t \Lambda X)^{-1} X^t \Lambda y \text{ and}$$

$$\hat{\sigma}^2 = (\mathbf{1}^t \Lambda \mathbf{1})^{-1} (y - X\hat{\beta})^t \Lambda (y - X\hat{\beta})$$

where Λ is a n by n diagonal matrix, whose elements are $\Lambda_i = \alpha_i^{0.5}\, i = 1 \ldots n$ and $\mathbf{1}$ is a column vector of ones.

In the case of Example 5.1, $\hat{\beta}_0 = -260.059, \hat{\beta}_1 = 5.041$ and $\hat{\sigma}^2 = 2897.647$. Note that they are computed separately and we have exactly same results as were obtained in the section 5.2. These estimates are influenced by the outliers (see section 5.3). Next we compute robustified estimates for ϕ.

7.3 Mixture Model

Again, we seek for a procedure that gives a good fit to the bulk of the data without being perturbed by a small proportion of outliers, and that does not require us to decide which observations are outliers. In order to achieve this goal, we consider the mixture model

$$p(y, X, \phi) = \lambda f(y) + (1 - \lambda) g(y) \tag{7.7}$$

where g is a dispersed parameter free function over the sample space and $1 - \lambda$ is a fixed small positive number which may be thought of as the proportion of contaminated data. We will often choose λ to be 0.95 or similar. Remember f is defined in (5.4.

For fixed choices of g and λ, we will consider the robustness properties of

$$\tilde{\phi} = \arg\max_{\phi} L_o(\phi) \tag{7.8}$$

where $L_o(\phi)$ is the observed likelihood function for ϕ.

$$L_o(\phi) = \Pi_{i=1}^{n} \left[\lambda f(y_i) + (1 - \lambda) g(y_i)\right]$$

We believe $\tilde{\phi}$ is our robustified estimator.

7.4 Calculating the Robustified Estimator

For mathematical simplification, we often consider complete likelihood function $L_c(\phi)$ to achieve (7.8).

$$\tilde{\phi} = \arg\max_{\phi} L_c(\phi) \tag{7.9}$$

where

$$L_c(\phi) = \Pi_{i=1}^{n} \left[[\lambda f(y_i)]^{z_i} \left[(1-\lambda) g(y_i) \right]^{1-z_i} \right]$$

$$z_i = \begin{cases} 1 & \text{if } y_i \in f \\ 0 & \text{if } y_i \in g \end{cases}$$

In general, z_i's are not observed, so z_i's are treated as unobserved random variables. They can be estimated using the E-step of the EM algorithm followed by estimation the parameters, β and σ^2 at the M-step.

7.4.1 E-Step

The z_i's are computed in a similar way to that explained in section 4.5.1. Hence

$$\tilde{z}_i = E[z_i | y_i, \eta_i, \sigma^2] = \frac{\lambda f(y_i)}{\lambda f(y_i) + (1-\lambda) g(y_i)} \tag{7.10}$$

where $\eta_i = x_i \beta$ is a linear predictor. Sometimes we use the notation $\tilde{z}_i = z(y_i, \eta_i, \sigma^2)$.

7.4.2 M-Step

The M-step is to maximize the complete likelihood function $L_c(\beta)$, with z_i replaced by $\tilde{z}_i$ $i = 1, \ldots n$.

$$\begin{aligned} l_c &= \log L_c(\beta) \\ &= \sum_{i=1}^{n} \hat{z}_i \log f(y_i) + \text{constant} \\ &= \sum_{i=1}^{n} \hat{z}_i \left[-\frac{1}{2} \left(\frac{y_i - x_i\beta}{\sigma} \right)^2 - \frac{1}{2} \log \sigma^2 \right] + \text{constant} \end{aligned}$$

A necessary condition for $\tilde{\beta}$ is

$$\sum_{i=1}^{n} \hat{z}_i \left(\frac{y_i - x_i\beta}{\sigma} \right) x_{ij} = 0 \ \forall j \tag{7.11}$$

A necessary condition for $\tilde{\sigma}^2$ is

$$\sum_{i=1}^{n} \hat{z}_i \left[\frac{(y_i - x_i\beta)^2}{2\,\sigma^4} - \frac{1}{2\,\sigma^2} \right] = 0 \tag{7.12}$$

In matrix form, we can write

$$\beta = (X^t Z X)^{-1} X^t Z y \tag{7.13}$$

and

$$\sigma^2 = (\mathbf{1}^t Z \mathbf{1})^{-1} (y - X\beta)^t Z (y - X\beta) \tag{7.14}$$

where Z is a diagonal matrix whose elements are $\tilde{z}_i$ $i = 1 \ldots n$. It is similar to the weighted form explained in section 7.2, but with the matrix Λ replaced by Z.

It is an iterative process algorithm. For a given starting value for β and σ^2, say $\tilde{\beta}^{(0)}$ and $\tilde{\sigma}^{2^{(0)}}$, we can immediately compute the elements of the matrix $Z^{(0)}$. Therefore, the new estimates $\tilde{\beta}^{(1)}$ and $\tilde{\sigma}^{2^{(1)}}$ can be derived using the fixed point equations (7.13) and (7.14). At the m^{th} iterative stage, we can write,

$$\tilde{\beta}^{(m+1)} = (X^t Z^{(m)} X)^{-1} X^t Z^{(m)} y$$

$$\tilde{\sigma}^{2^{(m+1)}} = (\mathbf{1}^t Z^{(m)} \mathbf{1})^{-1} (y - X\beta^{(m+1)})^t Z^{(m)} (y - X\beta^{(m+1)})$$

The limit of the sequence $\{\tilde{\beta}^{(m)}\}_{m=0}^{\infty}$ and $\{\tilde{\sigma}^{2^{(m)}}\}_{m=0}^{\infty}$ if they converge, are the estimates for β and σ^2. That is,

$$\tilde{\beta} = \lim_{m \to \infty} \tilde{\beta}^{(m)}$$

and

$$\tilde{\sigma}^2 = \lim_{m \to \infty} \tilde{\sigma}^{2^{(m)}}$$

That is, the estimate $\tilde{\phi}$ for ϕ is given below.

$$\tilde{\phi} = (\tilde{\beta}, \tilde{\sigma}^2)^t$$

7.5 Influence Function for $\tilde{\phi}$

This section is more complicated, because the updating function $h(\phi, F_n)$ has two components. That is $h = (h_1, h_2)^t$, where

$$h_1(\phi, F) = (X^t Z X)^{-1} X^t Z y \tag{7.15}$$

$$h_2(\phi, F) = (\mathbf{1}^t Z \mathbf{1})^{-1} (y - X\beta)^t Z (y - X\beta) \tag{7.16}$$

7.5.1 One - Step Influence Function for $\tilde{\phi}$

Here the one-step influence function has two components, one for $\tilde{\beta}$ and the other for $\tilde{\sigma}^2$. The first one can be obtained directly from (5.21). That is,

$$IF^1_{\tilde{\beta}}((x_0, y_0), F_n) = n\hat{z}(X^t ZX)^{-1} x_0^t (y_0 - x_0\tilde{\beta}) \tag{7.17}$$

Next, we compute the one-step influence function for $\tilde{\sigma}^2$ from the definition of influence function. Let $\breve{\sigma}^2(F_n)$ be a statistical functional defined in (7.18)

$$\breve{\sigma}^2(F_n) = \frac{\sum_{i=1}^n z_i (y_i - x_i\tilde{\beta})^2}{\sum_{i=1}^n z_i} \tag{7.18}$$

where F_n is the empirical distribution which places mass $\frac{1}{n}$ at the n points $\{(x_1, y_1), \ldots (x_n, y_n)\}$. Now consider the perturbed data set $(x_1, y_1), \ldots (x_n, y_n), (x_0, y_0)$ with weights $\frac{1-\epsilon}{n}, \ldots, \frac{1-\epsilon}{n}, \epsilon$. The new estimate $\breve{\sigma}^2_{new}$ is defined below.

$$\breve{\sigma}^2_{new} = \breve{\sigma}^2((1-\epsilon)F_n + \epsilon\Delta_{(x_0,y_0)}) \tag{7.19}$$

From (7.18)

$$\breve{\sigma}^2_{new} = \frac{\frac{1-\epsilon}{n}\sum_{i=1}^n z_i (y_i - x_i\tilde{\beta})^2 + \epsilon(y_0 - x_0\tilde{\beta})^2}{\frac{1-\epsilon}{n}\sum_{i=1}^n z_i + \epsilon z_0}$$

$$\begin{aligned}
\frac{\breve{\sigma}^2_{new} - \breve{\sigma}^2(F_n)}{\epsilon} &= \frac{[\sum_{i=1}^n z_i] z_0 (y_0 - x_0\tilde{\beta})^2 - z_0 \sum_{i=1}^n z_i (y_i - x_i\tilde{\beta})^2}{[\sum_{i=1}^n z_i][\frac{1-\epsilon}{n}\sum_{i=1}^n z_i + \epsilon z_0]} \\
&= \frac{z_0(y_0 - x_0\tilde{\beta})^2 - z_0\tilde{\sigma}^2}{\frac{1-\epsilon}{n}\sum_{i=1}^n z_i + \epsilon z_0} \\
\lim_{\epsilon\to 0}\frac{\breve{\sigma}^2_{new} - \breve{\sigma}^2(F_n)}{\epsilon} &= \frac{nz_0[(y_0 - x_0\tilde{\beta})^2 - \tilde{\sigma}^2]}{\sum_{i=1}^n z_i} \\
IF_{\breve{\sigma}^2}((x_0, y_0), F_n) &= \frac{nz_0[(y_0 - x_0\tilde{\beta})^2 - \tilde{\sigma}^2]}{\sum_{i=1}^n z_i}
\end{aligned}$$

We call $IF_{\breve{\sigma}^2}((x_0, y_0), F_n)$ the one step influence function for $\tilde{\sigma}^2$ and denote it as $IF^1_{\tilde{\sigma}^2}((x_0, y_0), F_n)$.

7.5.2 Jacobian Matrix

Once again, another complicated Jacobian matrix, J, is to be computed. The dimension of the Jacobian matrix is $(p+1)$ by $(p+1)$. For mathematical simplification, we partition the matrix J as follows:

$$J = \begin{bmatrix} J_1 & J_2 \\ J_3 & J_4 \end{bmatrix}$$

where J_1 is a $(p \times p)$ matrix, J_2 is a $(p \times 1)$ column vector, J_3 is a $(1 \times p)$ row vector and J_4 is a scalar. J_1 is similar to the (5.22). That is

$$J_1 = \frac{\partial h_1(\phi, F_n)}{\partial \beta} = (X^t Z X)^{-1} X^t V X \tag{7.20}$$

$$J_2 = \frac{\partial h_1(\phi, F_n)}{\partial \sigma^2} = \begin{bmatrix} \frac{\partial \beta_1}{\partial \sigma^2} \\ . \\ \\ . \\ \frac{\partial \beta_p}{\partial \sigma^2} \end{bmatrix} \tag{7.21}$$

$$J_3 = \frac{\partial h_2(\phi, F_n)}{\partial \beta} = \begin{bmatrix} \frac{\partial h_2(\phi, F_n)}{\partial \beta_1} & . & . & . & . & \frac{\partial h_2(\phi, F_n)}{\partial \beta_p} \end{bmatrix} \tag{7.22}$$

$$J_4 = \frac{\partial h_2(\phi, F_n)}{\partial \sigma^2} \tag{7.23}$$

From (7.15), we can write

$$(X^t Z X) h_1(\phi, F_n) = X^t Z y$$

Differentiate both sides with respect to σ^2

$$\begin{aligned} X^t \frac{\partial Z}{\partial \sigma^2} X h_1(\phi, F_n) + (X^t Z X) \frac{\partial h_1(\phi, F_n)}{\partial \sigma^2} &= X^t \frac{\partial Z}{\partial \sigma^2} y \\ (X^t Z X) \frac{\partial h_1(\phi, F_n)}{\partial \sigma^2} &= X^t \frac{\partial Z}{\partial \sigma^2} (y - X h_1(\phi, F_n)) \\ J_2 = \left[\frac{\partial h_1(\phi, F_n)}{\partial \sigma^2} \right]_{\phi = \tilde{\phi}} &= (X^t Z X)^{-1} X^t \frac{\partial Z}{\partial \sigma^2} (y - X \tilde{\beta}) \end{aligned}$$

