Robust Adaptively Weighted Estimators for Regression Models

by

Wei Tu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

This thesis introduces a new class of robust estimators for regression models. Specifically, a class of weighted least square estimators under linear regression models is introduced in Chapter 2, with a continuous adaptive weight function computed using the Kolmogorov-Smirnov statistic. Asymptotic properties, such as consistency and asymptotic normality, of the proposed estimator are established under the model. Simulation studies show that the proposed estimator attains almost full efficiency and have a better robustness properties than the initial estimators for finite sample sizes. An application to a real contaminated dataset shows that it's comparable to other robust estimators in practice.

In Chapter 3, a class of weighted maximum likelihood estimators under logistic regression models is introduced, again with a continuous adaptive weight function computed using Mahalanobis distances of exploratory variables. Asymptotic consistency of the proposed estimator is proved under the model, and finite-sample properties are also studied by simulation. In simulation studies, it is observed that the proposed estimator is almost as efficient as the maximum likelihood estimator under the model, and under point-mass contamination models, the proposed estimator shows a comparable robustness. This is also verified in an application to a real data set.

Chapter 4 contains some concluding remarks and future directions.

Acknowledgements

First and foremost I want to thank my supervisor Dr. Rohana Karunamuni, for supporting me during the last two years. I would thank him for guiding me in the road of research, showing me what qualities a good researcher should have, and all the encouragement and criticism. Also, I am going to pursue my PhD degree in statistics, and I couldn't have decided to devote myself to research if it's not his support and help.

I would also like to express my sincere thanks to Dr. Linglong Kong, Dr. Irina Dinu and Dr. Bei Jiang for serving in my dissertation examination committee. I would also thank all the professors who taught me before for the mentorship. Also, I couldn't finish this degree without all the help, financially and many others, from Department of Mathematical and Statistical Sciences.

Last but not least, I would like to thank all my friends, and my grandmother who passed away, my parents and my sister for supporting me unconditionally.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Almost all statistical methods rely on some background assumptions about datasets; for example, assuming the data points have a normal distribution. However, in practice sometimes observed data only satisfy the assumptions approximately. For instance, in the linear regression model, the assumed normal distribution model may only hold for majority of observations, and some of them may show different pattern than the others. Unfortunately, this kind of discrepancy will bring many difficulties to statistical methods used for analyzing these data sets. *Robust* statistical methods focus on deriving statistical methods that produce reliable results not only when the observed data follow the assumed assumptions, but also when they only hold approximately. Here the word "robust" refers to the ability of a method to retain its validity under a model misspecification and/or when outliers are present.

For parametric models, robust estimators have been studied extensively in the last fifty years, following the pioneering work of Tukey (1960), Huber

(1967) and Hampel (1968), among others. Marrona et al. (2006) summarizes the most up-to-date results about robust statistics theory in parametric models. For the linear regression model, robust methods will be discussed with some details in the next section. Strong parametric assumptions made on assumed models make it easier to study theoretical properties of robust estimators studied in the literature. However, strong assumptions also bring many restrictions in applying these methods in practice. So, when we have enough knowledge about some model features of the data parametrically but are not willing to assume other features of the models, semiparametric models can be used as an extension of classical parametric models. As robust estimators in semiparametric models, Bickel et al (1993) have introduced a class of generalized M (GM)-estimators, and they have provided general guidelines on how to construct asymptotic "efficient" estimators for regular semiparametric models.

When we extend a classical parametric model to a larger semiparametric model, we usually have to pay a price in efficiency. For example, when we use a threshold value to control observations with large residuals in linear regression models, we would lose some efficiency. This is because there still will be a small fraction of observations that will have large residuals, even when the observed data follows the assumed model perfectly. So, if a large threshold value is chosen then the estimators will lose some robustness against outliers, and if the threshold value is too small then the efficiency lost would be significant. To obtain full efficiency under the true model and maximum robustness under contaminated model simultaneously is a challenging problem in semiparametric models estimation.

1.2 Robust estimates for the linear regression model

Consider a multiple linear regression model with the response variable y_i is related to p explanatory variables $x_{i1}, x_{i2}, ..., x_{ip}$ by the the linear model

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + e_i \qquad i = 1, ..., n,$$

or in matrix form

$$Y = X\theta + e,$$

where e_i's are independent random errors with identical distribution F, $Y = (y_1, y_2, ..., y_n)'$, $X = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})'$, $\theta = (\theta_1, ..., \theta_p)'$, $e = (e_1, e_2, ..., e_n)'$ and the e_i's are independent of the x_i's. The goal is to find the "best" estimator of unknown parameter θ that captures the relationship between the response and explanatory variables. Suppose we already have an estimator $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$, then we can obtain

$$\hat{y}_i = \sum_{j=1}^{p} x_{ij}\hat{\theta}_j,$$

and $\hat{y}_i$ is called the predicted or estimated value of y_i. Residual r_i measures the difference between observed y_i and estimated $\hat{y}_i$:

$$r_i = y_i - \hat{y}_i, \ \ i = 1, ..., n.$$

Naturally, we want an estimator $\hat{\theta}$ to make the residuals as small as possible. The popular least squares (LS) estimators are obtained by minimizing the

residual sum of squares

$$\hat{\theta}_{LS} = \min_{\theta} \sum_{i=1}^{n} r_i^2, \tag{1.1}$$

or, by differentiating (1.1) and solving the following p equations:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\theta_j)x_{ik} = 0, \qquad k = 1, ..., p.$$

In matrix form, they can be written as

$$X^{'}X\hat{\theta}_{LS} = X^{'}Y.$$

If the design matrix X has full rank, then a unique solution can be obtained as

$$\hat{\theta}_{LS} = (X^{'}X)^{-1}X'Y.$$

There are many desirable equivariance qualities that we want to have for our estimators. For example, we call a regression estimator $\hat{\theta}$ *regression equivariant* if for all $\gamma \in \mathbb{R}^p$,

$$\hat{\theta}(X, Y + X\gamma) = \hat{\theta}(X, Y) + \gamma. \tag{1.2}$$

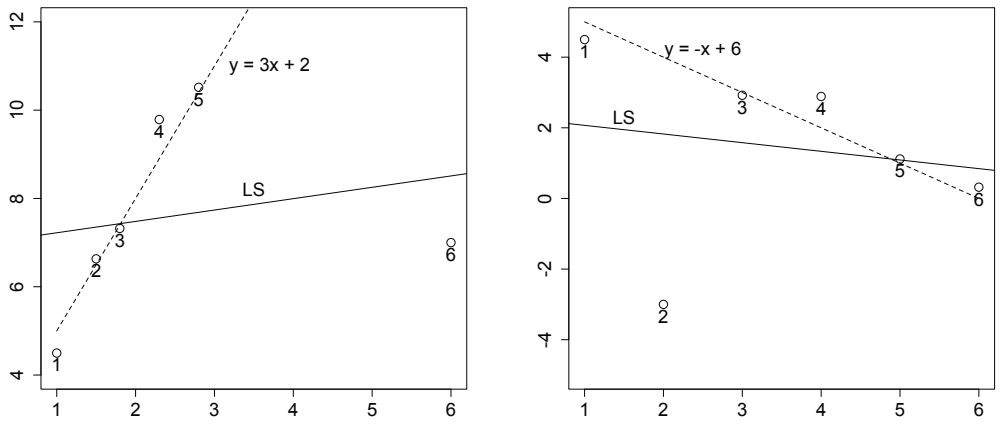
It is *scale equivariant* if for all $\lambda \in \mathbb{R}$

$$\hat{\theta}(X, \lambda Y) = \lambda\hat{\theta}(X, Y), \tag{1.3}$$

and *affine equivariant* if for all nonsingular $p \times p$ matrices A

$$\hat{\theta}(XA, Y) = A^{-1}\hat{\theta}(X, Y). \tag{1.4}$$

The LS estimators satisfy all of the above desirable properties. When the errors e_i are homoscedastic and uncorrelated, LS estimators achieves the lowest possible mean squared error among all estimators based on the linear combination of y_i. Also, when the errors are normally distributed, the LS estimator (LSE) equals to *maximum likelihood estimator* (MLE), which reaches the *Cramér Rao lower bound* (defined in Section 1.3) of the variance of estimators. Thus, the LSE is efficient under normal errors. However, when these assumptions fail to hold (e.g., when there exist outliers), it will bring dramatic changes on the LSE. We present two examples based on some artificial data to show how much damage a single atypical observation can do to the LSE; they are depicted on Figure 1.1.



(a) (b)

Figure 1.1: *Scatterplots of contaminated data set with LS fit line and the line used to simulate data*

From Figure 1.1a, we can see the LS fit line was dragged down by observation #2. In fact, the artificial data set was simulated from the model $y = -x + 6 + e$ with random error e from the $N(0, 1)$ distribution, and ob-

servation #2 is a gross error $(2, -3)$. This point is called an *outlier in the y-direction* because it has an atypical y-value compared to the bulk of the data set, and we can see it has a large residual.

For Figure 1.1b , observation #6 changes the LS regression line dramatically since its x-value is very different from the others. This artificial data set was simulated from the model $y = 3x + 2 + e$ with random error e from the $N(0, 1)$ distribution, and observation #6 is a gross error $(6, 7)$. This point is called an *outlier in the x-direction* or a *leverage point*, and actually this point doesn't have a very large residual, so it will be hard to diagnose by checking residuals. However, there are other diagnostic methods to check leverage points; for example, by checking the diagonal elements $h_1, h_2, ..., h_n$ of the so-called "hat matrix" $H = X(X'X)^{-1}X'$, and h_i can also be called a *leverage*, so x-outliers can also be called *leverage points*.

As we can see, both independent and dependent variables can be contaminated by outliers, and it brings many difficulties in robustifying the least squares method. Huber (1964) proposed a straightforward and popular *M-estimators* defined by minimizing a "bounded" version of sum of residuals:

$$\hat{\theta}_M = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(y_i - \sum_{j=1}^{p} x_{ij}\theta_j), \tag{1.5}$$

and $\hat{\theta}_M$ can be obtained by solving the following p equations:

$$\sum_{i=1}^{n} \psi(y_i - \sum_{j=1}^{p} x_{ij}\theta_j)x_{ik} = 0, \qquad k = 1, ..., p,$$

where ρ is a bounded nondecreasing function of $|x|$, with $\rho(0) = 0$,and $\psi = \rho'$. If ρ is bounded, it is also assumed that $\rho(\infty) = 1$. Huber's basic idea is

to control observations with too large residuals. To make the estimator scale invariant, a scale parameter σ is introduced, and then the estimating equations become

$$\sum_{i=1}^{n} \psi(\frac{r_i(\hat{\theta})}{\sigma})\mathbf{x_i} = 0. \tag{1.6}$$

In practice, σ can be estimated by *standardized Median Absolute Deviation* (MAD), which can be defined as

$$\hat{\sigma} = \frac{1}{0.675}\text{Median}_i(|r_i| \mid r_i \neq 0).$$

M-estimator $\hat{\theta}_M$ is regression, affine and scale equivariant, and its asymptotic distribution of $\hat{\theta}_M$ can also be derived. M-estimators are robust to y-outliers, however, they are not robust to x-outliers.

In order to quantify the robustness property of estimators, many methods have been proposed. Hampel (1971) proposed asymptotic *breakdown point* (BP), and he also proposed the *influence function technique* (Hampel, 1974) as an asymptotic version of the sensitivity curve. Roughly speaking, the breakdown point of an estimator is the proportion of incorrect observations an estimator can handle before giving an arbitrarily large result (Hampel 1971, Donoho and Huber 1983). *Asymptotic contamination* BP is formally defined as

Definition: The asymptotic contamination BP of the estimate $\hat{\theta}$ at F, denoted by $\epsilon^*(\hat{\theta}, F)$, is the largest $\epsilon^* \in (0, 1)$ such that for $\epsilon < \epsilon^*$, $\hat{\theta}_\infty((1-\epsilon)F+\epsilon G)$ as a function of G remains bounded, and also bounded away from the boundary of Θ.

The BP of M-estimators is 0 due to its vulnerability to leverage points.

There are many robust regression estimators achieving an asymptotic breakdown point 0.5. Rousseeuw (1984) proposed *Least Median of Squares* (LMS) estimator by replacing the sum in (1.1) with the median, and it can be defined as

$$\hat{\theta}_{LMS} = \min_{\theta} \text{Med } r_i^2.$$

The LMS estimator is very robust with respect to y outliers as well as x outliers, and it is regression, scale and affine equivariant. But it has an asymptotic convergence rate of $n^{-1/3}$, which means its relative efficiency with respect to the LS estimator is 0, since the asymptotic convergence rate of the LS estimator is $n^{-1/2}$. To overcome this, Rousseeuw (1984) introduced the *Least Trimmed Squares* (LTS) estimator by eliminating the observations with the largest residuals while minimizing the sum of squares, and it is defined as

$$\hat{\theta}_{LTS} = \min_{\theta} \sum_{i=1}^{h} (r^2)_{i:n},$$

where $(r^2)_{1:n} \leq ... \leq (r^2)_{n:n}$ are the ordered squared residuals. The LTS estimator achieves the highest breakdown point 0.5 when h is approximately $n/2$, and it has excellent robustness properties against y- and x- outliers. It's also regression, scale and affine equivariant. Further, it has a normal convergence rate of $n^{-1/2}$. But compared with the LS estimator, it only has a relative efficiency about 7% under the normal model.

To obtain an estimate with high efficiency and high BP simultaneously, Yohai (1987) proposed an MM-estimator by using an initial robust estimator with high breakdown point and an M-estimator. His idea is as follows: first obtain an initial consistent estimate $\hat{\theta}_0$ with a high BP but may be with low

efficiency (for example, LMS, LTS are possible candidates); second, compute a robust scale $\hat{\sigma}$ of the residuals $r_i(\hat{\theta}_0)$ (use MAD, for example) and then iteratively solve (1.6) starting with $\hat{\theta}_0$. The resulting estimator has a breakdown point of 0.5, as long as the initial estimator $\hat{\theta}_0$ has a BP of 0.5.

1.3 Preliminaries

In this subsection, we state a few definitions, some empirical processes results and other results that are used throughout the thesis. Most of them are taken from the monograph *Asymptotic Statistics* by van der Vaart (2000).

Suppose θ is an unknown deterministic parameter which is to be estimated from measurements X, distributed according to some probability density function $f(x;\theta)$. The variance of any unbiased estimator $\hat{\theta}$ of θ is then bounded by the reciprocal of the *Fisher information* $I(\theta)$:

$$var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the Fisher information $I(\theta)$ is defined by

$$I(\theta) = E[(\frac{\partial l(X;\theta)}{\partial \theta})^2] = -E[\frac{\partial^2 l(X;\theta)}{\partial \theta^2}]$$

and $l(x;\theta) = \log(f(x;\theta))$ is the natural logarithm of the likelihood function, and the inverse of Fisher information $I(\theta)^{-1}$ is known as the Cramér-Rao lower bound when estimating θ.

The *efficiency* of an unbiased estimator $\hat{\theta}$ is defined as

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{var(\hat{\theta})}.$$

Then the Cramér-Rao lower bound gives $e(\hat{\theta}) \leq 1$.

Assume that an estimator T_n of $\phi(\theta)$ based on n observations has the property that, as $n \to \infty$,

$$\sqrt{n}(T_n - \phi(\theta)) \rightsquigarrow N(0, \sigma^2(\theta)).$$

Let $T_{1,n}$ and $T_{2,n}$ denote two such estimators with asymptotic variances $\sigma_1^2(\theta)$ and $\sigma_2^2(\theta)$, respectively. Then the asymptotic efficiency of $T_{1,n}$ with respect to (w.r.t.) $T_{2,n}$ is given by

$$\frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

If the ratio is bigger than 1, then the second estimator needs proportionally as many observations more than the first one to achieve the same asymptotic precision.

Consider again the linear regression model $Y = X\theta + e$ with design matrix X. Assume that $\lim_{n\to\infty} \frac{1}{n} X^T X = C$, where C is a positive definite matrix. Let σ_e^2 be the variance of the error term e. Then the asymptotic distribution of the LSE can be written as

$$\sqrt{n}(\hat{\theta}_{LS} - \theta_0) \xrightarrow{d} N(0, \sigma_e^2 C^{-1}).$$

Let $X_1, ..., X_n$ be a random sample from a distribution function F. Then

the empirical distribution function of the sample is defined as

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq t),$$

which is an unbiased estimator of the distribution F. From the strong law of large numbers, it is consistent:

$$\mathbb{F}_n(t) \xrightarrow{as} F(t), \quad for\ every\ t.$$

Glivenko-Cantelli theorem extends this result from pointwise convergence to uniform convergence. This theorem is given next.