Let $R = y - X\beta$ be a column vector. From (7.16), we can write

$$(\mathbf{1}^t Z \mathbf{1}) h_2(\phi, F_n) = R^t Z R$$

Differentiate both sides with respect to β_j

$$\begin{aligned}
(\mathbf{1}^t \frac{\partial Z}{\partial \beta_j} \mathbf{1}) h_2(\phi, F_n) &+ (\mathbf{1}^t Z \mathbf{1}) \frac{\partial h_2(\phi, F_n)}{\partial \beta_j} \\
&= -x_{.j}^t Z R + R^t \frac{\partial Z}{\partial \beta_j} R + R^t Z(-x_{.j}) \\
(\mathbf{1}^t Z \mathbf{1}) \frac{\partial h_2(\phi, F_n)}{\partial \beta_j} &= R^t \frac{\partial Z}{\partial \beta_j} R - (\mathbf{1}^t \frac{\partial Z}{\partial \beta_j} \mathbf{1}) h_2(\phi, F_n) \\
&\quad -2R^t Z x_{.j} \\
&= h_2(\phi, F_n) \mathbf{1}^t Z' x_{.j} - (R^2)^t Z' x_{.j} - 2R^t Z x_{.j} \\
&= (h_2(\phi, F_n) \mathbf{1}^t Z' - (R^2)^t Z' - 2R^t Z) x_{.j} \\
\frac{\partial h_2(\phi, F_n)}{\partial \beta_j} &= (\mathbf{1}^t Z \mathbf{1})^{-1} \gamma x_{.j}
\end{aligned}$$

where $\gamma = h_2(\phi, F_n) \mathbf{1}^t Z' - (R^2)^t Z' - 2R^t Z$ is a row vector, $x_{.j}$ is the j^{th} column of the matrix X and $Z' = \frac{\partial Z}{\partial \eta}$.

$$\begin{aligned}
\frac{\partial h_2(\phi, F_n)}{\partial \beta} &= \left(\frac{\partial h_2(\phi, F_n)}{\partial \beta_1}, \ldots, \frac{\partial h_2(\phi, F_n)}{\partial \beta_p} \right) \\
&= (\mathbf{1}^t Z \mathbf{1})^{-1} \gamma X \\
J_3 = \left[\frac{\partial h_2(\phi, F_n)}{\partial \beta} \right]_{\phi = \tilde{\phi}} &= (\mathbf{1}^t Z \mathbf{1})^{-1} \tilde{\gamma} X
\end{aligned}$$

where $\tilde{\gamma} = \tilde{\sigma}^2 \mathbf{1}^t Z' - (\tilde{R}^2)^t Z' - 2\tilde{R}^t Z$ and $\tilde{R} = y - X\tilde{\beta}$.

$$(\mathbf{1}^t Z \mathbf{1}) h_2(\phi, F_n) = R^t Z R$$

Differentiate both sides with respect to σ^2

$$\begin{aligned}(\mathbf{1}^t\frac{\partial Z}{\partial\sigma^2}\mathbf{1})h_2(\phi,F_n)+(\mathbf{1}^tZ\mathbf{1})\frac{\partial h_2(\phi,F_n)}{\partial\sigma^2} &= R^t\frac{\partial Z}{\partial\sigma^2}R\\ (\mathbf{1}^tZ\mathbf{1})\frac{\partial h_2(\phi,F_n)}{\partial\sigma^2} &= R^t\frac{\partial Z}{\partial\sigma^2}R-(\mathbf{1}^t\frac{\partial Z}{\partial\sigma^2}\mathbf{1})h_2(\phi,F_n)\end{aligned}$$

$$\begin{aligned}J_4 &= \left[\frac{\partial h_2(\phi,F_n)}{\partial\sigma^2}\right]_{\phi=\tilde{\phi}}\\ &= (\mathbf{1}^tZ\mathbf{1})^{-1}[(y-X\tilde{\beta})^t\frac{\partial Z}{\partial\sigma^2}(y-X\tilde{\beta})-\mathbf{1}^t\frac{\partial Z}{\partial\sigma^2}\mathbf{1}\tilde{\sigma}^2]\\ &= \frac{\sum_{i=1}^n[(y_i-x_i\tilde{\beta})^2-\tilde{\sigma}^2]\frac{\partial z_i}{\partial\sigma^2}}{\sum_{i=1}^n z_i}\end{aligned}$$

where

$$\begin{aligned}z_i=z(y_i,\eta_i,\sigma^2) &= \frac{\lambda f(y_i,\eta_i,\sigma^2)}{\lambda f(y_i,\eta_i,\sigma^2)+(1-\lambda)g(y_i)}\\ \left[\lambda f(y_i,\eta_i,\sigma^2)+(1-\lambda)g(y_i)\right]z_i &= \lambda f(y_i,\eta_i,\sigma^2)\\ \frac{\partial z_i}{\partial\sigma^2} &= \frac{z_i(1-z_i)}{f(y_i,\eta_i,\sigma^2)}\frac{\partial f(y_i,\eta_i,\sigma^2)}{\partial\sigma^2}\end{aligned}$$

$$\begin{aligned}\log f(y_i,\eta_i,\sigma^2) &= -\log 2\pi-\frac{1}{2}\log\sigma^2-\frac{1}{2}\left(\frac{y_i-\eta_i}{\sigma}\right)^2\\ \frac{1}{f(y_i,\eta_i,\sigma^2)}\frac{\partial f(y_i,\eta_i,\sigma^2)}{\partial\sigma^2} &= \frac{1}{2\sigma^4}[(y_i-\eta_i)^2-\sigma^2]\end{aligned}$$

Hence

$$\frac{\partial z_i}{\partial\sigma^2}=\frac{z_i(1-z_i)}{2\sigma^4}[(y_i-\eta_i)^2-\sigma^2]$$

Table 7.1: Regression and scale estimates for the Belgium phone call data

	MLE	**MLE$_{-6}$**	$\boldsymbol{\lambda = 0.75}$	$\boldsymbol{\lambda = 0.90}$	$\boldsymbol{\lambda = 0.95}$
Estimate for $\boldsymbol{\beta_0}$	- 260.06	-63.48	- 51.66	-52.44	-52.54
Estimate for $\boldsymbol{\beta_1}$	5.04	1.30	1.08	1.10	1.10
Estimate for $\boldsymbol{\sigma}$	56.22	4.40	0.91	1.31	1.35

7.6 Numerical Results

In this section, Example 6.1 is used to illustrate our method numerically. The regression parameters and scale parameter are computed together, and results will be given in the Table 7.1. The same g, defined in section 5.8, is used here. As usual, the estimates for β and σ are computed for various $\lambda = 0.75, 0.9, 0.95$.

The estimates $\tilde{\sigma}$, $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are almost the same for the various λ. This means λ has little impact on the estimates. However, the mixture estimates $\tilde{\sigma}$, $\tilde{\beta}_0$ and $\tilde{\beta}_1$ heavily deviate from the maximum likelihood estimates $\hat{\sigma}$, $\hat{\beta}_0$ and $\hat{\beta}_1$, are very close to the MLE$_{-6}$ estimates, which are the MLE after deleting the six outlier set of observations from the original data set. These fits can be viewed in Figure 7.1. In particular, the case of $\lambda = 0.95$ is considered for the mixture model.

Next, we would like to have the influence functions for these estimates. Since three parameters were estimated, the dimension of the Jacobian matrix J is 3×3. In fact, the dimensions of $J1, J2, J3$ and $J4$ are $2 \times 2, 2 \times 1, 1 \times 2$ and 1×1 respectively. The result of the Jacobian matrix J for this case is given

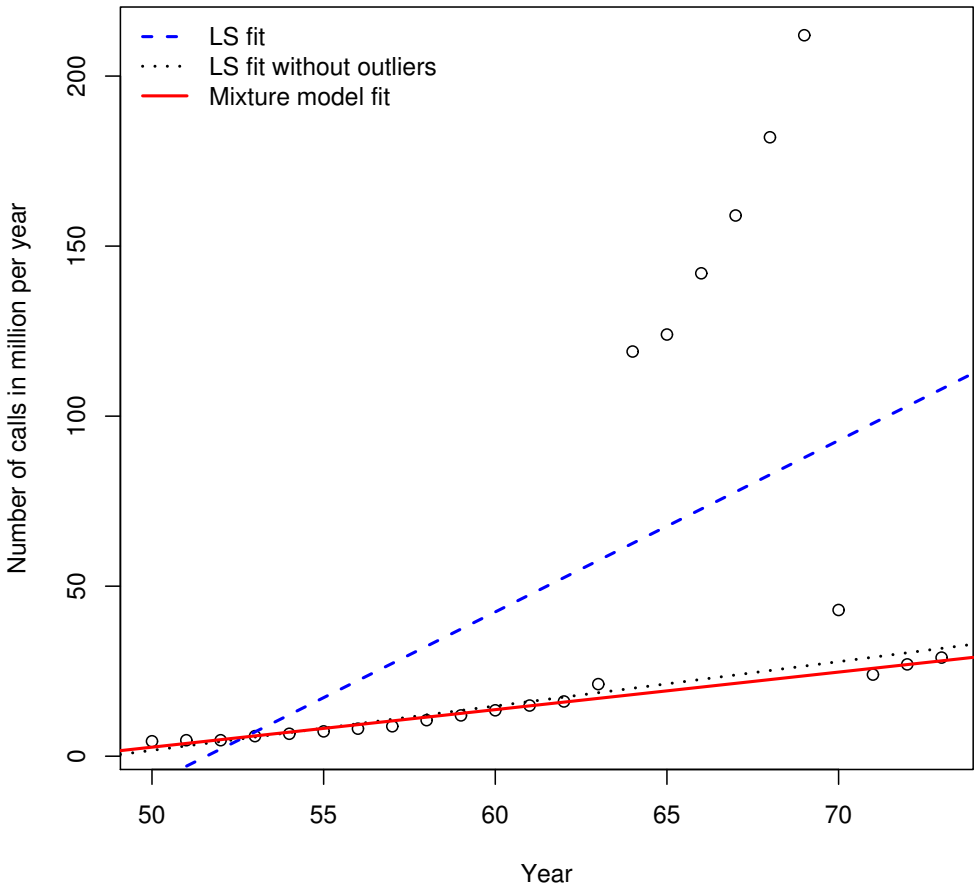

Figure 7.1: Number of international phone calls from Belgium in the years $1950 - 1973$ with LS fit,MLE_{-6} fit, and mixture model fit

Table 7.2: Regression and scale estimates are given by various methods using the Belgium phone call data

	Estimates for β_0	**Estimates for β_1**	**Estimates for σ**
MLE	- 260.059	5.041	53.83
Mixture	-52.542	1.104	1.35
Tukey	-52.302	1.098	1.65
Huber	-102.622	2.041	9.03

below when $\lambda = 0.95$.

$$J = \begin{bmatrix} -0.1334329 & -8.4569541 & -0.1372100 \\ 0.00288651 & 0.18267840 & 0.00296296 \\ -0.1439740 & -9.0697582 & 0.14398012 \end{bmatrix}$$

The true and one-step influence functions for the $\tilde{\beta}_0$, $\tilde{\beta}_1$ and $\tilde{\sigma}^2$ are given in the Figures 7.2 and 7.3 when $x = 70$ and $x = 100$ respectively. Using these figures, it can be seen that the estimates $\tilde{\beta}_0$, $\tilde{\beta}_0$ and $\tilde{\sigma}^2$ are insensitive to extreme values.

7.6.1 Comparison of $\tilde{\phi}$ with Standard Robust Estimates

Like other chapters, the estimates $\tilde{\beta}_0$, $\tilde{\beta}_1$ and $\tilde{\sigma}$ are compared with the traditional estimates obtained by Huber and Tukey methods, where MAD is computed as the scale estimate. Results are given in Table 7.2. Results based on the our method and on the Tukey method are very similar here too.