Theorem 1.1: *If $X_1, X_2, ...$ are i.i.d. random variables with distribution function F, then $\|\mathbb{F}_n - F\|_\infty = sup_t|\mathbb{F}_n - F| \xrightarrow{as} 0$.*

If the random sample $X_1, ..., X_n$ is from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$, then the empirical distribution, as a discrete uniform measure, can be defined as $\mathbb{P}_n = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$, where δ_x is the probability distribution that degenerates at x. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, the expectation of f under empirical measure $\mathbb{P}_n f$ can be defined as

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

and similarly the expectation of f under P is $Pf = \int f dP$. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n f - Pf\| = \sup_{f\in\mathcal{F}}|\mathbb{P}_n f - Pf| \xrightarrow{as} 0.$$

The empirical process evaluated at f is defined as $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$,

and a class $\mathcal{F}$ is called P-Donsker if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges in distribution to a tight limit process in the space $l^\infty(\mathcal{F})$.

Whether a class of functions is Glivenko-Cantelli or Donsker depends on the "size" of the class. A way to measure the size of a class $\mathcal{F}$ is items of entropy. Consider the bracketing entropy relative to the $L_r(P)$-norm

$$\|f\|_{P,r} = (P|f|^r)^{1/r}.$$

Given two functions l and u, the bracket $[l, u]$ is the set of all functions ψ_i with $l \leq \psi_i \leq u$. A ε-bracket in $L_r(P)$ is a bracket $[l, u]$ with $P(u - l)^r < \varepsilon^r$. The bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, L_r(P))$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. (l, u must have finite $L_r(P)$ norms, but need not belong to $\mathcal{F}$.) The following theorem gives a criteria to check whether a class is P-Glivenko-Cantelli, and it's a generalization of the classic Glivenko-Cantelli theorem (Theorem 1.1).

Theorem 1.2 (Glivenko-Cantelli): *Every class $\mathcal{F}$ of measurable functions such that $N_{[\,]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$ is P-Glivenko-Cantelli.*

If a class of functions $\mathcal{F}$ is shown to be Glivenko-Cantelli, then we have the uniform convergence within the class. Theorem 5.9 in van der Vaart (2000) provides a way to prove the consistency of an estimator. This result is given next.

Theorem 1.3: *Let Ψ_n be random vector-valued functions and let Ψ be a fixed vector-valued function of θ such that for every $\epsilon > 0$*

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{P} 0,$$

$$\inf_{\theta: d(\theta,\theta_0)\geq\epsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|.$$

Then any sequence of estimators $\hat{\theta}_n$ *such that* $\Psi_n(\hat{\theta}_n) = o_P(1)$ *converges in probability to* θ_0.

1.4 Summary and organization of the thesis

This thesis proposes a new class of robust estimators for the linear and logistic regression models. They are essentially "weighted estimators". Specifically, Chapter 2 proposes a class of weighted least squares estimators for the linear regression model. Instead of using cut off threshold values to eliminate the effects of extreme observations, the proposed estimator in Chapter 2 uses a continuous weight function, calculated from an initial robust estimators of the regression as well as the scale. The weighting step on the initial robust estimators improves the efficiency while still keeping the high breakdown point. The resulting estimator is named as *Robust and Efficient Continuous Weighted Least Squares* (RECWLS) estimator. I have proved that the proposed RECWLS estimator is asymptotically consistent. Furthermore, the asymptotically distribution of RECWLS estimator is also derived, and it attains the full efficiency under normal errors. Simulation studies presented in the thesis also verify that the proposed estimator exhibits full efficiency under normal errors for finite sample sizes. They have comparable efficiencies to threshold-value based estimators under t-error models. When the outliers have small and large leverage points, the proposed estimator shows a better robustness than the initial robust estimators, especially when the outliers are more extreme.

Chapter 3 proposes a class of weighted maximum likelihood estimators for the logistic regression model, with a continuous weight function computed using an adaptive Mahalanobis distances of exploratory variables. The asymptotic consistency of the proposed estimator is proved. A simulation study shows that the proposed estimator is almost as efficient as the maximum likelihood estimator under the clean model. Under point-mass contamination models, the proposed estimator has a comparable robustness, especially when the contaminated points are on the boundary of the exploratory space.

Some concluding remarks and future directions are stated in Chapter 4.

Chapter 2

Linear Regression Model

2.1 Preliminaries: the linear regression model and existing estimators

Consider a random sample of observations $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of p explanatory variables and y_i is the response variable. Assume that they are linked by the linear relationship

$$y_i = \mathbf{x}_i^T \theta + e_i, \quad i = 1, ..., n, \tag{2.1}$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$ is the unknown regression parameter that needs to be estimated, the error terms e_i's are i.i.d. unobservable random variables with unknown distribution $F_0(./\sigma_0)$ for some scale parameter $\sigma_0 > 0$, the e_i's are independent with the covariates $\mathbf{x}_i$'s. We also assume F is symmetric about 0 to simplify some theoretical proofs.

In order to find an estimator θ, say $\hat{\theta}$, which captures the linear relationship between $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ and $Y = (y_1, ..., y_n)^T$, the well-known least squares

method tries to find $\hat{\theta}$ minimizing $\sum(y_i - \mathbf{x}_i^T\theta)^2$. Then $\hat{\theta}$ is the solution to the estimating equation

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\theta)\mathbf{x}_i = 0. \tag{2.2}$$

The least square estimator (LSE) is efficient under normal error models, and it achieves the lowest possible variance among all unbiased estimators based on the linear combination of y values. From the estimating equation (2.2), we can see that LSE can be easily affected by (i) observations with extreme values of standardized residuals, (ii) observations with extreme values of leverages, and (iii) failures of the assumed model.

To obtain a more robust estimator, Rousseeuw (1984) proposed an equivariant regression estimator that attains the maximum asymptotic breakdown point 1/2, *least median of squares estimator* (LMS), as well as a *least trimmed squares estimator* (LTS). Both LMS and LTS estimators show great robustness properties against outliers and high leverage points. But the main disadvantage of them is that they are not very efficient. For the LMS estimator, the rate of convergence is $n^{-1/3}$, so its relative efficiency with respect to LSE is 0. The LTS estimator achieves the normal $n^{-1/2}$ convergence rate, but its relative efficiency w.r.t. the LSE is only about 7%.

The excellent robustness properties of LMS and LTS estimators are still useful to construct better estimators that are more efficient. Suppose $\hat{\theta}$ and $\hat{\sigma}$ denote a pair of initial robust estimators of regression and scale, respectively. Then initial standardized residuals can be calculated as

$$r_i = \frac{y_i - \mathbf{x}_i^T\hat{\theta}}{\hat{\sigma}}.$$

Now use a weighting step on the initial robust estimators to improve the efficiency while still keeping the high breakdown point (i.e., high robustness). Rousseeuw and Leroy (1987) proposed a weight function based on the initial standardized residuals

$$w_i = \begin{cases} 1 & \text{if } r_i < t_0 \\ 0 & \text{if } r_i \geq t_0 \end{cases}$$

with $t_0 = 2.5$ as a reasonable fixed threshold. The rationale for the choice $t_0 = 2.5$ is that if one assumes a normal-error model, then $r_i \geq t_0$ would indicate an outlier. The resulting weighted least squares estimator is of the form $\theta_{1n} = (X^T W X)^{-1} X^T W Y$, where the matrix $W = \text{diag } (w_1, ..., w_n)$. Even when the error distribution perfectly follows the assumed distribution, there still will be a small fraction of observations with $r_i > t_0$, which will be eliminated using fixed threshold weight functions. If a really large t_0 is chosen, then the robustness properties will be influenced. This fixed threshold based weighted least squares estimator does retain a high breakdown point. Further, He and Portnoy (1992) showed that it converges no faster than the initial estimator $\hat{\theta}$, which means that its asymptotic relative efficiency compared to the LSE could be still very low.

Gervini and Yohai (2001) proposed a class of *robust and efficient weighted least squares estimators* (REWLSE), using adaptive cut-off values. Their adaptive cut-off values are computed using the empirical distribution of the residuals of an initial robust estimator. They also proved that their REWLSE has the full asymptotic efficiency if the errors are normally distributed. Furthermore, the REWLSE also has asymptotic breakdown points no less than those of the initial estimators.

2.2 RECWLS estimator

Instead of using cut-off values to eliminate the observations with high residuals, I will propose a class of weighted least square estimators with a continuous weight function. I will refer these new estimators as *robust and efficient continuous weighted least squares* (RECWLS) estimators in what follows.

First, let $\hat{\eta}$ denote a test statistic that measures the goodness-of-fit of the observed data to the assumed model. We assume that $\hat{\eta}$ estimates some unknown parameter $\eta \geq 0$. In fact $\eta = 0$ if the model is correctly chosen; that is, $\hat{\eta}$ is a consistent estimator of 0 under the true model. There are many ways to construct such a statistic. Let the empirical distribution function of the standardized absolute residuals of initial robust estimators be

$$F_n^+(t) = \frac{1}{n} \sum_{i=1}^{n} I(|r_i| \leq t),$$

and let $F_0^+(t)$ be the assumed cumulative distribution function of the absolute errors under the model. Based on Kolmogorov-Smirnov statistic, by comparing $F_n^+(t)$ and $F_0^+(t)$, following Gervini and Yohai (2002) we define

$$\hat{\eta} = \sup_{t \geq t_0} \left| F_0^+(t) - F_n^+(t) \right|.$$

Then for $t_0 = 0$, $\hat{\eta}$ can be considered a measure of the goodness-of-fit of the observed data to the assumed model.

When $\left| F_0^+(t) - F_n^+(t) \right|$ is large for a large t, it means the outliers are present

in the sample, since a heavier tail appeared in the empirical distribution function $F_n^+(t)$. The value $t_0 = 2.5$ is a reasonable choice in application. Also $\hat{\eta}$ can be viewed as a measurement of percentage of outliers present in the sample. Since the true error distribution is usually unknown, a hypothetical distribution F is used instead of $F_0^+(t)$, and $F = \Psi$ is a recommended choice in practice (Gervini and Yohai, 2002), where Ψ denotes the cumulative distribution function of $N(0,1)$. It must be pointed out that even though it is assumed that $F = \Psi$, it doesn't mean that the error terms e_i are assumed to come from a normal distribution. We observed that $\hat{\eta}$ is a pretty robust to misspecification of F. I will show this fact in the simulation studies section below. There are many other ways to construct a goodness-of-fit of the observed data to the assumed model, such as Cramer-von Mises statistic, $\sqrt{n}\int(F_n(x) - F_0(x))^2 dF_0(x)$.

We define a weight function of the form

$$w_\beta(\mathbf{x}, y) = m(\eta\,|r_{\alpha,\sigma}|),$$

where $\beta = (\eta, \alpha, \sigma)$ denotes nuisance parameters, population residual $r_{\alpha,\sigma} = \frac{y - \mathbf{x}^T\alpha}{\sigma}$ with $\alpha \in \Theta$, $\sigma \in \mathbb{R}$, and m is an absolutely continuous non-increasing mapping from $\mathbb{R}^+$ to $(0,1]$ such that $m(0) = 1, \sup_{x>0}[xm(x)] < \infty$, and its first derivative is bounded with $m^{(1)}(0) = 0$. Define a objective function

$$\psi_{\theta,\beta}(\mathbf{x}, y) = w_\beta(\mathbf{x}, y)\phi_\theta(\mathbf{x}, y), \quad \mathbf{x}, \theta \in \mathbb{R}^p,\ y \in \mathbb{R}, \tag{2.3}$$

where $\phi_\theta = (y - \mathbf{x}^T\theta)\mathbf{x}$, $\theta \in \Theta$. Then I define the RECWLS estimator $\hat{\theta}_n$ as

the solution of estimating equation

$$\sum_{i=1}^{n} \psi_{\theta,\hat{\beta}}(\mathbf{x}_i, y_i) = 0, \tag{2.4}$$

where $\hat{\beta}$ is an initial estimator of the nuisance parameter β. The solution $\hat{\theta}_n$ of (2.4) can be written as a weighted least squares estimator:

$$\hat{\theta}_n = (X^T W X)^{-1} X W Y,$$

where $W = \text{diag}(w_{\hat{\beta}}(\mathbf{x}_1, y_1), ..., w_{\hat{\beta}}(\mathbf{x}_n, y_n))$. For robustness purposes of $\hat{\theta}_n$, $\hat{\beta}$ should be chosen as a robust estimator of β.

2.3 Asymptotic properties

This section studies asymptotic properties of the RECWLS estimator. I will show that under some general assumptions on the error distribution F_0 and the moments of explanatory variables, the estimator is asymptotically consistent and has an asymptotic normal distribution. For models with normally distributed errors, the proposed estimator achieves the full asymptotic efficiency.

Assume the random vector $(\mathbf{X}, Y)$ follows the central model if

$$(\mathbf{X}, Y) \sim H_0, \quad \textit{with } H_0(\mathbf{x}, y) = G_0(\mathbf{x}) F_0\{(y - \mathbf{x}^T\theta)/\sigma\}.$$

Asymptotic is established under the assumption that the sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ follows the linear regression model (2.1). Assume that θ_0 and σ_0 are the "true value" of θ and σ, respectively, where σ denotes the scale parameter of the error distribution.

We define functions

$$\mathbb{P}_n\psi_\theta = \Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\theta,\hat{\beta}}(\mathbf{x}_i, y_i)$$

and

$$P\psi_\theta = P\psi_{\theta,\beta} \;\; and \;\; \Psi(\theta) = P\phi_\theta = P\psi_{\theta,0},$$

where $\hat{\beta} = (\hat{\eta}, \hat{\alpha}, \hat{\sigma})$, $\psi_{\theta,0}$ is obtained from (2.3) with $\beta_0 = (0, \theta_0, \sigma_0)$, and P denotes the (unknown) joint probability distribution of the $(\mathbf{x}, y)$'s, H_0.

The consistency of $\hat{\theta}_n$ is established in Theorem 2.1 below. To prove this theorem, I first state and prove three lemmas. The following assumptions are needed in Theorem 2.1 and in the lemmas:

A1 $\hat{\alpha} \xrightarrow{p} \theta_0$ and $\hat{\sigma} \xrightarrow{p} \sigma_0$.

A2 $E(\mathbf{X}\mathbf{X}^T)$ is nonsingular.

A3 $E_{G_0}(\|\mathbf{x}\|^4) < \infty$ and F_0 has finite fourth moment.

A4 The weight function $m(x)$ is continuous, has bounded first derivative, $m(0) = 1$ and $m^{(1)}(0) = 0$.

For most popular initial robust estimators, like the least median squares and the least trimmed squares used in simulation studies, A1 is satisfied. In the theorems and lemmas stated below, asymptotic are understood to be as $n \to \infty$.

Lemma 2.1: *If A1 is satisfied, then* $\hat{\eta} = o_p(1)$.