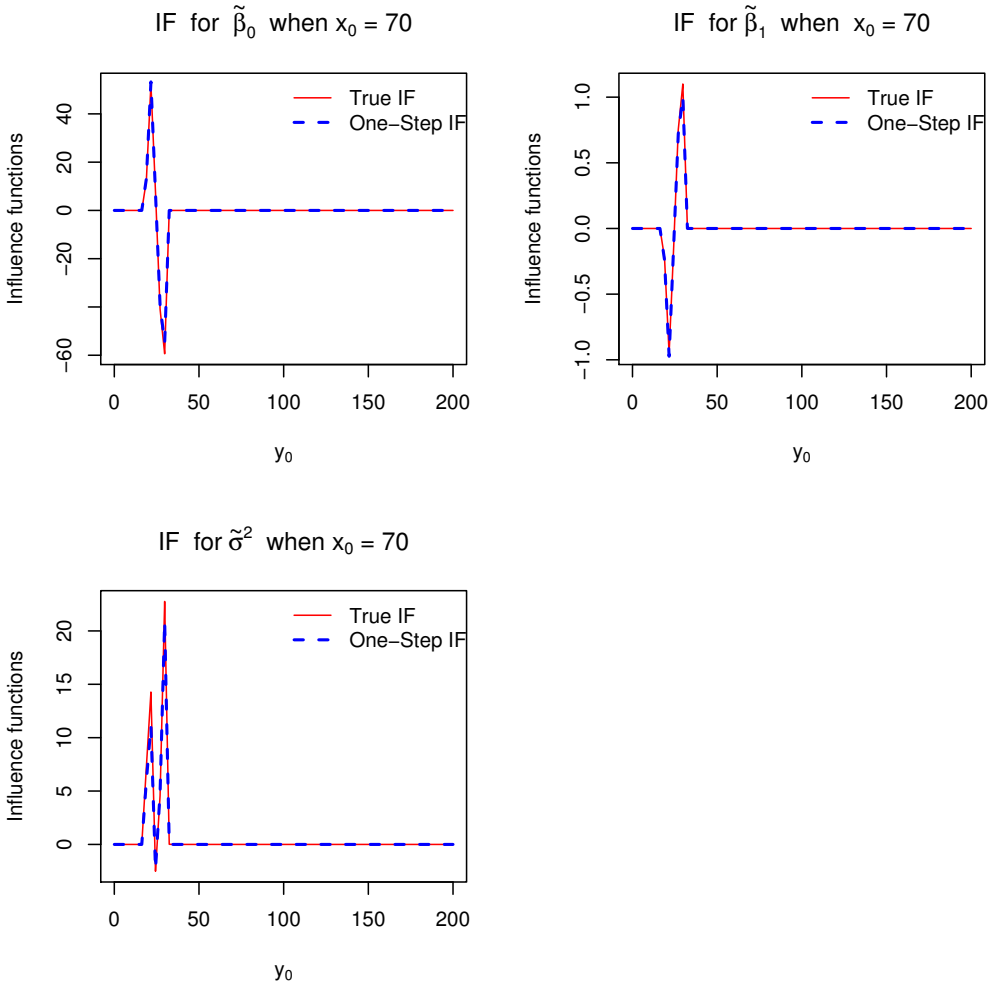

Figure 7.2: True and one-step influence functions for estimates when $x = 70$

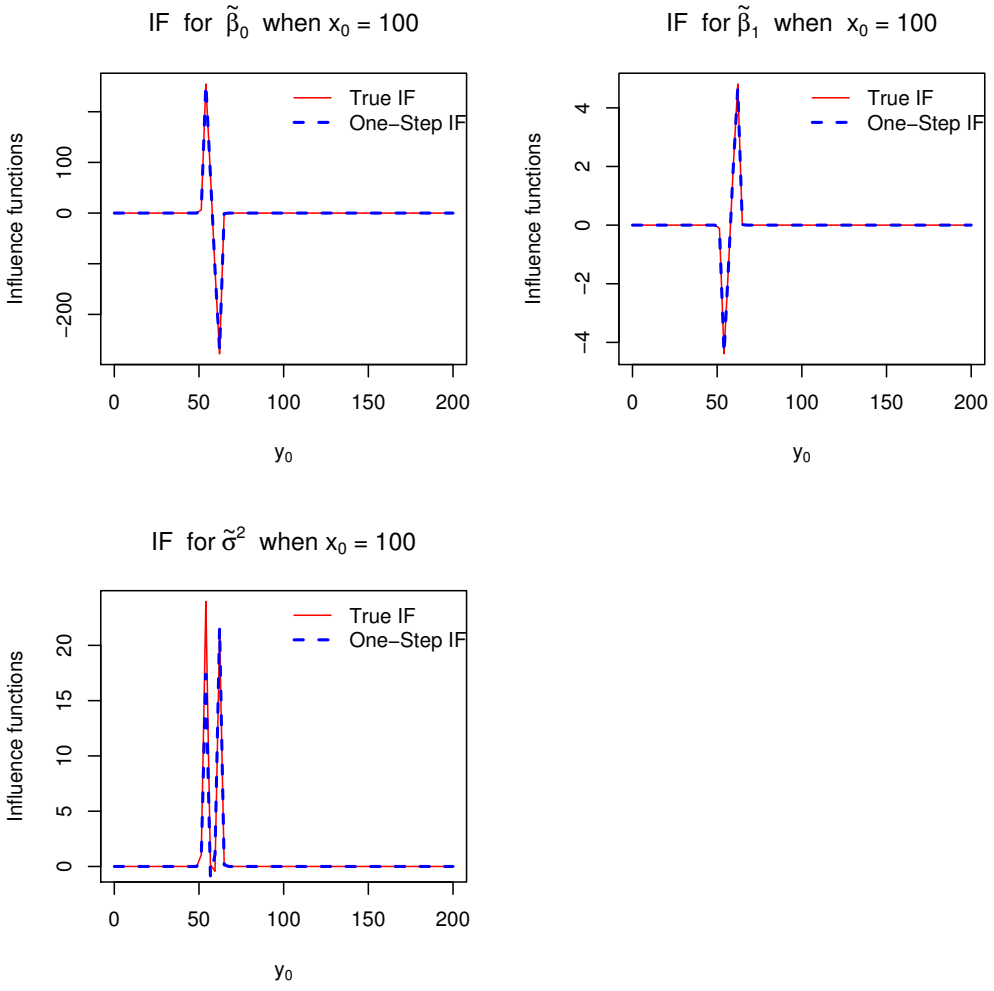

Figure 7.3: True and one-step influence functions for estimates when $x = 100$

Table 7.3: Location and scale estimates for the speed of light data

	MLE	**MLE$_{-2}$**	$\boldsymbol{\lambda = 0.90}$	$\boldsymbol{\lambda = 0.95}$	$\boldsymbol{\lambda = 0.97}$
Estimate for β	26.21	27.75	27.73	27.74	27.75
Estimate for σ	10.75 (4.45)	5.08 (5.19)	4.91	4.98	5.01

MAD values are given in the brackets.

7.7 Location and Scale

This section is a special case of this chapter. That is, we consider the same model of (7.1), where X is a column vector whose elements are 1. The β and σ are considered as location and scale parameters. Analytical results can be easily obtained from the sections 7.4 and 7.5 by replacing X by the column vector of 1, so that we are directly interested in numerical analysis. We use the data set of the speed of light, explained in Chapter 1, for this purpose.

For the mixture method, the g is defined as uniform distribution with parameters $a = -60$ and $b = 60$, because minimum and maximum value of the observations are -44 and 40. The results for MLE, MLE$_{-2}$ (compute the MLE after removing the two outliers), and our method for various λ such as 0.90, 0.95, and true $\lambda = 64/66 = 0.97$, are given in Table 7.3. The mixture model gives almost similar estimates for various λ, and these are very close to the MLE$_{-2}$. In addition, the mixture model estimates are definitely better than MLE.

Next, we are computing the Jacobian matrix J in order to compute the true influence functions of $\tilde{\beta} = 27.74$ and $\tilde{\sigma} = 4.98$. In this case, J is a 2 by 2 matrix and all $J1, J2, J3$ and $J4$ are single terms.

$$J_1 = \frac{\sum_{i=1}^{n} z_i'(y_i - \tilde{\beta})}{\sum_{i=1}^{n} z_i}$$

Table 7.4: Location and scale estimates are computed by various methods for speed of light data

	Estimates for $\boldsymbol{\beta}$	**Estimates for $\boldsymbol{\sigma}$**
MLE	26.21	10.75
Mixture	27.74	4.98
Tukey	27.67	5.19
Huber	27.39	5.03

$$J_2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta})\frac{\partial z_i}{\partial \sigma^2}}{\sum_{i=1}^{n} z_i}$$

$$J_3 = \frac{\tilde{\sigma}^2 \sum_{i=1}^{n} z_i' - \sum_{i=1}^{n} z_i'(y_i - \tilde{\beta})^2 - 2\sum_{i=1}^{n} z_i(y_i - \tilde{\beta})}{\sum_{i=1}^{n} z_i}$$

$$J_4 = \frac{\sum_{i=1}^{n}[(y_i - \tilde{\beta})^2 - \tilde{\sigma}^2]\frac{\partial z_i}{\partial \sigma^2}}{\sum_{i=1}^{n} z_i}$$

where $z_i' = \frac{\partial z_i}{\partial \beta}$. The numerical result for J is given below

$$J = \begin{bmatrix} 0.037304 & 0.000543 \\ \\ -0.062959 & 0.057262 \end{bmatrix}$$

Figures 7.4 and 7.5 give the true and one-step influence functions for the estimates $\tilde{\beta} = 27.74$ and $\tilde{\sigma} = 4.98$ respectively. The influence functions are bounded, which means estimates obtained by the mixture model are not unduly affected by outliers.

Next we compare these results with standard robust methods. The results are given in Table 7.4. Location and scale estimates are computed simultaneously for all these methods. We have here chosen MAD as the scale estimate

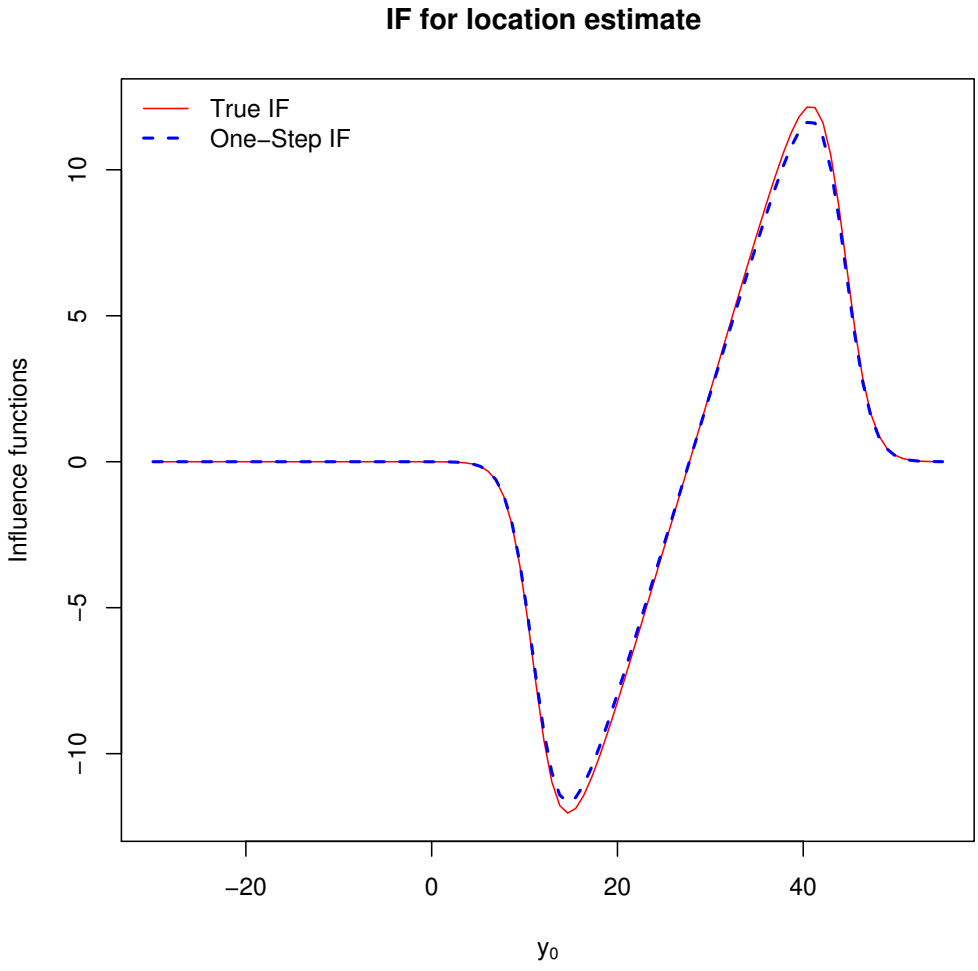

Figure 7.4: True and one-step influence functions for the location estimate

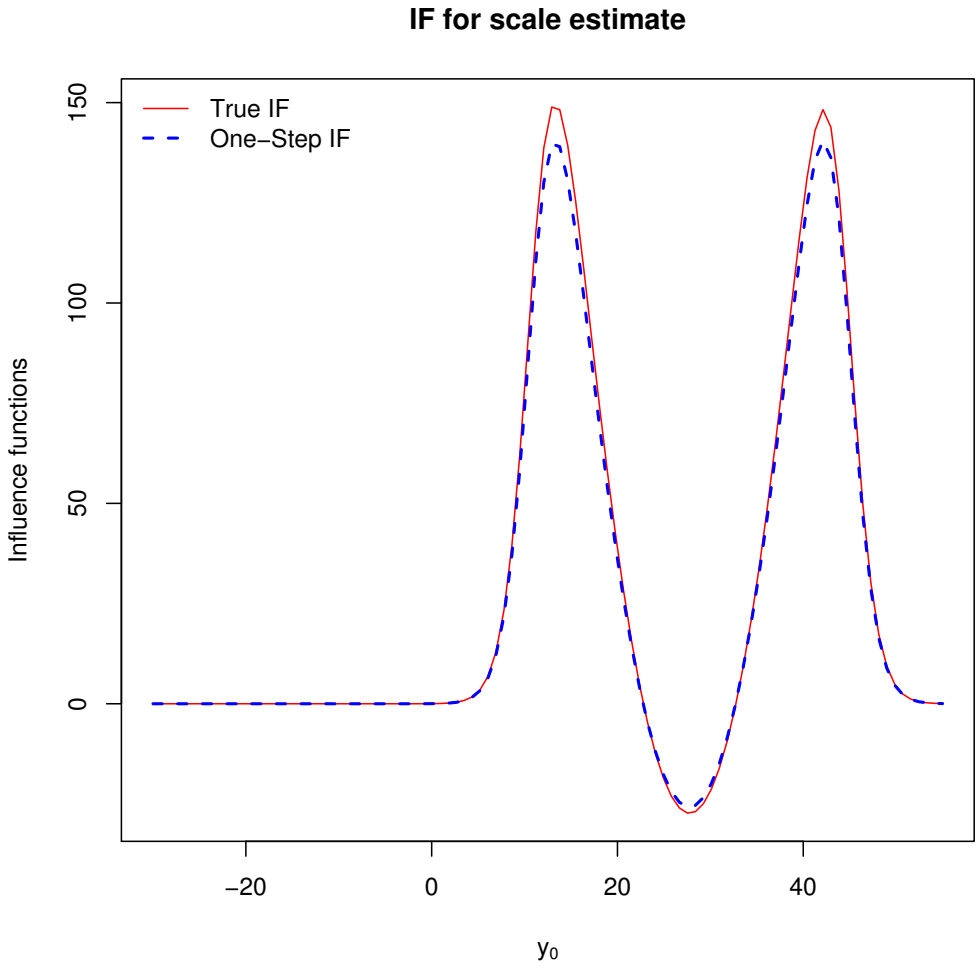

Figure 7.5: True and one-step influence functions for the scale estimate

for the cases of both the Tukey and Huber methods. Once again we found that mixture estimates and Tukey estimates are very similar.

Chapter 8

Robust Estimation for Non-Linear Models

The computation of robust statistics for non-linear model parameters is very limited in the robust literature. However the function 'nlrob' in the **R** library 'robustbase' fits non-linear regression using iteratively re-weighted least squares (IRWLS) method. In this Chapter, we are trying to apply our method to non-linear models.

8.1 The Model

The nonlinear model can be written as

$$y = \mu(\beta) + \epsilon \tag{8.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. It is similar form of (5.1), but $\mu(\beta)$ is not a linear combinations of covariates.

The definition of nonlinearity relates to the prediction equation, which form is nonlinearity on one or more unknown parameters. Note that, it is not the relationship between the response variable and the covariates. For example, the prediction equation of the form $\mu(\beta) = \frac{\beta_1 x}{\beta_2 + x}$ is considered non-linear model and the prediction equation of the form $\mu(\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ is still considered as linear model.

In this Chapter, we will investigate the Michaelis-Menten model (8.2) using treated Puromycin data, given in Appendix in section 8.6. Please refer [4] for further detail of the model,

$$\mu(\beta) = \frac{\beta_1 \, x}{\beta_2 + x} \tag{8.2}$$

where x is substrate concentration in an enzymatic and y is the reaction rates. The parameter $\beta = (\beta_1, \beta_2)$ is to be estimated. Figure 8.1 shows that the relationship between x and y is non-linear, and outliers are absent.

8.2 Calculating the Estimator

The estimate $\hat{\beta}$ for β can be obtained by

$$\hat{\beta} = \arg\max_{\beta} l(\beta)$$

where $l(\beta)$ is the log-likelihood and defined below

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i, \beta) \tag{8.3}$$

where $f_Y(y, \beta)$ is a probability density function of random variable Y. Since the $f_{Y_i}(y_i, \beta)$ belongs to exponential family, (8.3) becomes

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} (y_i b(\theta) + c(\theta) + d(y_i)) \\ &= \sum_{i=1}^{n} l_i(\beta) \end{aligned}$$

Since we are interested to do this investigation based on the normal distribution, $\mu = b(\theta) = \frac{\theta}{\sigma^2}$, $c(\theta) = -\frac{\theta^2}{2\sigma^2}$ and $d(y_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{y_i}{2\sigma^2}$, and σ^2 is treated as either known or 1 for simplicity.

For maximization,

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = 0 \tag{8.4}$$

where

$$U(\beta) = \begin{bmatrix} U_1(\beta) \\ . \\ . \\ U_p(\beta) \end{bmatrix} = \begin{bmatrix} \frac{\partial l(\beta)}{\partial \beta_1} \\ . \\ . \\ \frac{\partial l(\beta)}{\partial \beta_p} \end{bmatrix}$$

Consider

$$\begin{aligned} U_j(\beta) = \frac{\partial l(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \theta} \frac{\partial \theta}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} [y_i b'(\theta) + c'(\theta)][1] \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\sigma^2} \right) \frac{\partial \mu_i}{\partial \beta_j} \end{aligned}$$

Hence

$$\begin{aligned} U(\beta) &= \begin{bmatrix} \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\sigma^2} \right) \frac{\partial \mu_i}{\partial \beta_1} \\ . \\ . \\ \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\sigma^2} \right) \frac{\partial \mu_i}{\partial \beta_p} \end{bmatrix} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & . & . & \frac{\partial \mu_n}{\partial \beta_1} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial \mu_1}{\partial \beta_p} & . & . & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix} \begin{pmatrix} y_1 - \mu_1 \\ . \\ . \\ y_n - \mu_n \end{pmatrix} \\ &= \frac{1}{\sigma^2} D^t e \end{aligned}$$

The matrix D is called gradient matrix. In the linear model the matrix D must be the design matrix X, because $\frac{\partial \mu(\beta, x_i)}{\partial \beta_j} = x_{ij}$.

$(p,q)^{\text{th}}$ element of the expected information matrix of U is defined by

$$\begin{aligned}\Sigma_{pq} &= E(U_p(\beta)U_q(\beta)) \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{n} E(y_i-\mu_i)^2\frac{\partial\mu_i}{\partial\beta_p}\frac{\partial\mu_i}{\partial\beta_q} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_p}\frac{\partial\mu_i}{\partial\beta_q}\end{aligned}$$

Let Σ is the expected information matrix of U.

$$\Sigma = \frac{1}{\sigma^2}\begin{pmatrix} \frac{\partial\mu_1}{\partial\beta_1} & . & . & \frac{\partial\mu_n}{\partial\beta_1} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial\mu_1}{\partial\beta_p} & . & . & \frac{\partial\mu_n}{\partial\beta_p} \end{pmatrix}\begin{pmatrix} \frac{\partial\mu_1}{\partial\beta_1} & . & . & \frac{\partial\mu_1}{\partial\beta_p} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial\mu_n}{\partial\beta_1} & . & . & \frac{\partial\mu_n}{\partial\beta_p} \end{pmatrix} = \frac{1}{\sigma^2}D^tD \qquad (8.5)$$

In order to solve (8.4), use Taylor expansion to $U(\beta)$ and apply first order approximation, we will get

$$\beta^{(m+1)} = \beta^{(m)} + (-U'(\beta))^{-1}U(\beta^{(m)})$$

where m denotes m^{th} of the iteration. The observed information $-U'(\beta)$ is usually replaced by the expected information Σ. Hence

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + \Sigma^{-1}U(\beta^{(m)}) \\ \Sigma\,\beta^{(m+1)} &= \Sigma\,\beta^{(m)} + U(\beta^{(m)}) \\ &= D^tD\beta^m + D^te = D^t(D\beta^m + e) = D^tr \\ \beta^{(m+1)} &= (D^tD)^{-1}D^tr\end{aligned}$$

where $r = D\beta + e$. The estimate $\hat{\beta}$ is given by

$$\hat{\beta} = \lim_{m\to\infty}\beta^{(m)}$$

Back to our example,

$$\mu' = \begin{pmatrix} \frac{\partial\mu}{\partial\beta_1} \\ \\ \frac{\partial\mu}{\partial\beta_2} \end{pmatrix} = \begin{pmatrix} \frac{x}{\beta_2+x} \\ \\ \frac{\beta_1\,x}{(\beta_2+x)^2} \end{pmatrix}$$

Table 8.1: Results of the iteration for the example

$\boldsymbol{m}$	**0**	**1**	**2**	**3**	**4**
$\boldsymbol{\beta_1}$	205.0000	213.0289	212.6034	212.6754	212.6830
$\boldsymbol{\beta_2}$	0.0800	0.0629	0.0640	0.0641	0.0641

Hence,

$$D = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} \\ . & . \\ & \\ . & . \\ \frac{\partial \mu_n}{\partial \beta_1} & \frac{\partial \mu_n}{\partial \beta_2} \end{pmatrix} = \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} \begin{pmatrix} \frac{x}{\beta_2 + x} & \frac{\beta_1 x}{(\beta_2 + x)^2} \end{pmatrix} = X(\mu')^t$$

Since it is an iterative process with starting values $\beta_1^{(0)} = 205$ and $\beta 2^{(0)} = 0.08$, which are recommended in [4]. In Table 8.1, we have listed the value of $\beta_1^{(m)}$ and $\beta_2^{(m)}$ for various values of m. Therefore, this method gives estimates of the parameters β_1 and β_2 of the model (8.2), where $\hat{\beta}_1 = 212.683$ and $\hat{\beta}_2 = 0.064$. Figure 8.1 shows the prediction line based on these estimates. It seems very good fit for the data.

Our intention is to develop robustify method for estimating the parameters in non-linear models, so that we are now making new example by replacing 8^{th} observation of the y-value by just arbitrary value 220 and rest of them are exactly same as treated Puromycin data set.

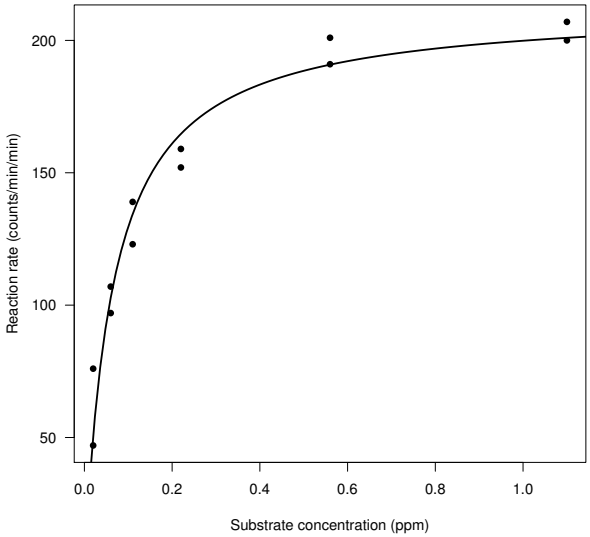

Figure 8.1: Treated Puromycin data with fitted Michaelis-Menten curves

8.3 The Mixture Model

We believe that robustify estimate for β of nonlinear models can be obtained using the mixture model, defined in (8.6).

$$p(y, X, \beta) = \lambda f(y, X, \beta) + (1 - \lambda)g(y) \tag{8.6}$$

where g is a known parameter free function over the sample space. Often $\lambda \approx 1$, but not exactly 1.