Proof: Recall that standardized error term $\epsilon_i = \frac{e_i}{\sigma_0} = \frac{y_i - \mathbf{x}_i^T\theta_0}{\sigma_0}$ and $r_i = \frac{y_i - \mathbf{x}_i^T\hat{\alpha}}{\hat{\sigma}}$, $F_0^+(t)$ is the distribution function of the absolute errors under central

model, and $\hat{F}_n^+(t) = \frac{1}{n}\sum_{i=1}^n I(|r_i| \leq t)$. We can write

$$\begin{aligned} r_i &= \frac{y_i - \mathbf{x}_i^T\theta_0 + \mathbf{x}_i^T\theta_0 - \mathbf{x}_i^T\hat{\alpha}}{\sigma_0}\frac{\sigma_0}{\hat{\sigma}} \\ &= \frac{\sigma_0}{\hat{\sigma}}\epsilon_i + \frac{1}{\hat{\sigma}}\mathbf{x}_i^T(\theta_0 - \hat{\alpha}), \end{aligned}$$

and so

$$\begin{aligned} Pr(r_i \leq t) &= Pr(\frac{\sigma_0}{\hat{\sigma}}\epsilon_i + \frac{1}{\hat{\sigma}}\mathbf{x}_i^T(\theta_0 - \hat{\alpha}) \leq t) \\ &= Pr(\epsilon_i \leq t\frac{\hat{\sigma}}{\sigma_0} + \frac{1}{\sigma_0}\mathbf{x}_i^T(\hat{\alpha} - \theta_0)) \\ &= F_0(t\frac{\hat{\sigma}}{\sigma_0} + \frac{1}{\sigma_0}\mathbf{x}_i^T(\hat{\alpha} - \theta_0)) \\ &= F_1(t). \end{aligned}$$

Define a class of measurable functions $\mathcal{F}_0 = \{f_t = 1_{[-t,t]} : t \in \mathbb{R}^+\}$, and also denote

$$\mathbb{P}_n f_t = \frac{1}{n}\sum_{i=1}^n I(|r_i| \leq t) = F_n^+(t),$$

$$Pf_t = Pr(r \leq t) = F_1^+(t) = F_1(t) - F_1(-t).$$

Now we show that $\mathcal{F}_0$ is P-Glivenko-Cantelli. Consider brackets of the form $[1_{[-t_{i-1},t_{i-1}]}, 1_{[-t_i,t_i]}]$ for a grid of points $0 = t_0 < t_1 < ... < t_k = \infty$ with the property that $F(t_i-) - F(t_{i-1}) < \varepsilon/2$ for each i, where $\varepsilon > 0$. These brakes have $L_1(F)$-size ε. Their total number k can be chosen smaller than $2/\varepsilon$, so for every $\varepsilon > 0$, $N_{[]}(\varepsilon, \mathcal{F}_0, L_1(P)) < \infty$. Now using (Glivenko-Cantelli) Theorem 19.4 in van der Vaart (2000), $\mathcal{F}_0$ is P-Glivenko-Cantelli, and so we have

$$\|\mathbb{P}_n f_t - Pf_t\|_{\mathcal{F}_0} = \sup_{f_t \in \mathcal{F}_0} |\mathbb{P}_n f_t - Pf_t| = \sup_{t>0} |F_n^+(t) - F_1^+(t)| \xrightarrow{as} 0.$$

Note that

$$\begin{aligned}\hat{\eta} &= \sup_{t\geq t_0}\left|F_0^+(t) - \hat{F}_n^+(t)\right| \\ &\leq \sup_{t\geq 0}\left|F_0^+(t) - \hat{F}_n^+(t)\right| \\ &= \sup_{t\geq 0}\left|F_0^+(t) - F_1^+(t)\right| + \sup_{t\geq 0}\left|F_n^+(t) - F_1^+(t)\right|.\end{aligned}$$

Now we only have to show $\sup_{t\geq 0}\left|F_0^+(t) - F_1^+(t)\right| = o_p(1)$ to prove $\hat{\eta} = o_p(1)$. Observe that

$$\begin{aligned}F_0^+(t) - F_1^+(t) &= F_0(t) - F_0(-t) - F_1(t) + F_1(-t) \\ &= (F_0(t) - F_0(t\frac{\hat{\sigma}}{\sigma_0} + \frac{1}{\sigma_0}\mathbf{x}_i^T(\hat{\alpha} - \theta_0))) \\ &\quad - (F_0(-t) - F_0(-t\frac{\hat{\sigma}}{\sigma_0} + \frac{1}{\sigma_0}\mathbf{x}_i^T(\hat{\alpha} - \theta_0))).\end{aligned}$$

Since $\hat{\alpha} \xrightarrow{p} \theta_0$, it is easy to see that for every $t > 0$, we have $|F_0^+(t) - F_1^+(t)| \to^P 0$. That F_0 is a continuous distribution function implies that this convergence is uniform in t and hence $\hat{\eta} = o_p(1)$.

Lemma 2.2: *If A2 is satisfied, then* $\|\Psi(\theta_n)\| \to 0$ *implies* $\|\theta_n - \theta_0\| \to 0$ *for any sequence* $\{\theta_n\} \in \Theta$.

Proof: Note that

$$\begin{aligned}\|\Psi(\theta_n)\| &= \|P\psi_{\theta_n,0}\| \\ &= \|P\phi_{\theta_n}\| \\ &= \left\|P(y - \mathbf{x}^T\theta_n)\mathbf{x}\right\| \\ &= \left\|P(\mathbf{x}^T\theta_0 + e - \mathbf{x}^T\theta_n)\mathbf{x}\right\| \\ &= \left\|P(\mathbf{x}^T(\theta_0 - \theta_n)\mathbf{x})\right\|.\end{aligned}$$

If $\|\Psi(\theta_n)\| \to 0$, then we have $\Psi(\theta_n) \to 0$. Since $\mathbf{x} = (x_1, x_2, ..., x_p)^T$, $\theta_0 - \theta_n = \theta_d = (\theta_{d1}, \theta_{d2}, ..., \theta_{dp})^T$, we obtain

$$P(\mathbf{x}^T(\theta_0 - \theta_n)\mathbf{x}) = P(\sum_{i=1}^{p} \theta_{di} \begin{pmatrix} x_1x_i \\ x_2x_i \\ ... \\ x_px_i \end{pmatrix}) = \sum_{i=1}^{p} \theta_{di} P(\begin{pmatrix} x_1x_i \\ x_2x_i \\ ... \\ x_px_i \end{pmatrix}) \to \mathbf{0}.$$

Since $P(\mathbf{X}\mathbf{X}^T) = \mathrm{E}(\mathbf{X}\mathbf{X}^T)$ is nonsingular, we have $\theta_0 - \theta_n \to \mathbf{0}$, and therefore we have $\|\theta_n - \theta_0\| \to 0$.

Lemma 2.3: *If A3 and A4 are satisfied, then the class* $\{\psi_{\theta,\beta} : \theta \in \Theta, \|\beta - \beta_0\| < \delta\}$ *is P-Glivenko-Cantelli for some* $\delta > 0$, *where* $\beta = (\eta, \alpha, \sigma)$ *and* $\beta_0 = (0, \alpha_0, \sigma_0)$ *is a fixed value.*

Proof: To show a class $\mathcal{F}$ of vector-valued functions $\psi : (\mathbf{x}, y) \mapsto \mathbb{R}^p$ to be Glivenko-Cantelli, we only need to show that each of the classes of coordinates $\psi^i : (\mathbf{x}, y) \mapsto \mathbb{R}$ with $\psi = (\psi^1, ..., \psi^p)^T$ ranging over $\mathcal{F}(i = 1, 2, ..., p)$ is Glivenko-Cantelli.

The class $\mathcal{F} = \{\psi_\gamma : \gamma = (\theta, \beta) = (\theta, \eta, \alpha, \sigma),\ \alpha, \theta \in \Theta,\ \eta \in (0, 1),\ \sigma^* < \sigma,\ \|\beta - \beta_0\| < \delta\}$ is a collection of measurable functions indexed by a bounded subset in $\Gamma = \Theta \times \Theta \times \mathbb{R} \times \mathbb{R} \subset \mathbb{R}^{2p+2}$. The parameter space consists of three nuisance parameters β and the target parameter θ. For any $\gamma_1 = (\theta_1, \beta_1)$ and

$\gamma_2 = (\theta_2, \beta_2)$,

$$\begin{aligned}
\left|\psi^i_{\gamma_1}(\mathbf{x}, y) - \psi^i_{\gamma_2}(\mathbf{x}, y)\right| &= \left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_1}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right| \\
&= |w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_1}(\mathbf{x}, y) - w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y) + \\
&\quad w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)| \\
&\leq \left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_1}(\mathbf{x}, y) - w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right| + \\
&\quad \left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right|.
\end{aligned} \tag{2.5}$$

Also we have

$$\begin{aligned}
&\left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_1}(\mathbf{x}, y) - w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right| \\
&= w_{\beta_1}(\mathbf{x}, y)\left|(y - \mathbf{x}^T\theta_1)x_i - (y - \mathbf{x}^T\theta_2)x_i\right| \\
&\leq \left|\mathbf{x}^T(\theta_2 - \theta_1)x_i\right| \\
&\leq |x_i|\,\|\mathbf{x}\|\,\|\theta_2 - \theta_1\|.
\end{aligned} \tag{2.6}$$

Then we obtain

$$\begin{aligned}
&\left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right| \\
&= |w_{\beta_1}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)|\left|y - \mathbf{x}^T\theta_2\right||x_i| \\
&= |w_{\eta_1,\alpha_1,\sigma_1}(\mathbf{x}, y) - w_{\eta_1,\alpha_2,\sigma_2}(\mathbf{x}, y) + w_{\eta_1,\alpha_2,\sigma_2}(\mathbf{x}, y) - w_{\eta_2,\alpha_2,\sigma_2}(\mathbf{x}, y)| \\
&\quad \left|y - \mathbf{x}^T\theta_2\right||x_i| \\
&\leq (|w_{\eta_1,\alpha_1,\sigma_1}(\mathbf{x}, y) - w_{\eta_1,\alpha_2,\sigma_2}(\mathbf{x}, y)| + |w_{\eta_1,\alpha_2,\sigma_2}(\mathbf{x}, y) - w_{\eta_2,\alpha_2,\sigma_2}(\mathbf{x}, y)|) \\
&\quad \left|y - \mathbf{x}^T\theta_2\right||x_i|.
\end{aligned} \tag{2.7}$$

Since $m(x)$ has a bounded first derivative, we have $\left|m^{(1)}(x)\right| < K$ for all x, for some finite constant K. Since $m(x)$ is an absolute continuous function, using

the Mean Value Theorem, for any x_1, x_2 we obtain

$$|m(x_1) - m(x_2)| = m^{(1)}(c)\,|x_1 - x_2| < K\,|x_1 - x_2|\,,\ c \in (x_1, x_2).$$

Then we have

$$\begin{aligned} |w_{\eta_1,\alpha_2,\sigma_2}(\mathbf{x}, y) - w_{\eta_2,\alpha_2,\sigma_2}(\mathbf{x}, y)| &= \left| m(\eta_1 \frac{\left|y - \mathbf{x}^T\alpha_2\right|}{\sigma_2}) - m(\eta_2 \frac{\left|y - \mathbf{x}^T\alpha_2\right|}{\sigma_2}) \right| \\ &< \frac{K}{\sigma_2}\left|y - \mathbf{x}^T\alpha_2\right|\,|\eta_1 - \eta_2|\,, \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} &|w_{\eta_1,\alpha_1,\sigma_1}(\mathbf{x}, y) - w_{\eta_1,\alpha_2,\sigma 2}(\mathbf{x}, y)| \\ &= \left| m(\eta_1 \frac{\left|y - \mathbf{x}^T\alpha_1\right|}{\sigma_1}) - m(\eta_1 \frac{\left|y - \mathbf{x}^T\alpha_2\right|}{\sigma_2}) \right| \\ &< K\eta_1 \left| \frac{\left|y - \mathbf{x}^T\alpha_1\right|}{\sigma_1} - \frac{\left|y - \mathbf{x}^T\alpha_2\right|}{\sigma_2} \right| \\ &\le K \left| \frac{\left|y - \mathbf{x}^T\alpha_1\right|}{\sigma_1} - \frac{\left|y - \mathbf{x}^T\alpha_1\right|}{\sigma_2} + \frac{\left|y - \mathbf{x}^T\alpha_1\right|}{\sigma_2} - \frac{\left|y - \mathbf{x}^T\alpha_2\right|}{\sigma_2} \right| \\ &\le K\left|y - \mathbf{x}^T\alpha_1\right| \left| \frac{1}{\sigma_1} - \frac{1}{\sigma_2} \right| + \frac{K}{\sigma_2}\left|(y - \mathbf{x}^T\alpha_1) - (y - \mathbf{x}^T\alpha_2)\right| \\ &\le \max(\frac{K}{(\sigma^*)}, \frac{K}{(\sigma^*)^2})(\left|y - \mathbf{x}^T\alpha_1\right|\,|\sigma_2 - \sigma_1| + \|\mathbf{x}\|\,\|\alpha_2 - \alpha_1\|). \end{aligned} \tag{2.9}$$

From (2.7), (2.8) and (2.9), we then have

$$\begin{aligned} &\left|w_{\beta_1}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y) - w_{\beta_2}(\mathbf{x}, y)\phi^i_{\theta_2}(\mathbf{x}, y)\right| \\ &< K_0\,|x_i|\left|y - \mathbf{x}^T\theta_2\right|(\|\mathbf{x}\|\,\|\alpha_2 - \alpha_1\| + \left|y - \mathbf{x}^T\alpha_2\right|\,|\eta_1 - \eta_2| + \\ &\quad \left|y - \mathbf{x}^T\alpha_1\right|\,|\sigma_2 - \sigma_1|), \end{aligned} \tag{2.10}$$

where $K_0 = \max(\frac{K}{(\sigma^*)}, \frac{K}{(\sigma^*)^2})$. Since $\|\theta_2 - \theta_1\|, \|\alpha_2 - \alpha_1\|, |\eta_1 - \eta_2| < \|\gamma_1 - \gamma_2\|$, from (2.5), (2.6) and (2.10), we now have

$$\begin{aligned}
\left|\psi^i_{\gamma_1}(\mathbf{x}, y) - \psi^i_{\gamma_2}(\mathbf{x}, y)\right| &< K_0 |x_i| \left|y - \mathbf{x}^T\theta_2\right| (\|\mathbf{x}\| \|\alpha_2 - \alpha_1\| + \left|y - \mathbf{x}^T\alpha_1\right| |\sigma_2 - \sigma_1| \\
&\quad + \left|y - \mathbf{x}^T\alpha_2\right| |\eta_1 - \eta_2|) + |x_i| \|\mathbf{x}\| \|\theta_2 - \theta_1\| \\
&< (K_0 |x_i| \left|y - \mathbf{x}^T\theta_2\right| \|\mathbf{x}\| + |x_i| \|\mathbf{x}\| + \\
&\quad K_0 |x_i| \left|y - \mathbf{x}^T\alpha_1\right| \left|y - \mathbf{x}^T\theta_2\right| + \\
&\quad K_0 |x_i| \left|y - \mathbf{x}^T\alpha_2\right| \left|y - \mathbf{x}^T\theta_2\right|) \|\gamma_2 - \gamma_1\| \\
&= L^i(\mathbf{x}, y) \|\gamma_2 - \gamma_1\|, \text{ every } \gamma_1, \gamma_2,
\end{aligned}$$

where $L^i(\mathbf{x}, y) = K_0 |x_i| \left|y - \mathbf{x}^T\theta_2\right| \|\mathbf{x}\| + |x_i| \|\mathbf{x}\| + K_0 |x_i| \left|y - \mathbf{x}^T\alpha_2\right| \left|y - \mathbf{x}^T\theta_2\right| + K_0 |x_i| \left|y - \mathbf{x}^T\alpha_1\right| \left|y - \mathbf{x}^T\theta_2\right|$. For every $\psi^i, i = 1, ..., p$, we have derived a Lipschitz condition, so for $\psi = (\psi^1, ..., \psi^p)^T$ we then have

$$\left|\psi^i_{\gamma_1}(\mathbf{x}, y) - \psi^i_{\gamma_2}(\mathbf{x}, y)\right| < L^i(\mathbf{x}, y) \|\gamma_2 - \gamma_1\|, \text{ every } \gamma_1, \gamma_2.$$

Consider the bracketing entropy relative to $L_r(P)$-norm

$$\|\psi_i\|_{P,r} = (P |\psi_i|^r)^{1/r}.$$

Use brackets of the type $[\psi^i_\gamma - \varepsilon L^i, \psi^i_\gamma + \varepsilon L^i]$ for γ ranging over a suitable chosen subset of Γ, and these brackets have $L_r(P)$-size $2\varepsilon \|L^i\|_{P,r}$. If γ ranges over a grid of mesh width ε over Γ, then the brackets $[\psi^i_\gamma - \varepsilon L^i, \psi^i_\gamma + \varepsilon L^i]$ ranges over $\mathcal{F}$. By the Lipschitz condition

$$\psi^i_{\gamma 1} - \varepsilon L \le \psi^i_{\gamma 2} \le \psi^i_{\gamma 1} + \varepsilon L, \text{ if } \|\gamma_2 - \gamma_1\| \le \varepsilon,$$

so we need as many brackets as we need balls of radius $\frac{\varepsilon}{2}$ to cover Γ, or we need less than $(\text{diam } \Gamma/\varepsilon)^{2p+2}$ cubes with size ε to cover parameter space Γ. If $P\left|L^i\right|^r < \infty$, then there exists a constant J, depending on Γ and p only, such that the bracketing numbers satisfy

$$N_{[\ ]}(\varepsilon, \mathcal{F}, L_r(P)) \leq J(\frac{\text{diam } \Gamma}{\varepsilon})^{2p+2}, \text{ every } 0 < \varepsilon < \text{diam } \Gamma.$$

Since all $\psi \in \mathcal{F}$ are continuous functions, so they are measurable. In order use to Theorem 19.4 (Glivenko-Cantelli) in van der Vaart (2000), it is now enough to verify that $P\left|L^i\right| < \infty$ for all $1 \leq i \leq p$. Then the class $\mathcal{F}$ is P-Glivenko-Cantelli. Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned} P(K_0|x_i|\left|y - \mathbf{x}^T\theta_2\right| \|\mathbf{x}\|) &\leq K_0P^{1/2}|x_i|^2 + K_0P^{1/2}(\left|y - \mathbf{x}^T\theta_2\right| \|\mathbf{x}\|)^2 \\ &\leq K_0P^{1/2}|x_i|^2 + K_0P^{1/4}(\left|y - \mathbf{x}^T\theta_2\right|)^4 + K_0P^{1/4}(\|\mathbf{x}\|)^4. \end{aligned} \tag{2.11}$$

Similarly,

$$P(K_0|x_i|\|\mathbf{x}\|) \leq K_0P^{1/2}|x_i|^2 + K_0P^{1/2}(\|\mathbf{x}\|)^2 \tag{2.12}$$

and

$$\begin{aligned} &P(K_0|x_i|\left|y - \mathbf{x}^T\alpha_2\right| \left|y - \mathbf{x}^T\theta_2\right|) \\ &\leq K_0P^{1/2}|x_i|^2 + K_0P^{1/2}(\left|y - \mathbf{x}^T\alpha_2\right| \left|y - \mathbf{x}^T\theta_2\right|)^2 \\ &\leq K_0P^{1/2}|x_i|^2 + K_0P^{1/4}(\left|y - \mathbf{x}^T\alpha_2\right|)^4 \\ &\quad + K_0P^{1/4}(\left|y - \mathbf{x}^T\theta_2\right|)^4. \end{aligned} \tag{2.13}$$

If A3 is satisfied, then the right hand sides of (2.11), (2.12) and (2.13) will all be finite, so we have $P|L^i| < \infty$. Thus the class $\mathcal{F}$ is P-Glivenko-Cantelli.