8.4 Calculating the Robustified Estimator

The robstified estimate $\tilde{\beta}$ may be defined as

$$\tilde{\beta} = \arg\max_{\beta} l_0 \tag{8.7}$$

where l_0 is observe log-likelihood and defined below

$$l_0 = \sum_{i=1}^{n} \log p(y_i, x_i, \beta)$$

After having mathematical simplification and algebraic operations, we will have complete log-likelihood, l_c,

$$l_c = \sum_{i=1}^{n} z_i l_i(\beta) + \text{constant} \tag{8.8}$$

where z_i' s are treated as missing values. It can be computed at the E-step of the EM algorithm.

$$\tilde{z}_i^{(m)} = \frac{\lambda f(y_i, x_i, \beta^{(m)})}{\lambda f(y_i, x_i, \beta^{(m)}) + (1-\lambda) g(y_i)}$$

Now (8.8) can be considered as a weighted form of the likelihood and new $U(\beta)$ and Σ are defined below.

$$U(\beta) = \begin{bmatrix} U_1(\beta) \\ . \\ . \\ U_p(\beta) \end{bmatrix} = \begin{bmatrix} \frac{\partial l_c(\beta)}{\partial \beta_1} \\ . \\ . \\ \frac{\partial l_c(\beta)}{\partial \beta_p} \end{bmatrix}$$

and

$$U_j(\beta) = \sum_{i=1}^{n} z_i \left(\frac{y_i - \mu_i}{\sigma^2} \right) \frac{\partial \mu_i}{\partial \beta_j}$$

$$\begin{aligned} U(\beta) &= \frac{1}{\sigma^2} \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & . & . & \frac{\partial \mu_n}{\partial \beta_1} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial \mu_1}{\partial \beta_p} & . & . & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix} \begin{pmatrix} z_i & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & . & . & z_n \end{pmatrix} \begin{pmatrix} y_1 - \mu_1 \\ . \\ . \\ y_n - \mu_n \end{pmatrix} \\ &= D^t Z e \end{aligned}$$

$$\Sigma_{pq} = \frac{1}{\sigma^2} \sum_{i=1}^{n} z_i^2 \frac{\partial \mu_i}{\partial \beta_p} \frac{\partial \mu_i}{\partial \beta_q}$$

$$\Sigma = \frac{1}{\sigma^2}\begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & . & . & \frac{\partial \mu_n}{\partial \beta_1} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial \mu_1}{\partial \beta_p} & . & . & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix} \begin{pmatrix} z_i & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & . & . & z_n \end{pmatrix} \begin{pmatrix} z_i & 0 & . & 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & . & . & z_n \end{pmatrix} \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & . & . & \frac{\partial \mu_1}{\partial \beta_p} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial \mu_n}{\partial \beta_1} & . & . & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix}$$

$$= \frac{1}{\sigma^2} D^t Z Z D$$

Hence the robustified estimate $\tilde{\beta}$ may be obtained by iterative process.

$$\begin{aligned} \beta^{(m+1)} &= \beta^m + \Sigma^{-1} U(\beta^m) \\ \Sigma\, \beta^{(m+1)} &= \Sigma\, \beta^{(m)} + U(\beta^{(m)}) \\ &= D^t Z Z D \beta^{(m)} + D^t Z e \\ &= D^t Z (Z D \beta^{(m)} + e) = D^t Z r \\ \beta^{(m+1)} &= (D^t Z Z D)^{-1} D^t Z r \end{aligned}$$

and

$$\tilde{\beta} = \lim_{m \to \infty} \beta^{(m)}$$

where $r = ZD\beta^{(m)} + e$.

8.5 Numerical Results

We use new set of data, explained in end of section 8.2 for this investigation. First, we compute the maximum likelihood estimator for the full data set and data set without 8^{th} observation. Later we fit the model for the full data set using mixture method. After investigate the data, we choose g as uniform distribution on $[40, 230]$, and $\lambda = 0.97$, and scale parameter is defined as $mad(y)$. Results are given in Table 8.2 and a number of fits based on various methods are given in Figure 8.2.

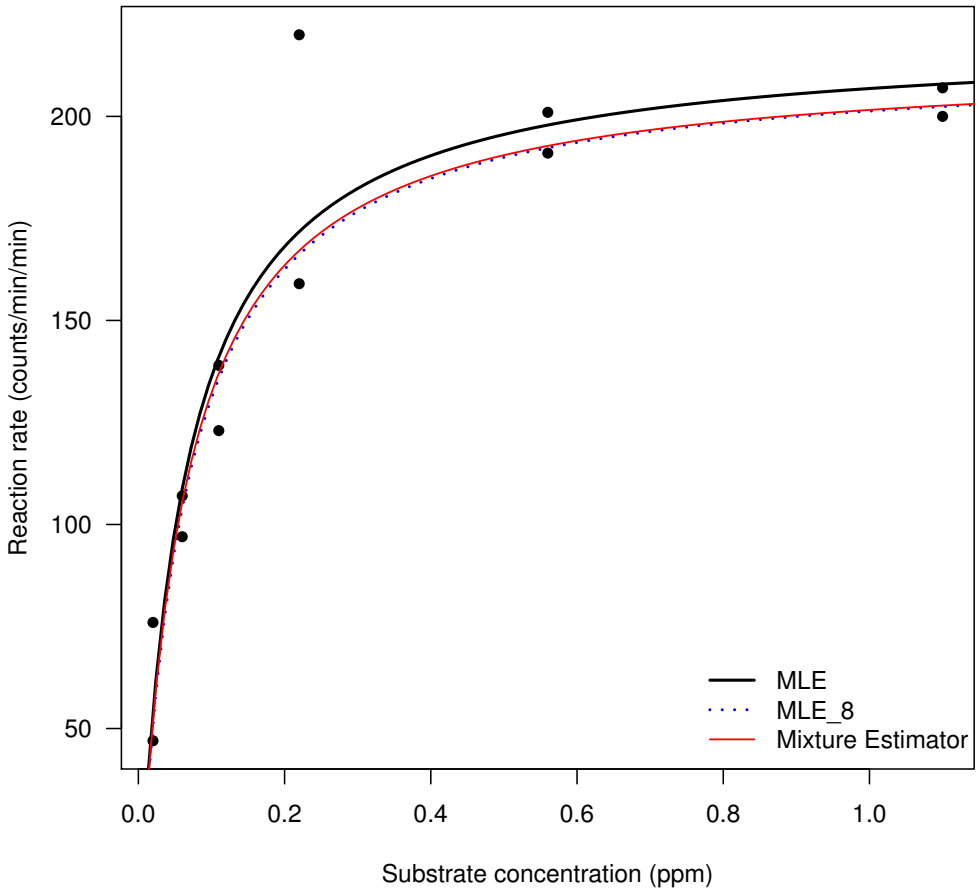

Figure 8.2: Treated Puromycin data with fitted Michaelis-Menten curves based on various methods

Table 8.2: Estimates based on various methods

	MLE	$\mathbf{MLE_{-8}}$	**Mixture**
$\boldsymbol{\beta_1}$	219.464	214.056	214.038
$\boldsymbol{\beta_2}$	0.061	0.063	0.0618

8.6 Appendix - Treated Puromycin Data

$\boldsymbol{x}$	$\boldsymbol{y}$
0.02	76
0.02	76
0.06	97
0.06	107
0.11	123
0.11	139
0.22	159
0.22	152
0.56	191
0.56	201
1.10	207
1.10	200

Chapter 9

Another Choice of g

Up to now, one of the mixture components g in our model was chosen as a probability density function. For example, g is assigned to as either uniform distribution or equal probability for success and failure in the previous chapters. We referred to this case as *proper* g. In this Chapter, we begin to define another form of g, which is not associated with distribution. This form of g is referred to as *improper density* g. The g may be defined as a constant. In this chapter, we are interested in estimating the parameters using the mixture model with a model density function f, which we refer to as a *regular component*, and an improper density function g.

The mathematical results given in the previous chapters will not be changed, because the derivations depend on g. But, in this chapter, we use particular g so that we have focused mainly on numerical results using previous examples.

9.1 Robust Location Estimate

This section illustrates the computation of robust statistics for the location parameter using the combination of proper and improper density functions. The example used here is generated data, explained in section 4.1.

We fit the mixture model with a model density function f and an improper component with the constant density $g = C_0$, which takes successively the values $0.005, 0.05, 0.1$ to estimate the parameters of the model density f when

Table 9.1: Location estimates (MLE and Mixture) for the generated data: $\hat{\theta}$ is a maximum likelihood estimator; $\tilde{\theta}_{0.95}$ is a mixture estimate when g is uniform distribution and $\lambda = 0.95$; and $\tilde{\theta}_{0.95,C_0}$ is an estimate for θ by mixture model with an improper constant density $g = C_0$ when $\lambda = 0.95$.

	$\hat{\boldsymbol{\theta}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95,0.005}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95,0.05}}$	$\tilde{\boldsymbol{\theta}}_{\mathbf{0.95,0.1}}$
Estimate for $\boldsymbol{\theta}$	8.7313	9.7854	9.7896	9.7564	9.7435

the non-contamination proportion $\lambda = 0.95$. The results are displayed in Table 9.1. In addition, MLE and the mixture estimate obtained from Table 4.1 are given in Table 9.1. The results obtained by the mixture models become more attractive, because they are very close to the true value of 10. When C_0 increases, the results deviate slightly from the true value. Particularly, $\tilde{\theta}_{0.95} \approx \tilde{\theta}_{0.95,0.005}$, because $\tilde{\theta}_{0.95}$ is obtained based on the uniform distribution on $[-80, 80]$ and proper density $g = 0.00625$, which is approximately closer to $C_0 = 0.005$.

9.2 Robust Regression Estimates

The data (Belgium Phone Call Data) for this example is described in section 5.1. We are trying to estimate the linear regression parameters using the mixtures of the proper density and the improper constant density $g = C_0$, which take the values $0.002, 0.05, 0.1$, when $\lambda = 0.95$. Results obtained by this model and other relevant results are displayed in Table 9.2.

The results obtained by all the mixture models are far better than MLE,as shown in Table 9.2. Figure 9.1 shows that the mixture model lines do fit the bulk of the data. It seems the results are not heavily affected by the choice of g. However, there is a small difference in the slope and intercept based on the comparison between the proper g and the improper g.

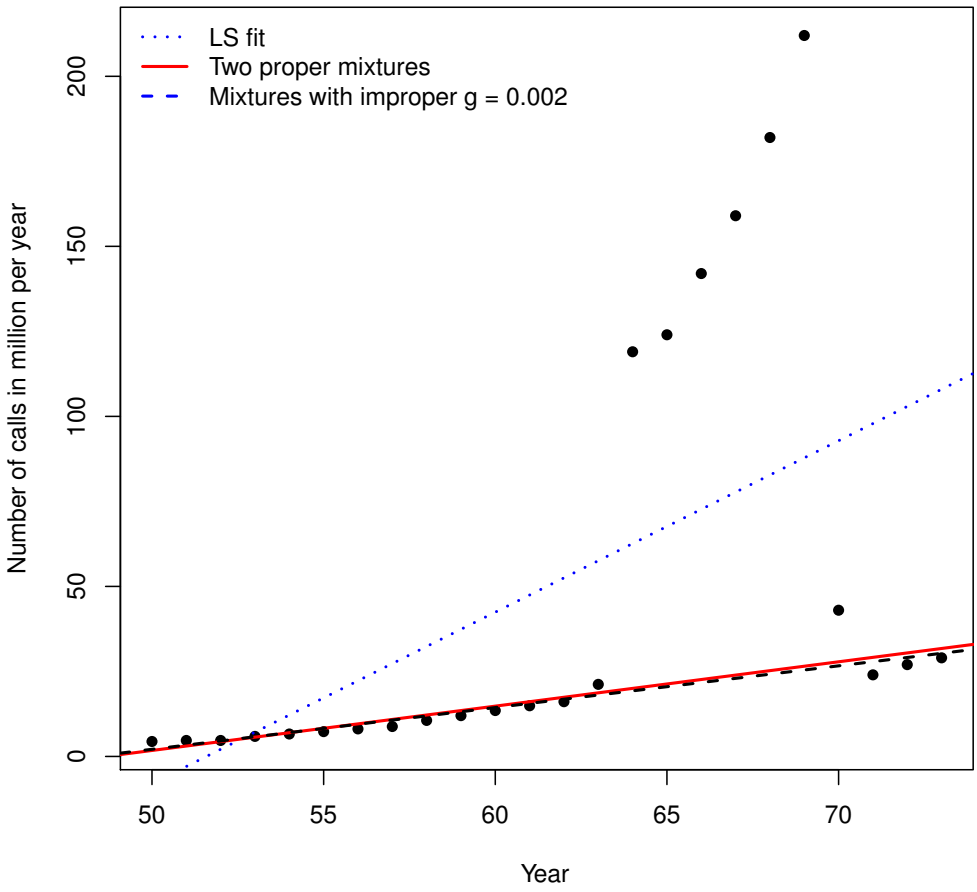

Figure 9.1: Number of international phone calls from Belgium in the years $1950 - 1973$ with LS fit and mixture model fits

Table 9.2: Regression estimates for the Belgium phone calls data when $\lambda = 0.95$

	MLE	g **is uniform on** $[0, 300]$	$g = 0.002$	$g = 0.05$	$g = 0.1$
Estimate for β_0	- 260.059	- 63.444	-59.243	-59.248	-59.252
Estimate for β_1	5.041	1.304	1.227	1.227	1.227

9.3 Robust Estimates for Poisson Regression

The data analyzed in this example is the number of voluntary associations to which respondents belonged in Quebec, Canada. More detail about the data is explained in section 6.9.2. We describe the response variable as a function of a number of explanatory variables such as gender, nationality and language skills under the Poisson model.