Theorem 2.1: *Assume that A1, A2 , A3 and A4 hold. Then any solution* $\hat{\theta}_n$ *of the estimating equation (2.4) converges in probability to* θ_0.

Proof: Denote

$$\Psi(\theta) = P\psi_{\theta,0} = P\phi_\theta$$

$$\Psi(\theta,\beta) = P\psi_{\theta,\beta} = P\phi_\theta w_\beta$$

$$\Psi_n(\theta) = \Psi_n(\theta,\beta) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\theta,\beta}$$

$$\Psi_n(\theta,\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\theta,\hat{\beta}}.$$

Note that $\Psi(\hat{\theta}_n) = (\Psi(\hat{\theta}_n) - \Psi_n(\hat{\theta}_n)) + \Psi_n(\hat{\theta}_n)$. Then if we can show that $\sup_{\theta\in\Theta}\|\Psi_n(\theta) - \Psi(\theta)\| = o_P(1)$, then $\left\|\Psi(\hat{\theta}_n)\right\| = o_P(1)$ since $\Psi_n(\hat{\theta}_n) = o_P(1)$. Then from Lemma 2.2, we have $\left\|\hat{\theta}_n - \theta_0\right\| = o_P(1)$. To show $\sup_{\theta\in\Theta}\|\Psi_n(\theta) - \Psi(\theta)\| = o_P(1)$, we factor it as follows

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\Psi_n(\theta,\hat{\beta}) - \Psi(\theta)\right\| \\
&= \sup_{\theta\in\Theta}\left\|\Psi_n(\theta,\hat{\beta}) - \Psi_n(\theta,\beta) + \Psi_n(\theta,\beta) - \Psi(\theta,\beta) + \Psi(\theta,\beta) - \Psi(\theta,0)\right\| \\
&\leq I_1 + I_2 + I_3,
\end{aligned}$$

where

$$I_1 = \sup_{\theta\in\Theta}\left\|\Psi_n(\theta,\hat{\beta}) - \Psi_n(\theta,\beta)\right\|$$

$$I_2 = \sup_{\theta\in\Theta}\|\Psi_n(\theta,\beta) - \Psi(\theta,\beta)\|$$

$$I_3 = \sup_{\theta\in\Theta}\|\Psi(\theta,\beta) - \Psi(\theta,0)\|.$$

From Lemma 2.3, we know that $\mathcal{F}$ is a P-Glivenko-Cantelli class, so $I_2 \xrightarrow{as} 0$.

For I_1,

$$\begin{aligned} I_1 &= \sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi_\theta(\mathbf{x}_i,y_i)(w_{\hat\beta}(\mathbf{x}_i,y_i)-w_\beta(\mathbf{x}_i,y_i))\right\| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}\left\|\phi_\theta(\mathbf{x}_i,y_i)(w_{\hat\beta}(\mathbf{x}_i,y_i)-w_\beta(\mathbf{x}_i,y_i))\right\| \\ &= \frac{1}{n}\sum_{i=1}^{n}(w_{\hat\beta}(\mathbf{x}_i,y_i)-w_\beta(\mathbf{x}_i,y_i))\sup_{\theta\in\Theta}\left\|\phi_\theta(\mathbf{x}_i,y_i)\right\|. \end{aligned}$$

A Taylor expansion of order one yields

$$\begin{aligned} |w_{\hat\beta}(\mathbf{x}_i,y_i)-w_\beta(\mathbf{x}_i,y_i)| &= |\nabla m(\beta+(\beta-\hat\beta)t)(\hat\beta-\beta)| \\ &\le \left\|\nabla m(\beta+(\beta-\hat\beta)t)\right\|\left\|\hat\beta-\beta\right\|, \end{aligned}$$

where $\nabla m(\beta)$ is the gradient of $m(\beta)$, and denote $\beta' = \beta+(\beta-\hat\beta)t$ with $t\in(0,1)$. Then we have

$$\begin{aligned} \nabla m(\beta') &= \left(\frac{\partial m(\eta'\frac{y-\mathbf{x}^T\alpha'}{\sigma'})}{\partial\eta'}, \frac{\partial m(\eta'\frac{y-\mathbf{x}^T\alpha'}{\sigma'})}{\partial\alpha'}^T, \frac{\partial m(\eta'\frac{y-\mathbf{x}^T\alpha'}{\sigma'})}{\partial\sigma'}\right)^T \\ &= m^{(1)}_{\beta'}(\mathbf{x},y)\left(\frac{y-\mathbf{x}^T\alpha'}{\sigma'}, \frac{1}{\sigma'}\mathbf{x}^T, -(y-\mathbf{x}^T\alpha')(\sigma')^2\right)^T. \end{aligned}$$

Since $\hat\beta-\beta = o_P(1)$ and $\left\|\nabla m(\beta')\right\|$ is finite, we have $w_{\hat\beta}(\mathbf{x}_i,y_i)-w_\beta(\mathbf{x}_i,y_i) = o_P(1)$. From assumption A3, we also have $\sup_{\theta\in\Theta}\left\|\phi_\theta(\mathbf{x},y)\right\|^2 < \infty$, and thus we obtain $I_1 \xrightarrow{p} 0$.

For I_3,

$$\begin{aligned} I_3 &= \sup_{\theta\in\Theta} \|P\phi_\theta(\mathbf{x},y)(w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y))\| \\ &\leq \sup_{\theta\in\Theta} P\,\|\phi_\theta(\mathbf{x},y)(w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y))\| \\ &\leq \sup_{\theta\in\Theta} P^{1/2}\,\|\phi_\theta(\mathbf{x},y)\|^2\, P^{1/2}|w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y)|^2 \\ &= P^{1/2}|w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y)|^2 \sup_{\theta\in\Theta} P^{1/2}\,\|\phi_\theta(\mathbf{x},y)\|^2\,, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Also

$$w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y) = 1 - m(\eta\frac{y-\mathbf{x}^T\alpha}{\sigma}) \xrightarrow{p} 0,$$

since $\eta = o_P(1)$, and $m(0) = 1$, and $|w_\beta(\mathbf{x},y) - w_{\beta_0}(\mathbf{x},y)|^2$ is bounded by 1. Now use the Dominated Convergence Theorem to obtain

$$\lim_{n\to\infty} P^{1/2}|w_\beta(\mathbf{x}) - w_{\beta_0}(\mathbf{x})|^2 = P^{1/2} \lim_{n\to\infty} |w_\beta(\mathbf{x}) - w_{\beta_0}(\mathbf{x})|^2 \xrightarrow{p} 0.$$

Also,

$$\sup_{\theta\in\Theta} P\,\|\phi_\theta(\mathbf{x},y)\|^2 = \sup_{\theta\in\Theta} P\left\|(y-\mathbf{x}^T\theta)\mathbf{x}\right\|^2 \leq P\,\|\mathbf{x}\|^2\,,$$

and from assumption A3 we have $\sup_{\theta\in\Theta} P\,\|\phi_\theta(\mathbf{x},y)\|^2 < \infty$. Then we have $I_3 \to 0$, and this completes the proof. ◀

The asymptotic normality of $\hat{\theta}_n$ is established in Theorem 2.2, which makes use of the following assumptions:

A5 $E_{G_0}(\|\mathbf{x}\|^8) < \infty$ and F_0 has finite eighth moment.

A6 $\hat{\theta}_n$ satisfies $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$.

Theorem 2.2: *Assume that A1 to A6 hold. Then we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $V^{-1}P\psi_{\theta_0}\psi_{\theta_0}^T(V^{-1})^T$, *where* $V^{-1} = P^{-1}(m(\eta\frac{y-\mathbf{x}^T\alpha}{\sigma})XX^T)$ *and* $P\psi_{\theta_0}\psi_{\theta_0}^T = P(m^2(\eta\frac{y-\mathbf{x}^T\alpha}{\sigma})(Y - X^T\theta_0)^2XX^T)$. *In particular if* $\eta = 0$, *then the covariance matrix equals to* $\sigma_0^{-1}E^{-1}(XX^T)$, *which means the RECWLS estimator is asymptotically efficient under normal errors.*

2.4 Monte Carlo studies

In this section a Monte Carlo Study was carried out to examine finite-sample efficiency and robustness properties of the proposed estimator, RECWLS. Two robust estimators, LMSE (least median of squares estimator) and LTSE (least trimmed squares estimator), were used as initial estimators for θ_{0n}. Further, the scales used to standardize residuals were the standardized MAD (median absolute deviation).

The following weight function $m(x)$ was used in the simulation:

$$m(x) = \frac{1}{(1+x^2)^5}. \tag{2.14}$$

For comparison purposes, I computed the following estimators:

1. Least squares (LS).

2. Least median of squares (LMS).

3. Least trimmed squares (LTS).

4. One-step weighted least squares with cut-off value $t_0 = 2.5$, starting from the LMSE (WLS-LMS).

5. Same as above, starting from the LTS (WLS-LTS).

6. REWLSE with hard-rejection weight $w(u) = I(u < 1)$ and $\eta = 2.5$, starting from LMSE (REWLS-LMS).

7. Same as above, starting from LTSE (REWLS-LTS).

8. One-step weighted least squares with weight function

$$w(\mathbf{x}_i, y_i) = m(\hat{\eta}\,\|r_i\|) = \frac{1}{(1 + (\hat{\eta}\left\|\frac{y_i - \mathbf{x}_i^T\theta_{0n}}{S_n}\right\|)^2)^5}$$

   starting from the LMSE (RECWLS-LMS).

9. Same as above, starting from LTSE (RECWLS-LTS).

10. For the case of linear regression with t-distributed errors in Section 2.4.2, I considered the corresponding maximum likelihood estimator.

2.4.1 Efficiencies with normal errors

I considered regression models with intercept, normal covariates and normal errors. Specifically, let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be a random model that follows the linear model (2.1) with $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip-1})^T$ and such that $(x_{i1}, ..., x_{ip-1})^T$ has a multivariate normal $N_{p-1}(\mu, \Sigma)$ distribution. Since all estimators are regression, affine and scale equivariant, without loss of generality I took $\mu = 0$, $\Sigma = I$ and $\theta_0 = 0$.

Samples with $n = 20, 50, 100, 200, 500, 1000$ and $p = 2, 5$ were considered. For each value of p and n, I generated 1000 samples, and for each estimator θ_n I computed the relative mean squared efficiency (REF) with respect to the LSE:

$$\text{REF} = \frac{\sum_{i=1}^{1000}(\|\theta_{ni}^{LS} - \theta_0\|)^2}{\sum_{i=1}^{1000}(\|\theta_{ni} - \theta_0\|)^2} = \frac{\sum_{i=1}^{1000}(\|\theta_{ni}^{LS}\|)^2}{\sum_{i=1}^{1000}(\|\theta_{ni}\|)^2},$$

where θ_{ni} and θ_{ni}^{LS} are the i-th generated values of θ_n and the LSE, respectively. Since the LSE is the most efficient estimator under normal models, an estimator with REF close to 1 would consider to be very efficient.

Table 2.1: *REF w.r.t. LSE for normal errors and* $p = 2$

estimator	n					
	20	50	100	200	500	1000
LMS	0.22	0.18	0.15	0.13	0.10	0.08
WLS-LMS	0.60	0.69	0.71	0.76	0.79	0.80
REWLS-LMS	0.60	0.69	0.74	0.81	0.88	0.93
RECWLS-LMS	0.79	0.96	0.98	0.99	1.00	1.00
LTS	0.24	0.17	0.13	0.12	0.09	0.09
WLS-LTS	0.60	0.68	0.70	0.75	0.79	0.83
REWLS-LTS	0.60	0.68	0.72	0.81	0.88	0.93
RECWLS-LTS	0.74	0.96	0.98	0.99	1.00	1.00

Tables 2.1 and 2.2 show the comparison of REF of different estimators for normal errors when $p = 2, 5$. As we can see from these tables, the initial estimators LMS and LTS are not very efficient: their relative efficiencies are decreasing as the sample size increases. When the number of parameters p increases, the REF decreases for the LMS, but it doesn't change much for the LTS. Further, the LTS is more efficient than the LMS estimator, especially when p is larger. But when n reaches 1000, both of them have very low efficiencies, about 7%. As I mentioned in Section 2.1, LMS estimators'

Table 2.2: *REF w.r.t. LSE for normal errors and* $p = 5$

estimator	n					
	20	50	100	200	500	1000
LMS	0.15	0.17	0.16	0.13	0.09	0.07
WLS-LMS	0.24	0.40	0.57	0.67	0.73	0.77
REWLS-LMS	0.24	0.40	0.59	0.72	0.84	0.90
RECWLS-LMS	0.28	0.61	0.95	0.99	0.99	1.00
LTS	0.24	0.18	0.15	0.12	0.08	0.07
WLS-LTS	0.34	0.41	0.57	0.66	0.72	0.76
REWLS-LTS	0.34	0.41	0.59	0.71	0.83	0.89
RECWLS-LTS	0.41	0.62	0.94	0.99	0.99	1.00

asymptotic relative efficiency with respect to the LSE is 0, and for the LTS's it is about 7%.

It is noticeable that even using a simple one-step weighted least squares with non-adaptive cut-off value greatly increases the efficiency. For $p = 2$ and sample size 100, WLS-LMS only has a REF about 69%, and for $p = 5$, it drops to 41%. We know ultimately when $n \to \infty$, the asymptotic relative efficiency is 0. For finite sample sizes, we can see that the REF of the weighted LSE with fixed threshold values are not that high, especially with the increasing of dimension p.

REWLS based estimators (i.e., the WLSE with adaptive threshold values) have very similar REF with WLSE with fixed threshold values, especially when the sample size is small ($n < 100$). The asymptotically efficiency of REWLS is 1, so we can see as n increases, the REF increases pretty fast.

Obviously, all estimators have better efficiencies with larger n, and REWLS-LMS showed a slightly better efficiency than WLS-LMS. For the proposed estimators, RECWLS-LMS greatly increases the efficiency. When $p = 2$,

RECWLS with sample size $n = 50$ and higher attains almost full efficiency, and this is also the case for $n = 100$ and higher when $p = 5$.

By comparing REWLS and RECWLS when $p = 2$, the efficiency of RECWLS with sample size $n = 20$ is almost as high as REWLS with sample size $n = 200$. When $p = 5$, RECWLS-LMS with sample size $n = 100$ have higher efficiencies than REWLS with sample size $n = 1000$. Implementing the LMS or LTS as initial estimators has similar effects in efficiencies, and when p increases, the efficiencies decrease as we have more parameters to estimate.

The proposed estimator, RECWLS, shows great efficiency properties overall. When $p = 2$, its REF reaches 96% even when sample size is $n = 50$, and it reaches about 95% when $n = 100$ for $p = 5$.

In summary, RECWLS has shown the best efficiency among the estimators considered under models with normal errors; its REF reaches close to 1 very fast. So, for small sample size data sets, the proposed estimator may be more preferred one to use in applications.

2.4.2 Efficiencies with heavy-tailed errors

The simulation settings in this section are the same as in Section 2.4.1, except that the errors e_i's were generated here according to a t distribution with 3 degrees of freedom. The relative efficiencies were calculated with respect to the maximum likelihood estimator in this case.

Samples with $n = 20, 50, 100, 200, 500, 1000$ and $p = 2, 5$ were considered. For each value of p and n, I generated 1000 samples, and for each estimator

θ_n, I computed the relative mean squared efficiency with respect to the MLE:

$$\text{REF} = \frac{\sum_{i=1}^{1000}(\left\|\theta_{ni}^{MLE} - \theta_0\right\|)^2}{\sum_{i=1}^{1000}(\|\theta_{ni} - \theta_0\|)^2} = \frac{\sum_{i=1}^{1000}(\left\|\theta_{ni}^{MLE}\right\|)^2}{\sum_{i=1}^{1000}(\|\theta_{ni}\|)^2},$$

where θ_{ni} and θ_{ni}^{MLE} are the i-th generated values of θ_n and the MLE, respectively. Since MLE's are the most efficient estimators under student t errors, an estimator with REF close to 1 would consider to be very efficient.

Table 2.3: *REF w.r.t. MLE for Student errors with 3 d.f. and* $p = 2$

estimator	n					
	20	50	100	200	500	1000
LSE	0.55	0.54	0.53	0.51	0.47	0.51
LMS	0.29	0.30	0.24	0.23	0.17	0.14
WLS-LMS	0.71	0.81	0.77	0.83	0.83	0.79
REWLS-LMS	0.71	0.81	0.79	0.84	0.86	0.81
RECWLS-LMS	0.76	0.83	0.83	0.84	0.79	0.77
LTS	0.60	0.54	0.43	0.45	0.42	0.40
WLS-LTS	0.74	0.81	0.78	0.81	0.84	0.85
REWLS-LTS	0.74	0.81	0.80	0.81	0.87	0.86
RECWLS-LTS	0.78	0.83	0.84	0.83	0.79	0.77

Tables 2.3 and 2.4 show the comparison of mean squared efficiencies of some estimators for student t errors with 3 d.f. when $p = 2, 5$. Unlike in the previous section, with the increase of sample n, most estimators' efficiencies do not increase, and most of them show a better efficiency when $n = 200$ than $n = 1000$. This is probably because with a higher sample size, the heavier tail of error distribution becomes more obvious.

All reweighed estimators show a better efficiency than the initial estimators, and LTS estimators show a significantly better efficiency than LMS estimators. Again WLS estimators and REWLS estimators show very similar

Table 2.4: *REF w.r.t. MLE for Student errors with 3 d.f. and* $p = 5$

estimator	n					
	20	50	100	200	500	1000
LSE	0.58	0.55	0.52	0.47	0.50	0.49
LMS	0.20	0.26	0.25	0.22	0.15	0.11
WLS-LMS	0.32	0.58	0.73	0.80	0.76	0.75
REWLS-LMS	0.32	0.58	0.74	0.81	0.80	0.78
RECWLS-LMS	0.34	0.74	0.86	0.83	0.79	0.76
LTS	0.55	0.51	0.46	0.45	0.30	0.21
WLS-LTS	0.41	0.59	0.70	0.80	0.78	0.74
REWLS-LTS	0.41	0.59	0.71	0.82	0.81	0.78
RECWLS-LTS	0.46	0.74	0.85	0.84	0.80	0.76

efficiencies. As the sample size n increases, the efficiencies of RECWLS estimators start to increase and then decrease. It appears that RECWLS estimators are more efficient than REWLS when $n < 200$.

In summary, under models with t errors, RECWLS still have a pretty good performance in efficiencies compared to other estimators, especially when the sample size n is small.

2.4.3 Model with normal errors and some fraction of outlier contamination

As in Section 2.4.1, a model with normal-error and normal-covariates were considered, But now h observations in each sample were replaced by identical outliers of the from $(\mathbf{x}_0, y_0)$. Because of the sphericity of the normal distribution, without loss of generality I took $\mathbf{x}_0 = (1, x_0, 0, ..., 0)^T$. I chose $x_0 = 1, 10$, which correspond to low and high leverage outliers, respectively, and varied y_0 in the grid $\{0.1jx_0 : j \text{ is a positive integer}\}$. For j, I considered integers

from 1 to 100, and then randomly chose 50 j from 101 to 1000, and another randomly chosen 50 j from 1001 to 10000. I only considered a small sample size $n = 50$, and I took $p = 2$ and $h = 3, 5, 8, 10$. For each estimator θ_{1n} and each value of h, x_0 and y_0, I estimated the mean squared error based on 1000 Monte Carlo replications, MSE $(\theta_{1n}, h, \mathbf{x}_0, y_0)$. Tables 4.19 to 4.12 report the values of $\max_{y_0}$ MSE$(\theta_{1n}, h, \mathbf{x}_0, y_0)$, which represents the worst performance of each estimator for that leverage and that number of outliers.

Table 2.5: *Maximum* MSE *with outliers with* $x_0 = 1$ *(not high leverage outliers)*

estimator	h			
	3	5	8	10
LMS	0.36	0.53	1.03	1.71
WLS-LMS	0.10	0.18	0.48	0.93
REWLS-LMS	0.10	0.18	0.48	0.93
RECWLS-LMS	0.12	0.18	0.39	0.75
LTS	0.39	0.68	1.63	2.87
WLS-LTS	0.09	0.48	1.33	2.47
REWLS-LTS	0.09	0.48	1.33	2.47
RECWLS-LTS	0.14	0.34	1.06	2.13

Table 2.6: *Maximum* MSE *with outliers with* $x_0 = 10$ *(high leverage outliers)*

estimator	h			
	3	5	8	10
LMS	0.35	0.53	1.14	1.91
WLS-LMS	0.19	0.38	0.95	1.67
REWLS-LMS	0.19	0.38	0.95	1.67
RECWLS-LMS	0.16	0.28	0.77	1.45
LTS	0.40	0.69	1.61	2.86
WLS-LTS	0.22	0.48	1.31	2.45
REWLS-LTS	0.22	0.48	1.32	2.45
RECWLS-LTS	0.18	0.35	1.05	2.11

Tables 2.5 and 2.6 show the maximum MSE for the normal error model with outliers $x_0 = 1, 10$. Initial estimators, LMS and LTS, already show great robustness properties since they are already robust, and WLS estimators and REWLS estimators have very similar performance, both increased the robustness significantly compared to initial estimators. When $x_0 = 1$, which means that the contaminated observations do not have high leverage, RECWLS estimators have a slightly larger MSE than REWLS estimators when $h = 3$. This is because the proposed estimators are more "gentle" than the threshold-value based estimators since its weight function is continuous. When h increases, RECWLS estimators show better robustness properties than the others, and the differences are increasing with increasing of h. When $x_0 = 10$, this means that the contaminated observations have a high leverage, the proposed RECWLS estimators in this case show better robustness properties even when the proportion of outliers is low ($h = 3$). Also, LMS-based estimators are more robust than LTS-based estimators judging from the MSE values.

In summary, the proposed RECWLS estimators show better robustness properties than the initial estimators and threshold-value based weighted estimators for finite sample sizes.

2.5 Real data analysis

I analyzed the aircraft data set given in Gray (1985). It records five measured characteristics of 23 single-engine aircrafts built over the years 1947-1979 recorded by the Office of Naval Research. The response variable is the cost (unit \$100,000) and the predictor variables are aspect ratio, lift-to-drag ratio, weight of plane (in pounds) and maximal thrust. There are 2 extreme outliers,

and there may be other potential moderate outliers and leverage points in the dataset. So it is a good dataset to illustrate robust estimators in a real world setting.

The LSE, REWLS and RECWLS estimators with LTS as the initial estimator, MM-estimators (Yohai, 1987) and S-estimators (Rousseeuw and Yohai, 1984) were calculated to compare the performance of these robust estimators. Table 2.7 shows the comparison of coefficients. As we can see from Table 2.7,

Table 2.7: *Coefficient estimates for Aircraft data*

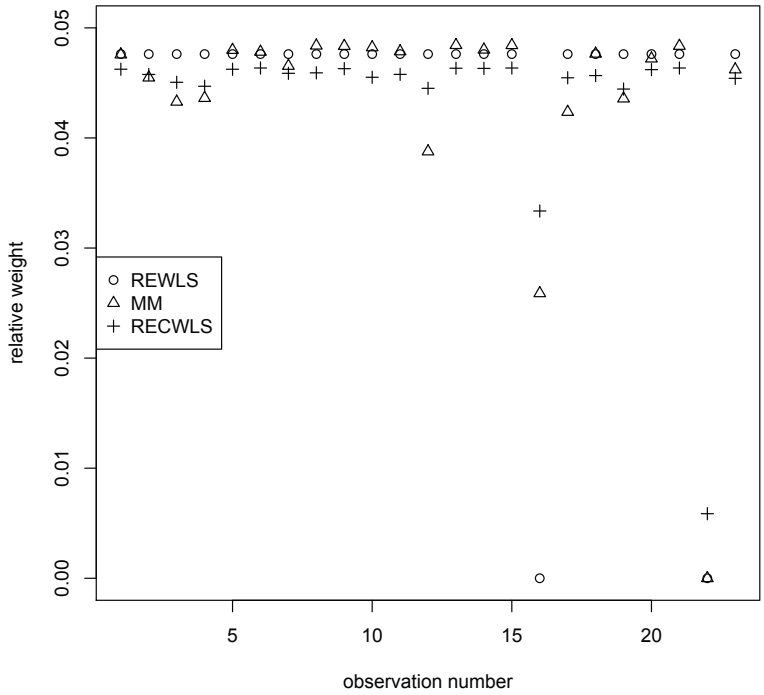
	Intercept	Aspect Ratio	Lift-to-drag Ratio	Weights ($\times 10^3$)	Thrust ($\times 10^3$)
LSE	-3.79 (10.12)	-3.85 (1.76)*	2.49 (1.19)	3.50 (0.48)**	-1.95 (0.50)**
REWLS	9.50 (5.58)	-3.05 (0.92)**	1.21 (0.65)	1.38 (0.39)**	-0.55 (0.33)**
S	13.37 (4.47)**	-4.02 (1.16)**	1.54 (0.44)**	1.70 (0.34)**	-0.98 (0.29)**
RECWLS	3.15 (7.71)	-3.41 (1.31)*	1.95 (0.89)*	2.39 (0.46)**	-1.24 (0.42)**
MM	6.14 (8.31)	-3.23 (0.86)**	1.67 (0.70)*	1.92 (0.79)*	-0.93 (0.51)

*standard errors in parentheses, two sided significance at level 0.05 (*) or level 0.01 (**)

all robust estimators are very different from the LSE. Among the robust estimators, the S-estimator and the REWLS estimator show a similar pattern, except that the variable lift-to-drag is not significant for the REWLS estimator (p-value is 0.08). The MM-estimator and the proposed estimator RECWLS are similar in performance, and the variable lift-to-drag is significant in both cases, and the variable thrust is not significant for the MM-estimator (p-value is 0.08).

The plots in Figure 2.1 exhibit the relative weights assigned to each observation under different estimators. Since the REWLS uses a threshold to assign 0/1 weight to observations, it gave 0 weight to observation #16 and #22, and the REWLS actually has exactly the same estimates as with its ini-

Figure 2.1: *Weights for each of the 23 observations in the Aircraft data*



tial estimator LTS in this case. All estimators downweighted observation #22, and the MM-estimator and the proposed estimator did not totally eliminate the observation #16, unlike the REWLS, but gave it a smaller weight. Both the MM-estimator and the proposed estimator also downweighted a little the moderate outliers and leverage points: #3, #4, #12 and #19.

Chapter 3

Logistic Regression Model

3.1 Preliminaries: the logistic regression model and existing estimators

Consider a random sample of observations $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i$ is a vector of p explanatory variables and $y_i \in \{0, 1\}$ is a dichotomous response variable, and assume the probability of positive response $\pi_i = P(y_i = 1|\mathbf{x}_i)$ is linked with the covariates via the relationship

$$g(\pi_i) = \mathbf{x}_i^T \beta,$$

where the link function is a quantile function of some probability distribution. For example, when the link function g is the logit link function $g(\pi_i) = \log\{\pi_i/(1-\pi_i)\}$, then the corresponding model is known as logistic regression. Another common link is $g = \Phi^{-1}$, the inverse of the standard normal distribution function, and the resulting model is called probit regression. Here I focus on estimating the regression parameter β for a given g.

The maximum likelihood estimator (MLE) of β of binomial regression model is defined as

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} d(y_i, \mathbf{x}_i, \beta),$$

where $d(y_i, \mathbf{x}_i, \beta) = -2y_i \log \pi(\mathbf{x}_i^T\beta) - 2(1-y_i)\log\{1-\pi(\mathbf{x}_i^T\beta)\}$ is the deviance and $\pi = g^{-1}$ is the inverse link function. Taking the derivative w.r.t. β, the MLE $\hat{\beta}$ satisfy the estimating equation

$$\sum_{i=1}^{n} \frac{\{y_i - \pi(\mathbf{x}_i^T\hat{\beta})\}\pi'(\mathbf{x}_i^T\hat{\beta})}{\pi(\mathbf{x}_i^T\hat{\beta})\{1-\pi(\mathbf{x}_i^T\hat{\beta})\}}\mathbf{x}_i = 0, \tag{3.1}$$

and if the link function is logit link, then the estimating equation becomes

$$\sum_{i=1}^{n}(y_i - \frac{e^{\mathbf{x}^T\beta}}{1+e^{\mathbf{x}^T\beta}})\mathbf{x}_i = 0.$$

For logistic and probit models the objective function in (3.1) is convex, so that if a finite minimizer exists, it is the unique solution of (3.1). Albert and Anderson (1984) showed that $\hat{\beta}$ exists if and only if the data overlap, in the sense that no hyperplane in the covariate space separates response from non-responses.

From the estimating equation (3.1) we can see that the MLE is affected by (i) observations with extreme values of explanatory variables $\mathbf{x}_i$'s, (ii) observations which are poorly predicted by the chosen model, usually with unreasonably large value of $y - \pi(\mathbf{x}^T\beta)$, and (iii) model misspecification.

Bianco and Yohai (1996) proposed robust estimators by controlling the deviances. However, in order to obtain theoretically unbiased estimators, an additional bias-correction term has to be added, and it makes the computation of their estimator very complicated and the estimator itself isn't straightfor-

ward.

Künsch et al.(1989) proposed Mallows-type estimators by controlling covariates and residuals in the estimating equation independently. For the logistic regression case, they were defined as solutions of

$$\sum_{i=1}^{n} w(\mathbf{x}_i;\hat{\eta})\Phi_b(y_i - \pi(\mathbf{x}_i^T\hat{\beta}) - c(\pi(\mathbf{x}_i^T\hat{\beta}), b))\mathbf{x}_i = 0,$$

where $\hat{\eta}$ is a vector of nuisance parameters (location, scatter estimate of covariates), and Φ_b is usually taken as Huber's function $\Phi_b(t) = (-b) \vee (t \wedge b)$, $c(t, b)$ is a bias-correcting function expressed as

$$c(t,b) = \begin{cases} b\pi(t)/\{1-\pi(t)\} - \pi(t) & \text{if } t < 0, b < 1 - \pi(t) \\ 1 - \pi(t) - b\{1-\pi(t)\}/\pi(t) & \text{if } t > 0, b < \pi(t) \\ 0 & \text{otherwise.} \end{cases}$$

The covariate weights $w(\mathbf{x}_i;\hat{\eta})$ usually depends only on continuous covariates. If we write $\mathbf{x}_i^T = (\mathbf{u}_i^T, \mathbf{z}_i^T)$, where $\mathbf{u}_i \in \mathbb{R}^{p-q}$ are the categorical covariates and $\mathbf{z}_i \in \mathbb{R}^q$ are the continuous covariates, then the weights are typically of the form $w(\mathbf{x}_i;\hat{\eta}) = \omega((\mathbf{z}_i - \hat{\mu})^T\hat{\Sigma}^{-1}(\mathbf{z}_i - \hat{\mu})/t)$, with $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ a non-increasing function, $\hat{\mu}$ and $\hat{\Sigma}$ are robust estimators of location and scatter of the $\mathbf{z}_i$'s, and t is a threshold value (usually $t = \chi^2_{q,1-\alpha}$ for some $\alpha \in (0,1)$).

The initial robust location and scale estimator of continuous covariates, $\hat{\mu}$ and $\hat{\Sigma}$, can be computed using the Minimum Covariance Determinant (MCD) methods. MCD is one of the first affine equivariant and highly robust estimators of multivariate location and scatter. It finds the $h(> n/2)$ observations

$x_{(i)}$ whose classical covariance matrix

$$V = \frac{1}{h}\Sigma_i(x_{(i)} - t)(x_{(i)} - t)^T$$

has the lowest possible determinant, where $t = \bar{x}$, the average of those h points.

As we can see, the covariate weight w and "residual" weight Φ_b are independent, and it will make the resulting estimators less efficient since the estimating equation will downweight observations with extreme covariates even if they are well-fitted.