Robust estimates for the Poisson regression parameters may be obtained by considering the mixtures of the Poisson distribution and the improper g with the constant values $0.005, 0.05, 0.1$. These results and other relevant comparable results are given in Table 9.3.

Once again, the results obtained by our mixture models give better estimates than MLE. There is a very small difference in the estimates based on comparisons between the proper g and the improper g. All estimates values are increased as C_0 increases except the intercept, which is the reference category here. The estimates based on the proper g (uniform distribution on $[0, 25]$) and improper g ($g = 0.05$) are almost similar.

Table 9.3: Regression estimates for Quebec data when $\lambda = 0.95$

Estimate	**MLE**	**Mixture $g$is uniform on $[0, 25]$**	**Mixture $g = 0.005$**	**Mixture $g = 0.05$**	**Mixture $g = 0.1$**
$\boldsymbol{\beta_0}$ **(Intercept)**	0.586	0.030	0.244	0.006	-0.076
$\boldsymbol{\beta_1}$ **(Men)**	0.079	0.110	0.101	0.109	0.101
$\boldsymbol{\beta_2}$ **(Canadian--Born)**	0.027	0.356	0.242	0.365	0.394
$\boldsymbol{\beta_3}$ **(English)**	0.357	0.495	0.435	0.506	0.545
$\boldsymbol{\beta_4}$ **(Other)**	-0.014	0.102	0.034	0.107	0.116

Table 9.4: Location and scale estimates for the speed of light data when $\lambda = 0.95$

	MLE	**MLE$_{-2}$**	**Mixture g is uniform on $[-60, 60]$**	**Mixture $g = 0.005$**	**Mixture $g = 0.05$**	**Mixture $g = 0.1$**
$\boldsymbol{\beta}$	26.21	27.75	27.74	27.75	27.70	27.64
$\boldsymbol{\sigma}$	10.75 (4.45)	5.08 (5.19)	4.98	5.004	4.67	4.35

MAD values are given in brackets.

9.4 Robust Location and Scale Estimates

The speed of light data is considered as an example to explain this case. Descriptions of the data are given in Chapter 1. We start to compute the robust statistics for the location and the scale parameters together by using the mixtures of the regular component f and the constant improper component $g = C_0$, where C_0 is assumed to take the following constant values $0.005, 0.05, 0.1$. Results are given in Table 9.4 including other appropriate results . The location parameter is denoted as β instead of the standard notation of μ.

The estimate MLE$_{-2}$, denoted in Table 9.4, indicates the maximum likelihood estimator of the data after removing two outliers from the original data set. The results obtained from the mixture models are very close to the MLE$_{-2}$, especially when $g = 0.005$. In addition, the results in columns three, four and five of the Table 9.4 are very similar.

9.5 Selection of C_0

If we choose large value of C_0, then all points may be assigned to the outlier component g. In contrast, all points may be assigned to the regular component f if C_0 is very small. These situations are not desired here. Hence we may say that the choice C_0 is not simple.

In our previous examples, if the density in the proper g (uniform distribution) is approximately the same as the constant in the improper g, then the results are almost same. Therefore, the desired C_0 may be found by the following approach: find the uniform density as explained in previous chapters and assign this constant density value to C_0, then define the C_0 on a real line to make an improper g. In other words, the suitable value for the constant C_0 is the reciprocal of the range, which is the domain of the uniform distribution.

Next, we would like to find out the range of C_0 based on the simulation study to have possible good estimates for parameters. Here the range of C_0 is defined from $k \times 10^{-3}$ to $k \times 10^3$, where k may be obtained by our suggestion, given before.

Location

In section 4.7, the proper component g is defined as the uniform distribution on $[-80, 80]$, so the range for the interval is 160. Therefore, $k = \frac{1}{160} = 0.00625$ is chosen to make the improper g here. Now we investigate the estimate when the C_0 value changing from 0.00625×10^{-3} to 0.00625×10^3. Results are given in Table 9.5. The estimates do not change much in this range. Note that k is in three decimal places and estimates are also similar in the range of $k \times 10^{-3}$ to $k \times 10^3$.

Linear Regression Estimates

As in the previous section, we choose $k = 0.003$ here, because the g is defined on $[0, 300]$ in section 5.8. The regression estimates, which may be still considered as location estimates, are given in Table 9.6. Again they are very similar in the range of $k \times 10^{-3}$ to $k \times 10^3$ when k is in the three decimal places.

Table 9.5: Estimates for location based on the range of C_0 when $\lambda = 0.95$

C_0	**Estimates**
$\mathbf{0.00625 \times 10^{-3}}$	10.00844
$\mathbf{0.00625 \times 10^{-2}}$	9.89238
$\mathbf{0.00625 \times 10^{-1}}$	9.83508
$\mathbf{0.00625 \times 10^{0}}$	9.78538
$\mathbf{0.00625 \times 10^{1}}$	9.75259
$\mathbf{0.00625 \times 10^{2}}$	9.70584
$\mathbf{0.00625 \times 10^{3}}$	9.67726

Table 9.6: Estimates for regression parameters based on the range of C_0 when $\lambda = 0.95$

C_0	Estimates for β_0	Estimates for β_1
$\mathbf{0.003 \times 10^{-3}}$	-59.238208	1.226444
$\mathbf{0.003 \times 10^{-2}}$	-59.242456	1.226521
$\mathbf{0.003 \times 10^{-1}}$	-59.243077	1.226533
$\mathbf{0.003 \times 10^{0}}$	-59.243383	1.226538
$\mathbf{0.003 \times 10^{1}}$	-59.245678	1.226580
$\mathbf{0.003 \times 10^{2}}$	-59.277966	1.227171
$\mathbf{0.003 \times 10^{3}}$	-59.117012	1.224226

Generalized Linear Estimates

According to section 6.9.2, $k = 0.04$ is chosen. The estimates for the range are given in Table 9.7. It is noted that estimates become worse when the C_0 become two digit number, but estimates are also similar in the range of $k \times 10^{-1}$ to $k \times 10^{1}$.

Location and Scale Estimates

In this case, three decimal places $k = 0.008$ is chosen because the g is defined on $[-60, 60]$ in section 7.7. Results are given in Table 9.8. Note that location parameter and scale parameter are estimated together. Reasonable good estimates for location and scale are obtained in the range from 0.008×10^{-1} to 0.008×10^{1}.

9.6 Discussion

Overall, we may say that the good robust estimates may be obtained in the possible range of $[k \times 10^{-1}, k \times 10^{1}]$ when $0 < k < 1$. Choice of k may be critical, but defined by our suggestion. It works very well in our examples.

The mixture with a regular component and an improper component works very well in our examples. It may be interpreted as one of the members of the mixture models. In addition, convergence of the EM algorithm does not seem to present any problems. The estimates produced by the mixture models are far better than the maximum likelihood estimator. The choice of g may not have a heavy impact on the estimates because the results are quite similar in cases of proper g and improper g.

It is unlikely to have good estimates if we do not choose a proper g correctly. In this context, the mixture with regular component f, and constant improper component g, is a better solution for this situation. In our examples, the proper choices of g in both cases give good estimates for the parameters when data is contaminated.

Table 9.7: Estimates for Poisson regression parameters based on the range of C_0 when $\lambda = 0.95$

C_0	Estimates for β_0	Estimates for β_1	Estimates for β_2	Estimates for β_3	Estimates for β_4
$\mathbf{0.04 \times 10^{-3}}$	0.4982334	0.07500358	0.08356528	0.4039187	0.03899754
$\mathbf{0.04 \times 10^{-2}}$	0.4226115	0.07861811	0.1296699	0.4290372	0.03585749
$\mathbf{0.04 \times 10^{-1}}$	0.2652892	0.09858947	0.2288189	0.4323557	0.02944058
$\mathbf{0.04 \times 10^{0}}$	0.02999085	0.110274	0.3555329	0.4951099	0.1021160
$\mathbf{0.04 \times 10^{1}}$	-0.3032804	0.08054619	0.4523816	0.6569848	0.1239638
$\mathbf{0.04 \times 10^{2}}$	-0.8445834	0.002091187	0.3562236	0.9638495	-0.1347507
$\mathbf{0.035 \times 10^{3}}$	-0.581695	-1.128759	-13.64093	-0.1248210	-33.71002

Table 9.8: Estimates for location and scale paprameters based on the range of C_0 when $\lambda = 0.95$

C_0	Estimates for β	Estimates for σ
$\mathbf{0.008 \times 10^{-3}}$	27.74745	5.050765
$\mathbf{0.008 \times 10^{-2}}$	27.74969	5.043625
$\mathbf{0.008 \times 10^{-1}}$	27.74931	5.037400
$\mathbf{0.008 \times 10^{0}}$	27.74324	4.981607
$\mathbf{0.008 \times 10^{1}}$	27.66608	4.470561
$\mathbf{0.008 \times 10^{2}}$	27.11393	2.806708
$\mathbf{0.008 \times 10^{3}}$	27.28663	1.244484

Chapter 10

Mixture Estimates for General Case

In this chapter, we demonstrate the method of the thesis in a general case, where the estimator that we are trying to make robust is a maximum likelihood estimator of a multiparameter multivariate distribution $f(y, \theta)$. We consider the calculation of the mixture estimator and its influence function, and describe conditions under which the mixture estimator has a bounded influence function, or in other words, is B-robust.

10.1 Model

The model here is simply a distribution $f(y, \theta)$, which we will sometimes write in cumulative form $F(y, \theta)$. We will often find it useful to take F to be in the form of a finite sum of point masses

$$F_w(y) = \frac{\sum_{i=1}^n w_i I(Y_i \geq y)}{\sum_{i=1}^n w_i} \tag{10.1}$$

where $w = (w_1, \ldots, w_n)$ is a vector of non-negative weights and $I(Y_i \geq y)$ is an indicator function.

10.2 Calculating the Estimator

The estimator for given a sample $y_1, \ldots, y_n$ from F, is the maximum likelihood estimator $\hat{\theta}_M$ taking this as our starting point with values of θ maxi-

mizing the likelihood

$$L(\theta; Y) = \Pi_{i=1}^{n}[f(y_i, \theta)]$$

In the case of multiple maxima, we concentrate on local maxima. For later convenience, we will consider the maximum likelihood estimator for a weighted sample $\{(y_1, w_1), \ldots, (y_n, w_n)\}$, where $w_1, \ldots, w_n$ are positive real numbers. A weighted sample generates a weighted likelihood with the form

$$L(\theta; Y, w) = \Pi_{i=1}^{n}[f(y_i, \theta)]^{w_i} \tag{10.2}$$

The maximum likelihood estimator $\hat{\theta}$ for θ satisfies

$$\frac{\partial}{\partial \theta} l(\theta; Y, w) = 0 \tag{10.3}$$

where $l(\theta; Y, w) = \log L(\theta; Y, w) = \sum_{i=1}^{n} w_i \log f(y_i, \theta)$. From (10.3), we can write

$$\sum_{i=1}^{n} w_i \frac{1}{f(y_i, \theta)} \frac{\partial f(y_i, \theta)}{\partial \theta} = 0$$

$$\sum_{i=1}^{n} w_i u_i = 0 \tag{10.4}$$

where $u(\theta, y) = \frac{1}{f(y,\theta)} \frac{\partial f(y,\theta)}{\partial \theta}$. If $y = y_i$, we write $u_i = u_i(\theta) = u(\theta, y_i)$, and the score function $U(\theta) = \sum_{i=1}^{n} w_i u_i(\theta)$. In fact, $U(\hat{\theta}) = 0$.