Gervini (2005) proposed a class of robust adaptive weighted maximum likelihood estimators for binary regression models. The adaptive weights he chose are based on adaptive cut-off thresholds to control observations with extreme covariates. He showed that the estimators based on adaptive thresholds are more efficient than those based on non-adaptive thresholds under the clean model and have comparable robustness under contaminated models.

3.2 RECMLE estimators

Similar to the strategy used in constructing the RECWLS estimator in the linear regression model of Chapter 2, here I construct a new class of weighted maximum likelihood estimators with a continuous weight function based on an estimator of a nuisance parameter as a function of the Kolmogorov-Smirnov statistic, without using cut-off values. I shall refer these estimators as RECMLE (*robust and efficient continuous maximum likelihood estimators*).

First construct two estimators, $\hat{\mu}^{(0)}$ and $\hat{\Sigma}^{(0)}$, that are initial location and scatter estimators of the $\mathbf{z}_i$'s. Then the squared Mahalanobis distances of the

$\mathbf{z}_i$'s can be defined as $m_i^2 = (\mathbf{z}_i - \hat{\mu}^{(0)})^T(\hat{\Sigma}^{(0)})^{-1}(\mathbf{z}_i - \hat{\mu}^{(0)})$, and the empirical distribution function of m_i^2 can be defined as

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(|m_i^2| \leq t).$$

When the $\mathbf{z}_i$'s are normally distributed, F_n converges to $F_{\chi_q^2}$ (χ_q^2 distribution function). Then the proportion of outliers in the covariates can be estimated by (Gervini, 2005)

$$\begin{aligned}\alpha_n &= \sup_{t \geq F_{\chi_q^2}^{-1}(1-\delta)} \{F_{\chi_q^2}(t) - F_n(t)\}_+ \\ &= \max_{i \geq i_0}\{F_{\chi_q^2}(m_{(i)}^2) - \frac{i-1}{n}\}_+,\end{aligned}$$

where $\{\cdot\}_+$ denotes the positive part, δ determines the length of the tail ($\delta = 0.25$ is a reasonable choice) and $i_0 = \min\{i : m_i^2 \geq F_{\chi_q^2}^{-1}(1-\delta)\}$. When $\left|F_{\chi_q^2}(t) - F_n(t)\right|$ is large for a large t, it means that the outliers (observations with extreme covariates) are present in the sample. Then an adaptive threshold can be defined as

$$\begin{aligned}t_n &= F_n^{-1}(1-\alpha_n) \\ &= m_{(n-[n\alpha_n])}^2.\end{aligned}$$

The adaptive threshold-type estimators were proposed by Gervini (2005). Specifically, he proposed a Mallows-type estimator with weights $w(\mathbf{x}_i; \hat{\eta}) = \omega(m_i^2/t_n)$, and they are essentially weighted maximum likelihood estimators.

The weight function I propose here can be defined as

$$w(x_i; \alpha_n) = m(\alpha_n m_i^2),$$

similar to the weight function defined in Section 2.2. I assume that m is an absolutely continuous non-increasing mapping from $\mathbb{R}^+$ to $(0, 1]$ such that $m(0) = 1, \sup_{x>0}[xm(x)] < \infty$, and first derivative is bounded with $m^{(1)}(0) = 0$. Define an objective function

$$\psi_{\beta,\eta}(\mathbf{x}, y) = w_\eta(\mathbf{x})\phi_\beta(\mathbf{x}, y),\ \mathbf{x}, \beta \in \mathbb{R}^p,\ y \in \{0, 1\}, \tag{3.2}$$

where $\eta = (\alpha, \mu, \Sigma)$ is a set of nuisance parameters (location, scale and goodness of fit of covariates), an adaptive weight function $w_\eta(\mathbf{x}) = m(\alpha(\mathbf{z} - \mu)^T\Sigma^{-1}(\mathbf{z} - \mu))$ and $\phi_\beta(\mathbf{x}, y) = (y - \pi(\mathbf{x}^T\beta))\mathbf{x}$, with $\mu \in \mathbb{R}^a, \Sigma$ is a $q \times q$ real matrix and $\mathbf{x}^T = (\mathbf{u}^T, \mathbf{z}^T)$. Then I define my adaptive estimator $\hat{\beta}_n$ of β as the solution to the estimating equation

$$\sum_{i=1}^{n} \psi_{\beta,\hat{\eta}}(\mathbf{x}_i, y_i) = 0,$$

where $\hat{\eta}$ is a consistent estimator of $\eta = (\alpha, \mu, \Sigma)$.

3.3 Asymptotic properties

This section studies asymptotic properties of the proposed estimator $\hat{\beta}_n$ defined in the previous section. I will show that, under some general assumptions on the moments of explanatory variables, the estimator is asymptotically consistent.

Assume that β_0, μ_0 and Σ_0 are the "true values" of β, μ and Σ, respectively, and the independent sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ follows the logistic

model $Pr(y_i = 1) = \pi(\mathbf{x}_i^T \beta_0)$, $i = 1, ..., n$. We define functions

$$\mathbb{P}_n \psi_\beta = \Psi_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \psi_{\beta,\hat{\eta}}(\mathbf{x}_i, y_i)$$

and

$$\Psi(\beta) = P\psi_{\beta,0} = P\phi_\beta,$$

where $\psi_{\beta,0}$ is obtained by (3.2) with η replaced by $\eta_0 = (0, \mu_0, \Sigma_0)$, and P denotes the (unknown) joint probability distribution of the $(\mathbf{x}, y)$'s.

The consistency of $\hat{\beta}_n$ is established in Theorem 3.1, which makes use of the results of Lemma 3.1, 3.2 and 3.3 stated below. The following assumptions are needed to prove lemmas and the theorem:

B1 $\hat{\mu} \xrightarrow{p} \mu_0$ and $\hat{\Sigma} \xrightarrow{p} \Sigma_0$.

B2 $E(\mathbf{X}\mathbf{X}^T)$ is nonsingular.

B3 $E_{G_0}(\|\mathbf{x}\|^4) < \infty$.

B4 the weight function $m(x)$ is continuous, has bounded first derivative, $m(0) = 1$ and $m^{(1)}(0) = 0$.

For almost all popular initial robust estimators, like the MCD (Minimum Covariance Determinant) used in the simulation studies, B1 is satisfied. In the lemmas and the theorem stated below, the asymptotic properties are understood to be as $n \to \infty$. The proof of Lemma 3.1 below is given in Gervini (2005).

Lemma 3.1: *If B1 is satisfied, then* $\alpha_n = o_p(1)$.

Lemma 3.2: *If B2 is satisfied, then* $\|\Psi(\beta_n)\| \to 0$ *implies* $\|\beta_n - \beta_0\| \to 0$ *for any sequence* $\{\beta_n\} \in \Theta$.

Proof: Consider

$$\begin{aligned}\|\Psi(\beta_n)\| &= \|P\psi_{\beta_n,0}\| \\ &= \|P\phi_{\theta_n}\| \\ &= \left\|P(y-\pi(\mathbf{x}^T\beta_n))\mathbf{x}\right\| \\ &= \left\|P(y-\pi(\mathbf{x}^T\beta_0)+\pi(\mathbf{x}^T\beta_0)-\pi(\mathbf{x}^T\beta_n))\mathbf{x}\right\| \\ &= \left\|P(y-\pi(\mathbf{x}^T\beta_0))\mathbf{x}+P(\pi(\mathbf{x}^T\beta_0)-\pi(\mathbf{x}^T\beta_n))\mathbf{x}\right\|.\end{aligned}$$

If $\|\Psi(\beta_n)\| \to 0$, then $\Psi(\beta_n) \to 0$. Since $P(y-\pi(\mathbf{x}^T\beta_0))\mathbf{x} = 0$, we have from the above equality that $P(\pi(\mathbf{x}^T\beta_0)-\pi(\mathbf{x}^T\beta_n))\mathbf{x} \to 0$. Note

$$\begin{aligned}P(\pi(\mathbf{x}^T\beta_0)-\pi(\mathbf{x}^T\beta_n))\mathbf{x} &= P(\pi^{(1)}(c)(\mathbf{x}^T\beta_n-\mathbf{x}^T\beta_0)\mathbf{x}) \\ &\leq \frac{1}{4}P(\mathbf{x}^T(\beta_n-\beta_0)\mathbf{x}),\end{aligned}$$

where $c \in (\mathbf{x}^T\beta_0, \mathbf{x}^T\beta_n)$ and $\pi^{(1)}$ is the first derivative, for logistic link, $\pi^{(1)} = \frac{e^x}{(e^x+1)^2} \in (0, 1/4]$. Also we have $\mathbf{x} = (x_1, x_2, ..., x_p)^T$, $\beta_0 - \beta_n = \beta_d = (\beta_{d1}, \beta_{d2}, ..., \beta_{dp})^T$.

$$P(\mathbf{x}^T(\beta_0-\beta_n)\mathbf{x}) = P(\sum_{i=1}^{p}\beta_{di}\begin{pmatrix}x_1x_i\\ x_2x_i\\ ...\\ x_px_i\end{pmatrix}) = \sum_{i=1}^{p}\beta_{di}P(\begin{pmatrix}x_1x_i\\ x_2x_i\\ ...\\ x_px_i\end{pmatrix}) \to \mathbf{0}.$$

Since $P(\mathbf{X}\mathbf{X}^T) = E(\mathbf{X}\mathbf{X}^T)$ is nonsingular, we have $\beta_0 - \beta_n \to \mathbf{0}$, and so we prove $\|\beta_n - \beta_0\| \to 0$.

Lemma 3.3: *Assume that B3 and B4 hold. Then the class* $\{\psi_{\eta,\beta} : \beta \in \Theta, \|\eta-\eta_0\| < \delta\}$ *is P-Glivenko-Cantelli for some* $\delta > 0$, *where* $\eta_0 = (0, \mu_0, \Sigma_0)$.

Proof: To show that a class $\mathcal{F}$ of vector-valued functions $\psi : (\mathbf{x}, y) \mapsto \mathbb{R}^p$ to be Glivenko-Cantelli, we need to show each of the classes of coordinates $\psi^i : (\mathbf{x}, y) \mapsto \mathbb{R}$ with $\psi = (\psi^1, ..., \psi^p)^T$ ranging over $\mathcal{F}$ $(i = 1, 2, ..., p)$ is Glivenko-Cantelli.

The class $\mathcal{F} = \{\psi_\gamma : \gamma = (\beta, \eta) = (\beta, \alpha, \mu, \Sigma),\ \alpha \in [0, 1],\ \mu \in \mathbb{R}^q,\ \Sigma \in \mathcal{S}_+^q,\ \beta \in \Theta,\ \|\eta - \eta_0\| < \delta\}$ is a collection of measurable functions indexed by a bounded subset in $\Gamma \subset \Theta \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}^{q\times q}$, and $\mathcal{S}_+^q$ denotes a set of positive semidefinite matrices defined in $\mathbb{R}^{q\times q}$. This is because Σ is a variance-covariance matrix of continuous explanatory variables essentially, so it is symmetric and positive semidefinite. For the norm, I use $\|\eta - \eta_0\| = (\alpha^2 + \|\mu - \mu_0\|^2 + \|\Sigma - \Sigma_0\|^2 + \|\beta - \beta_0\|^2)^{1/2}$, with $\|\cdot\|$ denotes Euclidean norm of for vectors μ, β, and for matrix Σ, $\|\cdot\|$ denotes the general induced norm (without specifying p). For two values $\gamma_i = (\beta_i, \eta_i)$, $i = 1, 2$ of γ, we have

$$\begin{aligned}
\left|\psi^i_{\gamma_1}(\mathbf{x}, y) - \psi^i_{\gamma_2}(\mathbf{x}, y)\right| &= \left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_1}(\mathbf{x}, y) - w_{\eta_2}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y)\right| \\
&= |w_{\eta_1}(\mathbf{x})\phi^i_{\beta_1}(\mathbf{x}, y) - w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y) + \\
&\quad w_{\eta_1}(\mathbf{x}, y)\phi^i_{\beta_2}(\mathbf{x}, y) - w_{\eta_2}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y)| \qquad (3.3) \\
&\le \left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_1}(\mathbf{x}, y) - w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y)\right| + \\
&\quad \left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y) - w_{\eta_2}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x}, y)\right|.
\end{aligned}$$

But

$$\begin{aligned}\left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_1}(\mathbf{x},y) - w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x},y)\right| &= w_{\eta_1}(\mathbf{x})\left|(y-\pi(\mathbf{x}^T\beta_1))x_i - (y-\pi(\mathbf{x}^T\beta_2))x_i\right| \\ &\leq \left|\pi(\mathbf{x}^T\beta_1) - \pi(\mathbf{x}^T\beta_2)\right||x_i| \\ &\leq K_0\,|x_i|\left|\mathbf{x}^T\beta_2 - \mathbf{x}^T\beta_1\right| \\ &\leq K_0\,|x_i|\,\|\mathbf{x}\|\,\|\beta_2-\beta_1\|\,,\end{aligned} \tag{3.4}$$

where K_0 is an upper bound of the first derivative of link function $\pi^{(1)}(x)$. Using the Mean Value Theorem, for any x_1 and x_2, there exists $c \in (x_1, x_2)$ such that

$$|\pi(x_1) - \pi(x_2)| = \pi^{(1)}(c)\,|x_1 - x_2| < K_0\,|x_1 - x_2|\,.$$

Furthermore, we have

$$\begin{aligned}\left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x},y) - w_{\eta_2}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x},y)\right| &= |w_{\eta_1}(\mathbf{x}) - w_{\eta_2}(\mathbf{x})|\left|y-\pi(\mathbf{x}^T\beta_2)\right||x_i| \\ &= |w_{\alpha_1,\mu_1,\Sigma_1}(\mathbf{x}) - w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x}) + \\ &\quad w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x}) - w_{\alpha_2,\mu_2,\Sigma_2}(\mathbf{x})|\left|y-\pi(\mathbf{x}^T\beta_2)\right||x_i| \\ &\leq (|w_{\alpha_1,\mu_1,\Sigma_1}(\mathbf{x}) - w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x})| + \\ &\quad |w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x}) - w_{\alpha_2,\mu_2,\Sigma_2}(\mathbf{x})|)\,|x_i|\,,\end{aligned} \tag{3.5}$$

$$\begin{aligned}|w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x}) - w_{\alpha_2,\mu_2,\Sigma_2}(\mathbf{x})| &= |m(\alpha_1(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)) - \\ &\quad m(\alpha_2(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2))| \\ &\leq K_1|\alpha_2-\alpha_1|(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2),\end{aligned} \tag{3.6}$$

and

$$
\begin{aligned}
|w_{\alpha_1,\mu_1,\Sigma_1}(\mathbf{x}) - w_{\alpha_1,\mu_2,\Sigma_2}(\mathbf{x})| &= |m(\alpha_1(\mathbf{z}-\mu_1)^T\Sigma_1^{-1}(\mathbf{z}-\mu_1)) - \\
&\quad m(\alpha_1(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2))| \\
&\leq K_1\alpha_1|(\mathbf{z}-\mu_1)^T\Sigma_1^{-1}(\mathbf{z}-\mu_1) - \\
&\quad (\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)| \\
&\leq K_1|(\mathbf{z}-\mu_1)^T\Sigma_1^{-1}(\mathbf{z}-\mu_1) - \\
&\quad (\mathbf{z}-\mu_1)^T\Sigma_2^{-1}(\mathbf{z}-\mu_1)| + \\
&\quad K_1|(\mathbf{z}-\mu_1)^T\Sigma_2^{-1}(\mathbf{z}-\mu_1) - \\
&\quad (\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)|.
\end{aligned} \tag{3.7}
$$

Since Σ is a positive semidefinite matrix, so does Σ^{-1}. Denote the eigenvalues of Σ_2^{-1} to be $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q \geq 0$. Then there is a set of orthonormal eigenvectors of Σ_2^{-1}, say $p_1, ..., p_q$, s.t. $\Sigma_2^{-1}p_i = \lambda_i p_i$. In matrix form, there is an orthogonal matrix Q s.t.

$$Q^{-1}\Sigma_2^{-1}Q = Q^T\Sigma_2^{-1}Q = \Lambda.$$

Then we obtain

$$
\begin{aligned}
&|(\mathbf{z}-\mu_1)^T\Sigma_2^{-1}(\mathbf{z}-\mu_1)-(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)| \\
&=|\sum_{i=1}^{q}\lambda_i(p_i^T(\mathbf{z}-\mu_1))^2-\sum_{i=1}^{q}\lambda_i(p_i^T(\mathbf{z}-\mu_2))^2| \\
&=|\sum_{i=1}^{q}\lambda_i p_i^T(2\mathbf{z}-\mu_1-\mu_2)p_i^T(\mu_2-\mu_1)| \\
&\leq\sum_{i=1}^{q}\lambda_i\left\|p_i\right\|^2\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\left\|\mu_2-\mu_1\right\| \qquad (3.8)\\
&\leq\lambda_1\left\|\mu_2-\mu_1\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\sum_{i=1}^{q}\left\|p_i\right\|^2 \\
&=\rho(\Sigma_2^{-1})\left\|\mu_2-\mu_1\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\sum_{i=1}^{q}\left\|p_i\right\|^2 \\
&\leq\left\|\Sigma_2^{-1}\right\|\left\|\mu_2-\mu_1\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|,
\end{aligned}
$$

where $\rho(\Sigma_2^{-1})=\max|\lambda_i|$ is the spectral radius of Σ_2^{-1}, and $\rho(A)\leq\|A\|$ holds for any induced norm.

Similarly, $\Sigma_2^{-1}-\Sigma_1^{-1}$ is symmetric too, thus denote the eigenvalues of $\Sigma_2^{-1}-\Sigma_1^{-1}$ to be $\lambda_1^*\geq\lambda_2^*\geq...\geq\lambda_q^*$, and the orthonormal eigenvectors of $\Sigma_2^{-1}-\Sigma_1^{-1}$ as $p_1^*,...,p_q^*$ s.t. $(\Sigma_2^{-1}-\Sigma_1^{-1})p_i^*=\lambda_i^*p_i^*$. In matrix form, there is an orthogonal Q^* s.t.

$$(Q^*)^{-1}(\Sigma_2^{-1}-\Sigma_1^{-1})Q^*=(Q^*)^T(\Sigma_2^{-1}-\Sigma_1^{-1})Q^*=\Lambda^*.$$

Then we have

$$
\begin{aligned}
&|(\mathbf{z}-\mu_1)^T\Sigma_2^{-1}(\mathbf{z}-\mu_1)-(\mathbf{z}-\mu_1)^T\Sigma_1^{-1}(\mathbf{z}-\mu_1)| \\
&=|(\mathbf{z}-\mu_1)^T(\Sigma_2^{-1}-\Sigma_1^{-1})(\mathbf{z}-\mu_1)| \\
&=|(\mathbf{z}-\mu_1)^TQ^*\Lambda^*(Q^*)^T(\mathbf{z}-\mu_1)| \\
&=|[(Q^*)^T(\mathbf{z}-\mu_1)]^T\Lambda^*[(Q^*)^T(\mathbf{z}-\mu_1)]| \\
&=|\sum_{i=1}^{q}\lambda_i((p_i^*)^T(\mathbf{z}-\mu_1))^2| \\
&\leq \max_{i=1,\ldots,q}|\lambda_i^*|\sum_{i=1}^{q}((p_i^{'})^T(\mathbf{z}-\mu_1))^2 \\
&\leq \rho(\Sigma_2^{-1}-\Sigma_1^{-1})\left\|\mathbf{z}-\mu_1\right\|^2\sum_{i=1}^{q}\left\|(p_i^*)^2\right\| \\
&\leq \left\|\mathbf{z}-\mu_1\right\|^2\left\|\Sigma_2^{-1}-\Sigma_1^{-1}\right\|.
\end{aligned}
\tag{3.9}
$$

Then from (3.5), (3.6), (3.7), (3.8) and (1.6), we obtain

$$
\begin{aligned}
&\left|w_{\eta_1}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x},y)-w_{\eta_2}(\mathbf{x})\phi^i_{\beta_2}(\mathbf{x},y)\right| \\
&<K_1|\alpha_2-\alpha_1|(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)|x_i| \\
&\quad+K_1\left\|\Sigma_2^{-1}\right\|\left\|\mu_2-\mu_1\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\||x_i| \\
&\quad+K_1\left\|\mathbf{z}-\mu_1\right\|^2\left\|\Sigma_2^{-1}-\Sigma_1^{-1}\right\||x_i|.
\end{aligned}
\tag{3.10}
$$

Since $|\alpha_2-\alpha_1|,\ \left\|\mu_2-\mu_1\right\|,\ \left\|\Sigma_2-\Sigma_1\right\|,\ \left\|\beta_2-\beta_1\right\|<\left\|\gamma_2-\gamma_1\right\|$, from (3.3),

(3.4) and (3.11), we can now give a bound for ψ :

$$\begin{aligned}
\left|\psi^i_{\gamma_1}(\mathbf{x},y)-\psi^i_{\gamma_2}(\mathbf{x},y)\right| &< K_0\left\|\beta_2-\beta_1\right\|\left|x_i\right|\left\|\mathbf{x}\right\|+\\
&\quad K_1|\alpha_2-\alpha_1|(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)|x_i|+\\
&\quad K_1\left\|\Sigma_2^{-1}\right\|\left\|\mu_2-\mu_1\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\left|x_i\right|+\\
&\quad K_1\left\|\mathbf{z}-\mu_1\right\|^2\left\|\Sigma_2^{-1}-\Sigma_1^{-1}\right\|\left|x_i\right|\\
&\leq \left(K_0\left|x_i\right|\left\|\mathbf{x}\right\|+K_1(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)|x_i|+\right.\\
&\quad K_1\left\|\Sigma_2^{-1}\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\left|x_i\right|+\\
&\quad \left.K_1\left\|\mathbf{z}-\mu_1\right\|^2\left|x_i\right|\right)\left\|\gamma_2-\gamma_1\right\|\\
&= L^i(\mathbf{x})\left\|\gamma_2-\gamma_1\right\|,\ \text{every } \gamma_1,\gamma_2.
\end{aligned}$$

For each $\psi^i, i=1,...,p$, we have derived a Lipschitz condition, so for ψ then we have

$$\left|\psi_{\gamma_1}(\mathbf{x},y)-\psi_{\gamma_2}(\mathbf{x},y)\right|<L(\mathbf{x})\left\|\gamma_2-\gamma_1\right\|,\ \text{every } \gamma_1,\gamma_2,$$

where $L(\mathbf{x})=K_0\mathbf{x}\left\|\mathbf{x}\right\|+K_1(\mathbf{z}-\mu_2)^T\Sigma_2^{-1}(\mathbf{z}-\mu_2)\mathbf{x}+K_1\left\|\Sigma_2^{-1}\right\|\left\|2\mathbf{z}-\mu_1-\mu_2\right\|\left|x_i\right|+K_1\left\|\mathbf{z}-\mu_1\right\|^2\mathbf{x}$.

Now consider the bracketing entropy relative to $L_r(P)$-norm

$$\left\|\psi_i\right\|_{P,r}=\left(P\left|\psi_i\right|^r\right)^{1/r}.$$

Use brackets of the type $[\psi_\gamma-\varepsilon L,\psi_\gamma+\varepsilon L]$ for γ ranging over a suitable chosen subset of Γ, and these brackets have $L_r(P)$-size $2\varepsilon\left\|L\right\|_{P,r}$. If γ ranges over a grid of mesh width ε over Γ, then the brackets $[\psi_\gamma-\varepsilon L,\psi_\gamma+\varepsilon L]$ range over

$\mathcal{F}$. By the Lipschitz condition, we obtain

$$\psi_{\gamma 1} - \varepsilon L \leq \psi_{\gamma 2} \leq \psi_{\gamma 1} + \varepsilon L, \text{ if } \|\gamma_2 - \gamma_1\| \leq \varepsilon,$$

so we need as many brackets as we need balls of radius $\frac{\varepsilon}{2}$ to cover Γ, or we need less than $(\text{diam } \Gamma/\varepsilon)^{2p+2}$ cubes with size ε to cover parameter space Γ. If $P|L^i|^r < \infty$, then there exists a constant J, depending on Γ and p only, such that the bracketing numbers satisfy

$$N_{[\,]}(\varepsilon, \mathcal{F}, L_r(P)) \leq J(\frac{\text{diam } \Gamma}{\varepsilon})^p, \text{ every } 0 < \varepsilon < \text{diam } \Gamma.$$

Since all $\psi \in \mathcal{F}$ are continuous functions, they are measurable. If B3 is satisfied, then $P|L| < \infty$, and thus the class $\mathcal{F}$ is P-Glivenko-Cantelli from Theorem 19.4 (Glivenko-Cantelli) in van der Vaart (1998).

Theorem 3.1: *If B1, B2, B3 and B4 are satisfied, then estimators $\hat{\beta}_n$ as the solution to the estimating equation $\Psi_n(\hat{\beta}_n) = 0$ converges in probability to β_0.*

Proof: Denote

$$\Psi(\beta) = P\psi_{\beta,0} = P\phi_\beta$$
$$\Psi(\beta, \eta) = P\psi_{\beta,\eta} = P\phi_\beta w_\eta$$
$$\Psi_n(\beta) = \Psi_n(\beta, \eta) = \frac{1}{n}\sum_{i=1}^{n} \psi_{\beta,\eta}$$
$$\Psi_n(\beta, \hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \psi_{\beta,\hat{\eta}}.$$

Note that $\Psi(\hat{\beta}_n) = (\Psi(\hat{\beta}_n) - \Psi_n(\hat{\beta}_n)) + \Psi_n(\hat{\beta}_n)$. Then if we can show that $\sup_{\beta\in\Theta} \|\Psi_n(\beta) - \Psi(\beta)\| = o_P(1)$, then $\left\|\Psi(\hat{\beta}_n)\right\| = o_P(1)$ since $\Psi_n(\hat{\beta}_n) = o_P(1)$. Then from Lemma 2.2, we have $\left\|\hat{\beta}_n - \beta_0\right\| = o_P(1)$. To show $\sup_{\beta\in\Theta} \|\Psi_n(\beta) - \Psi(\beta)\| =$

$o_P(1)$, we factor it as follows:

$$\begin{aligned}
&\sup_{\beta\in\Theta} \|\Psi_n(\beta,\hat{\eta}) - \Psi(\beta)\| \\
&= \sup_{\beta\in\Theta} \|\Psi_n(\beta,\hat{\eta}) - \Psi_n(\beta,\eta) + \Psi_n(\beta,\eta) - \Psi(\beta,\eta) + \Psi(\beta,\eta) - \Psi(\beta,0)\| \\
&\leq J_1 + J_2 + J_3,
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \sup_{\beta\in\Theta} \|\Psi_n(\beta,\hat{\eta}) - \Psi_n(\beta,\eta)\| \\
J_2 &= \sup_{\beta\in\Theta} \|\Psi_n(\beta,\eta) - \Psi(\beta,\eta)\| \\
J_3 &= \sup_{\beta\in\Theta} \|\Psi(\beta,\eta) - \Psi(\beta,0)\| .
\end{aligned}$$

From Lemma 3.3, we know that $\mathcal{F}$ is a P-Glivenko-Cantelli class, so $J_2 \xrightarrow{as} 0$. For J_1,

$$\begin{aligned}
J_1 &= \sup_{\beta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} \phi_\beta(\mathbf{x}_i, y_i)(w_{\hat{\eta}}(\mathbf{x}_i) - w_\eta(\mathbf{x}_i)) \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\beta\in\Theta} \|\phi_\beta(\mathbf{x}_i, y_i)(w_{\hat{\eta}}(\mathbf{x}_i) - w_\eta(\mathbf{x}_i))\| \\
&= \frac{1}{n}\sum_{i=1}^{n} |w_{\hat{\eta}}(\mathbf{x}_i) - w_\eta(\mathbf{x}_i)| \sup_{\beta\in\Theta} \|\phi_\beta(\mathbf{x}_i, y_i)\| .
\end{aligned}$$

From (3.11), we obtain

$$\begin{aligned}
|w_{\hat{\eta}}(\mathbf{x}_i) - w_\eta(\mathbf{x}_i)| < K_1|\hat{\alpha}-\alpha|(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu) \\
+ K_1 \left\|\Sigma^{-1}\right\| \|\hat{\mu}-\mu\| \|2\mathbf{z}-\hat{\mu}-\mu\| \\
+ K_1 \|\mathbf{z}-\hat{\mu}\|^2 \left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\| .
\end{aligned} \tag{3.11}$$

Since $\hat{\eta}-\eta = (\hat{\alpha}-\alpha, \hat{\mu}-\mu, \hat{\Sigma}-\Sigma) = o_P(1)$, we have $\hat{\alpha}-\alpha = o_p(1)$, $\hat{\mu}-\mu = o_p(1)$

and $\hat{\Sigma} - \Sigma = o_p(1)$. Also, $\hat{\Sigma}^{-1} - \Sigma^{-1} = \hat{\Sigma}^{-1}(\Sigma - \hat{\Sigma})\Sigma^{-1}$, and so we have $\hat{\Sigma}^{-1} - \Sigma^{-1} = o_p(1)$. Then from (3.10) it follows that $|w_{\hat{\eta}}(\mathbf{x}_i) - w_\eta(\mathbf{x}_i)| = o_p(1)$. Since B2, B3 are satisfied we also have $\sup_{\beta\in\Theta} \|\phi_\beta(\mathbf{x}, y)\|^2 < \infty$. Then we obtain $J_1 \xrightarrow{p} 0$.

For J_3,

$$\begin{aligned} J_3 &= \sup_{\beta\in\Theta} \|P\phi_\beta(\mathbf{x}, y)(w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x}))\| \\ &\leq \sup_{\beta\in\Theta} P\,\|\phi_\beta(\mathbf{x}, y)(w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x}))\| \\ &\leq \sup_{\beta\in\Theta} P^{1/2}\,\|\phi_\beta(\mathbf{x}, y)\|^2\, P^{1/2}|w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x})|^2 \\ &= P^{1/2}|w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x})|^2 \sup_{\beta\in\Theta} P^{1/2}\,\|\phi_\beta(\mathbf{x}, y)\|^2\,, \end{aligned}$$

and the last inequality used the Cauchy-Schwarz inequality. We know

$$w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x}) = 1 - m(\eta(\mathbf{z} - \mu)^T\Sigma^{-1}(\mathbf{z} - \mu)) \xrightarrow{p} 0.$$

Then since $\hat{\eta} = o_P(1)$, $m(0) = 1$ and $|w_{\hat{\eta}}(\mathbf{x}) - w_{\eta_0}(\mathbf{x})|^2$ is bounded by 1, using the Dominated Convergence Theorem we obtain

$$\lim_{n\to\infty} P^{1/2}|w_\eta(\mathbf{x}) - w_{\eta_0}(\mathbf{x})|^2 \xrightarrow{p} 0.$$

Also we have,

$$\sup_{\beta\in\Theta} P\,\|\phi_\beta(\mathbf{x}, y)\|^2 = \sup_{\beta\in\Theta} P\,\left\|\pi(y - \mathbf{x}^T\beta)\mathbf{x}\right\|^2 \leq P\,\|\mathbf{x}\|^2\,,$$

and so from assumption B3, we get $\sup_{\beta\in\Theta} P\,\|\phi_\beta(\mathbf{x}, y)\|^2 < \infty$. It then follows that $J_3 \to 0$. This completes the proof. ◄

3.4 Monte Carlo studies

In this section a Monte Carlo Study was carried out to examine the finite-sample efficiency and robustness properties of the proposed estimator $\hat{\beta}_n$ given in Section 3.2. For the initial robust estimators of the location and scatter of the covariates, $\hat{\mu}$ and $\hat{\Sigma}$, minimum covariance determinant (MCD) estimators were used. The weight function $m(x)$ used in the simulation is the same as (2.14), $m(x) = \frac{1}{(1+x^2)^5}$.

For comparison purposes, I computed the following estimators:

1. Maximum likelihood estimator (MLE).

2. Weighted maximum likelihood estimator with fixed hard rejection weight (F-WMLE) with $\alpha = 0.10$.

3. Weighted maximum likelihood estimators with adaptive hard rejection weight (A-WMLE) with $\sigma = 0.10$.

4. The proposed estimator, weighted maximum likelihood estimator with continuous weight (RECMLE).

3.4.1 Efficiencies under the clean model

We considered clean logistic regression models with intercept and normally distributed covariates. Specifically, let $(\mathbf{x}_1, y_1), ..., (x_n, y_n)$ be a random model that follows the logistic model with $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip-1})^T$ and such that $(x_{i1}, ..., x_{ip-1})^T$ has a $N_{p-1}(\mu, \Sigma)$ distribution, and $\mu = \mathbf{0}$, $\Sigma = I$ was chosen in the following simulation studies. Unlike LSE of linear regression models, there isn't model equivariance in the logistic regression. Different sets of regression

parameter β_0 will influence the performance of estimators greatly, so β_0 must be chosen carefully. I considered different kind of conditions, when $\beta_0 = (0, 1.79)$, $P_0(y = 1) = 0.5$, so the distribution of $\pi(\mathbf{x}^T\beta_0)$ is symmetric; when $\beta_0 = (-1, 1.18)$, $P_0(y = 1) \approx 0.3$, so the distribution of $\pi(\mathbf{x}^T\beta_0)$ is skewed.