10.3 Influence Function for $\hat{\theta}$

To compute the influence function, we begin with a contaminated distribution function $F_{w,\epsilon}$ defined in (10.5).

$$F_{w,\epsilon}(y) = (1 - \epsilon) F_w + \epsilon I_y \tag{10.5}$$

where I_y denotes the cumulative distribution function giving mass 1 to y. The new maximum likelihood estimator $\hat{\theta}^{\epsilon}$ based on $F_{w,\epsilon}$ satisfies

$$(1 - \epsilon) \sum_{i=1}^{n} w_i u_i(\hat{\theta}^{\epsilon}) + \epsilon \left[\sum_{i=1}^{n} w_i \right] u(\hat{\theta}^{\epsilon}) = 0 \tag{10.6}$$

analogously to (10.4). Based on the definition of the influence function (2.1), we can write

$$\hat{\theta}^{\epsilon} \approx \hat{\theta} + \epsilon IF_{\hat{\theta},F_w}(y)$$

Application of Taylor expansion will give

$$\begin{aligned} u_i(\hat{\theta}^{\epsilon}) &= u_i(\hat{\theta}) + \epsilon \left[u_i'(\theta)\right]_{\theta=\hat{\theta}} IF_{\hat{\theta},F_w}(y) \ + \cdots \\ &= u_i(\hat{\theta}) + \epsilon \left[u_i'(\theta)\right]_{\theta=\hat{\theta}} IF_{\hat{\theta},F_w}(y) \ + O(\epsilon^2) \end{aligned}$$

where $u_i'(\theta) = \frac{\partial u_i}{\partial \theta}$. Substitute this into (10.6),

$$\begin{aligned} (1-\epsilon)\sum_{i=1}^{n} w_i \left[u_i(\hat{\theta}) + \epsilon \left[u_i'(\theta)\right]_{\theta=\hat{\theta}} IF_{\hat{\theta},F_w}(y) \ + O(\epsilon^2)\right] & \\ + \epsilon \left[\sum_{i=1}^{n} w_i\right] \left[u(\hat{\theta}) + \epsilon \left[u'(\theta)\right]_{\theta=\hat{\theta}} IF_{\hat{\theta},F_w}(y) \ + O(\epsilon^2)\right] &= 0 \end{aligned}$$

Since $O(\epsilon^2) \to 0$ as $\epsilon \to 0$ and (10.4), we have

$$\begin{aligned} \sum_{i=1}^{n} w_i u_i' IF_{\hat{\theta},F_w}(y) + \sum_{i=1}^{n} w_i u &= 0 \\ \sum_{i=1}^{n} w_i^* u_i' IF_{\hat{\theta},F_w}(y) + u &= 0 \end{aligned}$$

where $w_i^* = \frac{w_i}{\sum_{i=1}^{n} w_i}$ and hence,

$$IF_{\hat{\theta},F_w}(y) = -\left[\sum_{i=1}^{n} w_i^* u_i'\right]^{-1}_{\theta=\hat{\theta}} u(\hat{\theta}, y) \tag{10.7}$$

The first part of the right hand side of (10.7) is constant in y and $u(\hat{\theta}, y)$ is often unbounded. For a simple example of the estimate of θ in $\mathcal{N}(\theta, \sigma^2)$ for known σ^2, $\hat{\theta} = \bar{y}$ and $u(\hat{\theta}, y) = \frac{y-\hat{\theta}}{\sigma^2}$ is unbounded. Hence the influence function for $\hat{\theta}$ is unbounded. Therefore, $\hat{\theta}$ is not, in general, B-robust.

10.4 Mixture Model

We will base our robustified estimate of θ on the mixture model given below in (10.8).

$$p(y, \theta) = \lambda f(y, \theta) + (1-\lambda) g(y) \tag{10.8}$$

where λ is an appropriately chosen fixed value such as 0.95, and $g()$ is a fixed, parameter free (possibly improper) density over the sample space of the model. In practice, $g(.)$ may be chosen with regard to subject area considerations and the type of bad data anticipated.

10.5 Calculating the Robustified Estimator $\tilde{\theta}$

The robustified estimator is taken to be the maximum likelihood estimator of θ in the mixture model (10.8), so θ is estimated by maximizing the observed likelihood function $L_o(\theta) = \Pi_{i=1}^n [p(y_i, \theta)]$. This is often complicated to maximize, so that the parameter θ is estimated using the EM algorithm. The complete likelihood function L_c is often considered.

$$L_c(\theta, Z) = \Pi_{i=1}^n [p(y_i, \theta)]^{z_i} \tag{10.9}$$

where $z_i = 1$ if y_i is sampled from f and $z_i = 0$ if it is sampled from g. For mathematical simplification, $l_c(\theta, Z) = \log L_c(\theta, Z)$ is considered and given in (10.10).

$$l_c(\theta, Z) = \sum_{i=1}^{n} z_i f(y_i, \theta) + \text{constant} \tag{10.10}$$

The expectation of $l_c(\theta, Z)$ for given y and θ is found in the E-step of the EM algorithm. As z_i enters linearly into l_c it suffices to calculate its expectation. If $\theta^{(m)}$ is the current value of θ, the expectation is

$$z_i^{(m)} = \frac{\lambda f(y_i, \theta^{(m)})}{\lambda f(y_i, \theta^{(m)}) + (1 - \lambda) g(y_i)} \tag{10.11}$$

At the M-step, an updated value $\theta^{(m+1)}$ of θ is calculated to satisfy

$$U(\theta^{(m+1)}) = \sum_{i=1}^{n} z_i^{(m)} u_i(\theta^{(m)} = 0 \tag{10.12}$$

If $\theta^{(m)} \to \tilde{\theta}$ as $m \to \infty$, then $\tilde{\theta}$ is considered as our robustified estimator for θ. We also define

$$\tilde{z}_i = \frac{\lambda f(y_i, \tilde{\theta})}{\lambda f(y_i, \tilde{\theta}) + (1 - \lambda) g(y_i)} \ , \tag{10.13}$$

the converged values of the weights. In addition

$$U(\tilde{\theta}) = \sum_{i=1}^{n} \tilde{z}_i u_i(\tilde{\theta}) = 0 \tag{10.14}$$

10.6 One-Step Influence Function for $\tilde{\theta}$

The true influence function of $\tilde{\theta}$ will be computed indirectly through first calculating the one-step influence function $\breve{\theta}$. Let $\breve{\theta}$ be the mixture estimator based on the sample $y_1, \ldots, y_n$ and let $\tilde{z}_1, \ldots \tilde{z}_n$ be the weights found by the E-step of the EM algorithm evaluated at $\tilde{\theta}$. We will use $\breve{\theta}(y_1, \ldots, y_n)$ to be the mixture estimator when $y_1, \ldots, y_n$ varies but $\tilde{z}_1, \ldots, \tilde{z}_n$ remains fixed. The calculation of $\breve{\theta}$ takes a single EM step. Note that in this case $\breve{\theta} = \tilde{\theta}$. In addition, if we have observation y, we can able to compute the weights z_y using (10.13).

Let $\breve{\theta}^\epsilon$ be the mixture estimator based on the weighted sample $y_1, \ldots, y_n, y$ having weights $(1-\epsilon)\tilde{z}_1, \ldots (1-\epsilon)\tilde{z}_n, \epsilon\tilde{z}$ then $\breve{\theta}^\epsilon$ satisfies

$$(1-\epsilon)\sum_{i=1}^{n} \tilde{z}_i u_i(\breve{\theta}^\epsilon) + \epsilon \tilde{z} u(\breve{\theta}^\epsilon) = 0 \tag{10.15}$$

The influence function of $\breve{\theta}$, which we define to be the one-step influence function of $\tilde{\theta}$, is denoted by $IF^1_{\tilde{\theta},F_n}(y)$ and we can write

$$\breve{\theta}^\epsilon \approx \breve{\theta} + \epsilon IF_{\breve{\theta},F_n}(y)$$

By the Taylor expansion, we have

$$u_i(\breve{\theta}^\epsilon) = u_i(\breve{\theta}) + \epsilon \left[u_i'(\theta)\right]_{\theta=\breve{\theta}} IF_{\breve{\theta},F_n}(y) + O(\epsilon^2)$$

$$u_i(\breve{\theta}^\epsilon) = u_i(\tilde{\theta}) + \epsilon \left[u_i'(\theta)\right]_{\theta=\tilde{\theta}} IF^1_{\tilde{\theta},F_n}(y) + O(\epsilon^2)$$

From (10.14) and (10.15)

$$\epsilon \sum_{i=1}^{n} \tilde{z}_i \left[u_i'(\theta)\right]_{\theta=\tilde{\theta}} IF^1_{\tilde{\theta},F_n}(y) + \epsilon \tilde{z} u(\tilde{\theta}, y) = 0$$

$$IF^1_{\tilde{\theta},F_n}(y) = -\left[\sum_{i=1}^{n} \tilde{z}_i \left[u_i'(\theta)\right]_{\theta=\tilde{\theta}}\right]^{-1} \tilde{z} u(\tilde{\theta}, y) \tag{10.16}$$

The one-step influence function of $\tilde{\theta}$ is bounded if the product $\tilde{z}u(\tilde{\theta}, y)$ is bounded, because the first part of right hand side in (10.16) is constant. Consider the following examples:

1. Location estimate (please refer the section 4.6.1 for further details)

$$\tilde{\theta} = \frac{\sum_{i=1}^{n} \tilde{z}_i y_i}{\sum_{i=1}^{n} \tilde{z}_i} \text{ and } u(\tilde{\theta}, y) = \frac{y - \tilde{\theta}}{\sigma^2}$$

where

$$\begin{aligned}
\tilde{z} &= \frac{\lambda f(y, \tilde{\theta})}{\lambda f(y, \tilde{\theta}) + (1-\lambda) g(y)} \\
&\quad \text{Since } g \text{ is a constant and let } k = \left(\frac{1-\lambda}{\lambda}\right) g \\
&= \frac{f(y, \tilde{\theta})}{f(y, \tilde{\theta}) + k} \\
&\quad \text{Since } f(y, \tilde{\theta}) = c \exp\left(-\frac{1}{2}\left[\frac{y-\tilde{\theta}}{\sigma}\right]^2\right) \text{ and let } k^* = \frac{k}{c} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\frac{y-\tilde{\theta}}{\sigma}\right]^2\right)}{\exp\left(-\frac{1}{2}\left[\frac{y-\tilde{\theta}}{\sigma}\right]^2\right) + k^*} \\
&= \frac{\exp\left(-\frac{1}{2}r^2\right)}{\exp\left(-\frac{1}{2}r^2\right) + k^*}; \quad r = \frac{y-\tilde{\theta}}{\sigma} \text{and} \\
u(\tilde{\theta}, y) &= \frac{1}{\sigma}\, r
\end{aligned}$$

Hence

$$\lim_{y \to \pm\infty} \tilde{z} u(\tilde{\theta}, y) = \frac{1}{\sigma} \lim_{r \to \pm\infty} \frac{r \ \exp\left(-\frac{1}{2}r^2\right)}{\exp\left(-\frac{1}{2}r^2\right) + k^*} = 0$$

Hence, we can say that the one-step influence function of $\tilde{\theta}$ is bounded. Therefore, our estimate $\tilde{\theta}$ is B-robust.

2. Location and scale estimates (please refer the section 7.5.1 for further details)

$$\tilde{\phi} = \begin{bmatrix} \tilde{\theta} \\ \tilde{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} \tilde{z}_i y_i}{\sum_{i=1}^{n} \tilde{z}_i} \\ \\ \frac{\sum_{i=1}^{n} \tilde{z}_i (y_i - \tilde{\theta})^2}{\sum_{i=1}^{n} \tilde{z}_i} \end{bmatrix}$$

and

$$u(\tilde{\phi}, y) = \begin{bmatrix} u_1(\tilde{\phi}, y) \\ \\ u_2(\tilde{\phi}, y) \end{bmatrix} = \begin{bmatrix} \frac{y-\tilde{\theta}}{\tilde{\sigma}^2} \\ \\ \frac{1}{2\,\tilde{\sigma}^4}[(y-\tilde{\theta})^2 - \tilde{\sigma}^2] \end{bmatrix}$$

By using the similar approach in 1 to the case of $\tilde{z}u_1(\tilde{\phi}, y)$, we can easily show that $\tilde{z}u_1(\tilde{\phi}, y) \to 0$ as $y \to \pm\infty$.

$$\begin{aligned} u_2(\tilde{\phi}, y) &= \frac{1}{2\,\tilde{\sigma}^4}[(y-\tilde{\theta})^2 - \tilde{\sigma}^2] \\ &= \frac{1}{2\,\tilde{\sigma}^2}[r^2 - 1] \end{aligned}$$

Hence

$$\begin{aligned} \lim_{y\to\pm\infty} \tilde{z}u_2(\tilde{\phi}, y) &= \frac{1}{2\,\tilde{\sigma}^2} \lim_{r\to\pm\infty} (r^2-1)\left(\frac{1}{1 + k^* \exp\left(\frac{1}{2}r^2\right)}\right) \\ &= \frac{1}{2\,\tilde{\sigma}^2} \lim_{r\to\pm\infty} \left[\frac{r^2}{1 + k^* \exp\left(\frac{1}{2}r^2\right)} - \frac{1}{1 + k^* \exp\left(\frac{1}{2}r^2\right)}\right] \\ &= \frac{1}{2\,\tilde{\sigma}^2} \lim_{r\to\pm\infty} \frac{r^2}{1 + k^* \exp\left(\frac{1}{2}r^2\right)} - 0 \end{aligned}$$

By using L'Hospital rule,

$$\begin{aligned} &= \frac{1}{2\,\tilde{\sigma}^2} \lim_{r\to\pm\infty} \frac{2}{k^* \exp\left(\frac{1}{2}r^2\right)} \\ &= 0 \end{aligned}$$

Thus the product of $\tilde{z}u(\tilde{\phi}, y)$ is bounded. Hence, the one-step influence functions of $\tilde{\phi}$ is bounded. Therefore our estimate $\tilde{\phi}$ is B-robust.