Samples with $n = 100, 500$ and $p = 2, 5$ were considered (when $p = 5$, $\beta_0 = (0, 1.79, 0, 0, 0)$ and $(-1, 1.18, 0, 0, 0)$). For each value of p and n, 1000 samples were generated, and for each estimator β_n, I computed $\sqrt{n} * bias$ and $n * var$. Since the MLE is asymptotically efficient, I computed the relative squared efficiency with respect to the MLE:

$$\text{REF} = \frac{\sum_{i=1}^{1000}(\left\|\beta_{ni}^{MLE} - \beta_0\right\|)^2}{\sum_{i=1}^{1000}(\|\beta_{ni} - \beta_0\|)^2},$$

where β_{ni} and β_{ni}^{MLE} are the i-th generated values of β_n and the MLE, respectively. Any estimator with REF close to 1 would be considered very efficient. Also since the MLE is only asymptotically efficient, the simulated efficiency heavily depends on the sample size and the number of parameters that need to be estimated.

Tables 3.1 and 3.2 exhibit comparisons of bias and relative efficiency of the estimators considered in the simulation when $p = 2, 5$. As we can see, for all estimators, the relative efficiency increases as the sample size n increases. For all MLE based estimators, the relative efficiency for the symmetric target parameter β_0 (distribution of $\pi(\mathbf{x}^T\beta_0)$ is symmetric, for example, $\beta_0 = (0, 1.79)$) is larger than skewed ones (for example, $\beta_0 = (-1, 1.18)$). Further, under all settings, the performance in efficiency of different estimators have the following order: MLE $\approx$ RECMLE $>$ A-WMLE $>$ F-WMLE. As for the bias, all MLE based estimators have almost similar biases. In summary, the proposed

Table 3.1: *Bias and variance of estimators of β_{02} for clean logistic models ($p = 2$)*

estimator	$\beta_0 = (0, 1.79)$			$\beta_0 = (-1, 1.18)$		
	$\sqrt{n} * bias$	$n * var$	REF	$\sqrt{n} * bias$	$n * var$	REF
$n = 100$						
MLE	0.88	16.76	1.00	0.59	10.58	1.00
F-WMLE	0.88	19.95	0.84	0.60	14.81	0.71
A-WMLE	0.87	17.97	0.93	0.57	12.03	0.88
RECMLE	0.88	16.77	1.00	0.59	10.56	1.00
$n = 500$						
MLE	0.31	13.97	1.00	0.29	9.32	1.00
F-WMLE	0.27	16.17	0.86	0.27	12.94	0.72
A-WMLE	0.30	14.20	0.98	0.26	9.76	0.96
RECMLE	0.31	13.97	1.00	0.29	9.32	1.00

estimator attains almost the full efficiency (relative to the MLE), while keeping the bias comparable for finite samples.

When comparing the respective cases of $p = 2$ and $p = 5$, that is comparing values of Table 3.1 to those of Table 3.2, we can see that both biases and variances increased when $p = 5$. This is probably because when $p = 5$, one has to estimate more parameters than in the case $p = 2$. However, the trend between different estimators seems to be the same.

3.4.2 Robustness under contaminated models

Consider point-mass contamination models defined by

$$P_*(y = 1|\mathbf{x}) = (1 - \epsilon)\pi(\mathbf{x}^T\beta_0) + \epsilon I\{\pi(\tilde{\mathbf{x}}^T\beta_0) \leq 0.5\},$$

where $\epsilon \in [0, 0.5)$ is the proportion of misclassified observations with possibly outlying covariates $\tilde{\mathbf{x}} = (\mathbf{u}, \tilde{\mathbf{z}})$ (only the continuous covariates are changed),

Table 3.2: *Bias and variance of estimators of β_{02} for clean logistic models ($p = 5$)*

estimator	$\beta_0 = (0, 1.79, 0, 0, 0)$			$\beta_0 = (-1, 1.18, 0, 0, 0)$		
	$\sqrt{n} * bias$	$n * var$	REF	$\sqrt{n} * bias$	$n * var$	REF
$n = 100$						
MLE	1.73	22.23	1.00	1.08	12.85	1.00
F-WMLE	1.76	25.89	0.86	1.10	16.62	0.77
A-WMLE	1.71	23.38	0.95	1.05	14.29	0.90
RECMLE	1.67	22.57	0.99	1.04	13.13	0.98
$n = 500$						
MLE	0.80	15.08	1.00	0.48	9.85	1.00
F-WMLE	0.82	16.56	0.91	0.46	11.69	0.84
A-WMLE	0.80	15.26	0.99	0.50	10.07	0.98
RECMLE	0.80	15.06	1.00	0.48	9.84	1.00

$\tilde{\mathbf{x}} \in \mathbb{R}^p$ and β_0 is the target parameter.

In the simulation, I have chosen $\tilde{x}(k) = (1, k, 0_{p-2})$ with $k = 2$ and $k = 5$, and $\epsilon = 0.1$ and $\epsilon = 0.2$, and the sample size was taken to be $n = 100$. Tables 3.3 and 3.4 display the comparison of mean squared errors of estimators of β_{02}.

When $k = 2$, which means that the leverage of contaminated point is not very large, the difference of MSEs between all estimators are not very large. This means that when the contaminated points are not in the boundary of the covariate space, it's very hard to detect it, and thus the robust estimators do not have a better performance than the MLE. On the other hand, it can be seen that when $k = 5$, all robust estimators have significantly smaller MSE than the MLE. Also, by comparing conditions of different β_0, the skewed case ($\beta_0 = (-1, 1.18)$) has a smaller MSE than the symmetric case ($\beta_0 = c(0, 1.79)$).

When $k = 5$, robustness of three robust estimators examined have the following order in performance: RECWLS > A-WMLE > F-WMLE. So the proposed estimator at least has a comparable robustness than the adaptive

Table 3.3: *Mean squared errors* $\times 10$ *of estimators of* β_{02} *under point-mass contaminations at* $\tilde{x} = (1, k)$

estimator	$\beta_0 = (0, 1.79)$		$\beta_0 = (-1, 1.18)$	
	$k = 2$	$k = 5$	$k = 2$	$k = 5$
$\epsilon = 0.10$				
MLE	14.54	30.63	4.56	12.56
F-WMLE	12.58	2.05	4.51	1.42
A-WMLE	10.49	1.99	3.85	1.29
RECMLE	13.66	1.55	4.32	1.10
$\epsilon = 0.20$				
MLE	25.26	38.02	9.15	16.66
F-WMLE	29.83	2.45	11.20	1.52
A-WMLE	30.29	2.42	11.33	1.47
RECMLE	25.35	2.29	9.19	1.54

and fixed threshold weighted MLEs. Comparing the two cases $p = 2$ and $p = 5$, the trends are the same, and the proposed estimator still has the best performance overall. Also, by looking at the MSE values for the two cases $\epsilon = 0.10$ and $\epsilon = 0.20$ separately, we see that the comparisons between different estimators are the same. However, we can see that the MSE is increased considerably when ϵ increased when $k = 2$ but not so much when $k = 5$. This is because when $k = 5$, all robust estimators easily detect most of the contaminated observations, so the increase of proportions of contamination would not influence the estimators that much, and that is not the case when $k = 2$.

In summary, all robust estimators didn't show very good robustness properties when the contamination points are not in the boundary of the covariate space. Further, when the contaminated points have high leverage, the proposed estimator has the best robustness performance.

Table 3.4: *Mean squared errors* $\times 10$ *of estimators of* β_{02} *under point-mass contaminations at* $\tilde{x} = (1, k, 0, 0, 0)$

estimator	$\beta_0 = (0, 1.79, 0, 0, 0)$		$\beta_0 = (-1, 1.18, 0, 0, 0)$	
	$k = 2$	$k = 5$	$k = 2$	$k = 5$
$\epsilon = 0.10$				
MLE	14.59	30.99	4.41	12.63
F-WMLE	17.69	4.07	5.69	2.06
A-WMLE	16.05	3.97	5.00	2.00
RECMLE	17.92	2.92	5.89	1.65
$\epsilon = 0.20$				
MLE	25.39	39.41	9.31	16.95
F-WMLE	25.43	4.35	9.43	2.67
A-WMLE	25.68	3.98	9.41	2.25
RECMLE	26.03	3.43	9.35	2.04

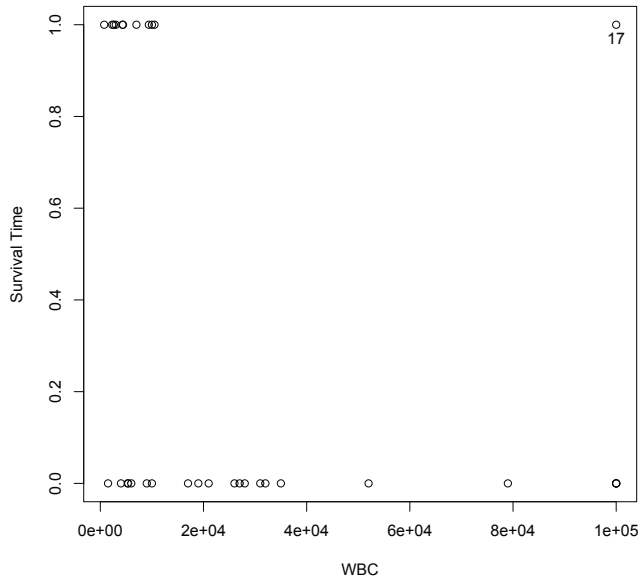
3.5 Real Data Analysis

The dataset analyzed here from Cox and Oakes (1984, p. 9.), and it is based on 33 patients who died from acute myelogenous leukemia. Three variables were measured for each patient: WBC, AG and Time. The response variable time is patient's survival time in weeks, and we transformed it into a binary variable with $Y = 1$ signifying patients with survival time longer than 52 weeks, and $Y = 0$ to those who did not. WBC measured the patient's white blood cell count at the time of diagnosis, and AG (present = 1, absent = 0) was about the test result indicating the presence or absence of a morphologic characteristic of white blood cells. AG present patients were identified by the presence of Auer rods and/or significant granulation of the leukemia cells in the bone marrow at the time of diagnosis.

The variable WBC is associated with low survival time, and a plot of time against WBC in Figure 3.1 shows that patient No. 17 is atypical, with a very

high WBC number and $Y = 1$. A logistic regression model was fit using binary

Figure 3.1: *Scatterplot of survival time against WBC for Leukemia data*



*Survival time has been transformed into a binary variable ($Y = 1$ if survival time longer than 52, $Y = 0$ if not)

.

survival time Y as response and WBC and AG as covariates. The estimators used here are the MLE, MLE_{17} (MLE after deletion of observation No.17), F-WMLE (weighted MLE with fixed threshold values), A-WMLE (weighted MLE with adaptive threshold values) and the proposed estimator RECMLE. In Table 3.5, the estimated parameters and their estimated standard errors (in parenthesis) are reported. As we can see, the MLE is very sensitive to atypical observations, and the observation No. 17 reduced the effect of WBC close to 0. All other robust estimators show a similar behaviour to that of the MLE_{17}.

Table 3.5: *The estimated regression parameters for Leukemia data with standard errors in parentheses*

Estimate	Intercept	WBC($\times 10^{-4}$)	AG
MLE	-1.3(0.81)	-0.32(0.18)	2.26(0.95)
MLE_{17}	0.21(1.08)	-2.35(1.35)	2.56(1.23)
F-WMLE	0.17(1.11)	-2.27(1.45)	2.54(1.23)
A-WMLE	0.18(1.11)	-2.28(1.44)	2.54(1.23)
RECMLE	0.15(1.20)	-2.18(1.55)	2.47(1.29)

*WBC measures the patient's white blood cell count at the time of diagnosis, and AG is the test result indicating the presence or absence of a morphologic characteristic of WBC.

Chapter 4

Concluding Remarks and Future Directions

4.1 Concluding Remarks

This thesis has proposed robust estimators for regression models that achieve high efficiency and high robustness properties.

In Chapter 2, a weighted least squares estimator (RECWLS) has been proposed for the linear regression model with an adaptive weight function. Theoretical results such as the consistency and asymptotic normality of the proposed estimator have been established. The derived asymptotic distribution suggests that the proposed RECWLS estimator achieves the full efficiency under normal errors models. In order to study finite sample properties of the proposed estimator, a simulation study was conducted. For model with normal errors, the RECWLS estimator shows relative efficiency of almost close to 1 with respect to the LSE, even for small sample sizes ($n = 50$ for $p = 2$, and $n = 100$ for $p = 5$), and this is not achieved by other threshold value based weighted

least square estimators. For models with heavy-tailed errors (student-t_3 was used), RECWLS estimator has comparable performance in efficiency when compared with other weighted least square estimators. All weighted LSE estimators studied increased their efficiency of the initial high BP robust estimators significantly. For models with normal errors and some fraction of outlier contamination, the maximum MSE suggests that the RECWLS estimator increased the robustness properties of the initial robust estimators, and it has better performance than the threshold value based estimators when the contamination of y-value or the leverage is more extreme. The proposed RECWLS estimator is applied to a real data set and the result showed that it had similar performance as the MM-estimator.

In Chapter 3, a RECMLE has been proposed for the logistic regression model. The asymptotic consistency of the proposed estimator was proved. A simulation study was conducted to examine the finite sample properties of the proposed RECMLE. For the clean model, the proposed estimator shows a relative efficiency of almost 1 with respect to the MLE, without worsening the bias. For point-mass contamination models, the proposed estimator, along with other weighted MLE estimators with threshold cutoff values, didn't show a significant difference in robustness properties compared with the MLE when the contamination has a small leverage. However, for high leverage contaminations, the proposed RECMLE significantly improved the robustness performance, and it did exhibit a much better performance than the threshold value based estimators. The performance of the proposed RECMLE on a real data set showed that it was comparable to the existing robust estimators in practice.

In summary, given the overall excellent performance in efficiency and com-

parable robustness properties provided by the proposed estimators, I believe that my proposed estimators are very useful in practical applications over the existing methods.

4.2 Future Directions

The following future research directions can be studied based on the work of this thesis:

1. Other asymptotic properties of the proposed estimators still need to be developed. Asymptotic breakdown points and influence functions need to be derived to assess the robustness property theoretically. The distributional properties of proposed estimators can be analyzed with more details, and corresponding confidence interval and robust hypothesis testing methods can also be studied.

2. More simulation studies can be done. Since the MLE isn't regression, affine and scale equivariant, so more target parameter values can be chosen to access the performance of proposed estimators in a more comprehensive way. Other kinds of contaminated models except than point-mass contaminated model, like multi different points contamination, can be studied to test the robustness of proposed estimators under different conditions.

3. The strategy used here to construct robust estimators can also be generalized to generalized linear regression models, linear mixed model and other different models.

4. The proposed estimator is applied on two small data sets in this thesis, and it can be used on some more complicated real data sets, especially those unstructured data sets with many potential outliers or errors.

Bibliography

[1] Albert, A., Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1), 1-10.

[2] Bianco, A. M., Yohai, V. J. (1996). Robust estimation in the logistic regression model (pp. 17-34). Springer, New York.

[3] Cox, D. R., Oakes, D. (1984). Analysis of survival data (Vol. 21). CRC Press.

[4] Donoho, D. L., Huber, P. J. (1983). The notion of breakdown point. A Festschrift for Erich L. Lehmann, 157184.

[5] Gervini, D., Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. The Annals of Statistics, 583-616.

[6] Gervini, D. (2005). Robust adaptive estimators for binary regression models. Journal of Statistical Planning and Inference, 131 (2), 297-311.

[7] Gray, J. B. (1985). *Graphics for Regression Diagnostics*. American Statistical Association Proceedings of the Statistical Computing Section, ASA, Washington, D.C., pp. 102-107.

[8] Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. University of California Press.

[9] Hampel, F. R. (1971). A general qualitative definition of robustness. The Annals of Mathematical Statistics, 1887-1896.

[10] Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69 (346), 383-393.

[11] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., Stahel, W. A. (1986). *Robust Statistics: the Approach based on Influence Functions*. Series in Probability and Mathematical Statistics.

[12] He, X., Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. The Annals of Statistics, 2161-2167.

[13] Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73-101.

[14] Huber, P. J. (2011). Robust statistics (pp. 1248-1251). Springer, Berlin Heidelberg.

[15] Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., Ritov, Y. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press.

[16] Künsch, H. R., Stefanski, L. A., Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association, 84 (406), 460-466.

[17] Leroy, A. M., Rousseeuw, P. J. (1987). *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

[18] Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American statistical association, 79 (388), 871-880.

[19] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. Contributions to probability and statistics, 2, 448-485.

[20] van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

[21] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, 642-656.