Note that the bounded one-step influence functions and true influence functions are graphically obtained for the situations described in Chapters 4, 5, 6, and 7.

10.7 True Influence Function of $\tilde{\theta}$

The true influence function is computed using Jorgensen's method. That is

$$IF_{\tilde{\theta},F_n}(y) = (I - J)^{-1} IF^1_{\tilde{\theta},F_n}(y) \tag{10.17}$$

where I is an identity matrix and J is a Jacobian matrix of the updating function evaluated at $\theta = \tilde{\theta}$, which depends on the data and constant matrix. Therefore, the true influence function is bounded if the one-step influence function is bounded. This implies that the boundedness of $\tilde{z}u(\tilde{\theta}, y)$ is the necessary and sufficient condition for B-robustness of the mixture estimator.

I would like to acknowledge to my chief supervisor for providing his input to complete some of the sections of this chapter.

Chapter 11

Summary and Concluding Remarks

Statistics tools such as generalized linear models, mixed models, and time series are popular for making decisions in the fields of engineering, medicine, agricultural science and business, because they help in extracting information from data. The method of likelihood based statistical modelling is a dominant frame work to estimate the model parameters and makes inferences about them. This gives us a unified framework to extend models for the various situations. Bayesian modelling also gives us a unified method for modelling data.

Robust methods became popular after the "Princeton Robustness Year" in 1971-1972, but the application of the robust methods was limited initially because of the unavailability of high powered computers, but nowadays usage of robust methods is not uncommon. One of the main shortcomings in the robust literature is that there are no general principles for creating robust estimates in new situations. Often, the methodology of robust statistics is *ad hoc* and lacks a unified approach. This is in contrast to classical statistical modelling and Bayesian modelling. For example, a number of robust procedures are available for estimating the generalized linear model parameters. The methods are the weighted maximum likelihood estimator [9], CUBIF [35] and later the robust estimation based on the quasi-likelihood [8].

This thesis is attempting to extend the range of robust statistics by making the unified method of likelihood-based statistical modelling more robust.

The main tool for achieving this goal is the two components mixture model, which reduces the influence of "bad" data by tending to assign them to the non-regular component. Note that this additional component would have no intrinsic interest and does not represent a serious attempt to model the outliers. In fact, the finite mixture form is being used as a mathematical tool to obtain a tractable form of analysis, but is not being regarded as the actual data-generating mechanism.

We refer to the parameter free non-regular component as an *outlier component*. This will be defined either based on the data, or prior information, or experience. In the next section, we will give possible choices for the outlier component in some situations. It is possible to think that more than one outlier component is to be added to the regular component. This is not recommended because of the following two reasons: (i) a mixture of outlier components may be considered as a single outlier component; (ii) it seems to make a model for the outliers, in which we are not interested. Adding one outlier component to the regular component is sufficient for our purpose.

The mixture model parameters are estimated by maximizing the observed likelihood function, which is a function of observations and parameters. Weights for each observation are estimated at the E-step of the EM algorithm for a given value of the parameters. We are using the EM algorithm, so that it is an iterative process. The parameter estimates for the regular (non-outlier) mixture component are considered as our robustified estimator for the model parameter.

11.1 Concluding Remarks

In the first part of this section we would like to explain how to choose the outlier component. First, we begin by describing the outlier component chosen for the thesis and follow with other options to define the outlier component.

We choose a parameter free uniform distribution for the outlier component in our examples in the sections 4.7, 5.8, 6.9.2, and 7.6, and in section 6.9.1 we took equal probabilities for success and failure. In addition, the improper g is chosen for the outlier component in Chapter 9. These choices for g give good estimates in our examples.

Apart from these choices for the outlier component, we give brief suggestions for other options. If random variable Y takes the value y as real, we might take the outlier component to be the normal distribution with mean μ (same mean of the regular component) and variance σ^2 (large variance compared with the regular component).

In simple language, choosing an appropriate outlier component is a little like choosing a prior distribution in the Bayesian frame work, but the outlier component is defined over the sample space, not the parameter space.

A small proportion of bad data is generated by unknown mechanisms and rest of them are good data. In our mixture context, $1 - \lambda$ is the proportion of contamination. We believe that in most observations assigned to the regular component, the parameters of which are the main interest in this study, occasionally undesirable observations appear and these are assigned to the outlier component. In that sense, we prefer to have λ close to 1. These assignment are made by the mixture fit.

It is not uncommon to define the contamination proportion $(1-\lambda)$ as a fixed value in the robust measures. For example, trimmed-mean and Winsoried-mean. In addition, it is common to choose the turning constant, which depends on the $(1 - \lambda)$ in the robust literature. We also decided to give a fixed value for the λ. The advantage of fixing λ is that our robustified estimate may be defined as weighted MLE, and the parameter θ of interest is only estimated by our method. In addition, it is easy to drive the influence function for the robustified estimate

Often we take λ to be fixed at 0.90, 0.95 and (1 - true contamination probability) for our numerical illustrations in the sections 4.7, 5.8, 6.9.2, and 7.6. We found that the estimates for these λ values are similar in most cases.

There is scope for extending the research using the mixture method. The standard errors of the estimates need to be investigated. In addition, it could be useful to examine the situation when λ is to be estimated. These explorations may lead to improvements the mixture method. However, our method is still very useful to compute robustified estimates for the model parameters even without these further investigations.

Bibliography

[1] Anderson, R. (2008), *Modern methods for robust regression*, Sage, Los Angeles.

[2] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., Tukey, J. W., (1972), *Regression diagnostics*, Princeton University Press, Princeton.

[3] Barnett, V., Lewis, T. (1994), *Outliers in statistical data*, Wiley, New York.

[4] Bates, D. M., Watts, D. G. (1988), *Nonlinear regression analysis and its applications*, Wiley, New York.

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

[6] Belsey, D. A, Kuh, E., Welsch, R. E. (1980), *Regression diagnostics*, Wiley, New York.

[7] Binaco, A. M., Yohai, V. J. (1996), Robust Estimation in the logistic regression model, *Robust Statistics, Data Analysis and Computer Intensive Methods, Proceedings of the workshop in honor of Peter J Huber*, Lecture Notes in Statistics 109, 17-34, Springer, New York

[8] Cantoni, E., Ronchetti, E. (2001), Robust inference for generalized linear models, *Journal of the American Statistical Association*, 96, 1022-1030.

[9] Carroll, R. J., Pederson, S. (1993), On robustness in the logistic regression model, *Journal of the Royal Statistical Society (B)*, 55, 693-706.

[10] Cook, R. D., Weisberg, S. (1982), *Residuals and influence in regression*, Chapman and Hall, London.

[11] Cook, R. D., Weisberg, S. (1999), *Applied regression including computing and graphics*, Wiley, New York.

[12] Croux, C., Haesbroeck, G. (2003), Implementing the Bianco and Yohai estimator for logistic regression, *Computational Statistics and Data Analysis*, 44, 273-295.

[13] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977),Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J.R. Statist. Soc., (B)*, 39, 1-38.

[14] Dempster, A. P., Laird, N. M., Rubin, D. B. (1980), Iteratively reweighted least squares for linear regression when errors are normal / independent distributed, *Multivariate Analysis*, 5, P.R.Krishnaiah (Ed.), Amsterdam: North Holland, 35-57.

[15] Dobson, A. J. (2002), *An introduction to generalized linear models*, Chapman and Hall, London.

[16] Everitt, B. S., Hand, D. J. (1981), *Finite mixture distributions*, Chapman and Hall, London.

[17] Fox, J. (1997), *Applied regression analysis, linear models, and related methods*, Sage, Los Angeles.

[18] Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2004), *Bayesian data analysis*, Chapman and Hall, London.

[19] Green, P. G. (1984), Iteratively rewighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion), *Journal of the Royal Statistical Society, Series (B)*, 46, 149-192.

[20] Hampel, F. R. (1974), The influence curve and its role in robust estimation, *J.Amer.Statist.Ass.*, 62, 1179-1186.

[21] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986), *Robust statistics: The approach based on influence functions*, Wiley, New York.

[22] Hawkins, D. M. (1980), *Identification of outliers*, Chapman and Hall, London.

[23] Hinkley, D. V. (1977), Jackknifing in unbalanced situations, *Technometrics*, 19, 285-292.

[24] Holland, P. W., Welsch, R. E. (1977), Robust regression using iteratively reweighted least-squares, *Commun. Statist. - Theor. Meth. A6*, 9, 813-827.

[25] Huber, P. J. (1964), Robust estimation of a location parameter, *Ann. Math. Statist.*, 35, 73-101.

[26] Huber, P. J. (1973), Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Stat.*, 1, 799-821.

[27] Huber, P. J. (1981), *Robust statistics*, Wiley, New York.

[28] Hunt, L. A. (1996), *Clustering using finite mixture models*, DPhil. thesis, University of Waikato, New Zealand.

[29] Iglewicz, G., Hoaglin, D. C. (1993), *How to detect and handle outliers*, ASQC Quality Press, Milwaukee.

[30] Jorgensen, M. A. (1993), Influence functions for iteratively defined statistics, *Biometrika*, 80, 2, 253-265.

[31] Jureckova, J., Picek, J. (2006), *Robust statistical methods with R*, Chapman and Hall, London.

[32] Kunsch, H. R., Stefanski, L. A., Carroll, R. J. (1989), Condinally unbiased bounded influence estimation in general regression models, with applications to generalized linear models, *Journal of the American Statistical Association*, 84, 460-466.

[33] Little, R. J. A., Rubin, D. B. (1987), *Statistical analysis with missing data*, Wiley, New York.

[34] Louis, T. A. (1982), Finding the observed information matrix when using EM algorithm, *Journal of the Royal Statistical Society (B)*, 44, 226-233.

[35] Maronna, R. A., Martin, R. D., Yohai, V. J. (2006), *Robust statistics*, Wiley, England.

[36] Mallows, C. L. (1973), Influence functions, *National Bureau of Economic Research Conference on Robust Regression*, Cambridge, MA.

[37] Mallows, C. L. (1975), On some topics in robustness, Technical memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

[38] McLachlan, G. J., Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, Dekker, New York.

[39] McLachlan, G. J., Peel, D. (2000), *Finite mixture models*, Wiley, England.

[40] McLachlan, G. J., Krishnan, T. (1996), *The EM algorithm and extensions*, Wiley, England.

[41] Meng, X. L., Rubin, D. B. (1991), Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *Journal of the American Statistical Association*, 86, 899-909.

[42] Nelder, J. A., Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of the Royal Statistical Society, Series A*, 135, 370-384.

[43] Oakes, D. (1999), Direct calculation of the information matrix via the EM algorithm, *Journal of the Royal Statistical Society, Series (B)*, 61, 479-482.

[44] Peters, B. C., Walker, H. F. (1978a), An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions, *SIAM J.Appl. Math.*, 35, 362-378.

[45] Peters, B. C., Walker, H. F. (1978b), The numerical evaluation of the maximum likelihood estimate of a subset of mixture proportions, *SIAM J.Appl. Math.*, 35, 447-452.

[46] Pregibon, D. (1981), Logistic regression diagonstics, *Ann. Statist.*, 89, 705-724.

[47] Preisser, J. S., Qaqish, B. F. (1999), Robust regression for clustered data with applications to binary regression, *Biometrics*, 55, 574-579.

[48] Rousseeuw, P. J., Leroy, A. M. (1987), *Robust regression & outlier detection*, Wiley, New York.

[49] Stigler, S. M. (1977), Do robust estimators work with real data? *Ann. Statist.*, 5, 1055-1078.

[50] Titterington, D. M., Smith, A. F. M., Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Wiley, New York.